Inference With Unequal Cluster Sizes

Lisa Hermans

Promotor: Prof. dr. Geert Molenberghs

Co-Promotor: Prof. dr. Marc Aerts

Acknowledgements

I realise how lucky I am to have the opportunity to accomplish this PhD, and most of all how lucky I am with all the wonderful people who supported me along the way.

A sincere word of thank you goes to my supervisor, Prof. dr. Geert Molenberghs. It has been a pleasure and honour working with you. I'm very grateful for all you have taught me. Thank you for all your good advice and the wonderful opportunities you have given me. You are the best promotor I could ever wish for! Prof. dr. Marc Aerts, my co-promotor, Prof. dr. Geert Verbeke and Prof. dr. Michael G. Kenward, thank you for the interesting discussions and valuable suggestions. Also to the other members of the jury, Prof. dr. Christel Faes, dr. Cristina Sotto and dr. Wim Van der Elst, I would like to express my sincere gratitude for your help throughout the years.

I met a lot of people at I-BioStat and it was a pleasure working with you. Especially my fellow (former and current) teaching assistants, I believe we were a great team. Kim, I really missed you in our office at CenStat. Fortunately, we continue our statistical discussions and heart-warming conversations when we meet for lunch. Thomas, you have a great motivation which inspires people, thank you for the great laughs we shared. Ruth, I learned a lot from you and you are always there for good advice, thank you for the many nice talks. Thank you to all my dear colleagues from I-BioStat, you are all such great people!

Finally, I would like to thank the people who are closest to my heart. I have such wonderful family and friends that support me. To my friends, your friendship is a special gift; generously given, happily accepted, and deeply appreciated. Eline, my sister, really is the best. You are always there for me and well, as often, great minds do think alike ;). Mom and dad, you taught me important values in life and I really appreciate that a lot.

You stood by my side at every step of the way and encouraged me no matter what. I really hope I can be the same kind of parent. Yoeri, thank you for loving me and bringing out the best in me. You are a remarkable man. I love you and our baby girl, Fem.

Many thanks to all of you!

Lisa Hermans
19 April 2019
Diepenbeek

Publications

The materials presented here, are based on the following publications:

Hermans, L., Molenberghs G., Aerts, M., Kenward, M.G. and Verbeke, G. (2018). A tutorial on the practical use and implication of complete sufficient statistics. *International Statistical Review*, 86(3), 403–414.

Hermans, L., Nassiri, V., Molenberghs, G., Kenward, M.G., Van der Elst, W., Aerts, M., and Verbeke, G. (2018). Clusters with unequal size: maximum likelihood versus weighted estimation in large samples. *Statistica Sinica*, 28(3), 1107–1132.

Hermans, L., Nassiri, V., Molenberghs, G., Kenward, M.G., Van der Elst, W., Aerts, M., and Verbeke, G. (2018). Fast, closed-form, and efficient estimators for AR(1) models. *Communications in Statistics: Simulation and Computation*, 47(5), 1492–1505.

Hermans, L., Ivanova, A., Sotto, C. Molenberghs, G., Kenward, M.G. and Verbeke, G. (2018). Doubly robust pseudo-likelihood for incomplete hierarchical binary data. *Statistical Modeling*, 00, 000–000.

Hermans, L., Molenberghs, G., Verbeke, G., Kenward, M.G., Mamouris, P. and Vaes, B. (2018) Optimal weighted estimation versus Cochran-Mantel-Haenszel. *Communications in Statistics: Simulation and Computation*. Submitted.

Other publications:

Hermans, L., Molenberghs, G., Kenward, M.G., Van der Elst, W., Nassiri, V., Aerts, M. and Verbeke, G. (2015). Clusters with random size: maximum likelihood versus weighted estimation. *Proceedings of the 30th International Workshop on Statistical Modelling*. Linz, Austria. H. Friedl and H. Wagner (eds). Johannes Kepler University Linz. p. 215–220.

Raymaekers, V., Brenard, C., **Hermans, L.**, Frederix, I., Staessen, J.A. and Dendale, P. (2019). How to reliably diagnose arterial hypertension: lessons form 24h blood pressure monitoring. *Blood Pressure.*, 28(2), 93–98.

Van der Elst, W., **Hermans, L.**, Verbeke, G., Kenward, M.G., Nassiri, V. and Molenberghs, G. (2016). Unbalanced cluster sizes and rates of convergence in mixed-effects models for clustered data. *Journal of Statistical Computation and Simulation.*, 11, 2123–2139.

Contents

List of Tables

List of Figures

List of Abbreviations

AC	Available Cases
AR	Autoregressive
CC	Complete Cases
CP	Complete Pairs
CS	Compound-Symmetry
DR	Doubly Robust
GEE	Generalized Estimating Equations
L	Likelihood
MAR	Missing At Random
MCAR	Missing Completely At Random
MH	Mantel-Haenszel Estimator
MI	Multiple Imputation
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimator
MNAR	Missing Not At Random
MSE	Mean Squared Error
PL	Pseudo-Likelihood
SD	Standard Deviation
SR	Single Robust
WGEE	Weighted Generalized Estimating Equations

PART I

Introduction and Background

Chapter 1

General Introduction

The vast majority of settings for which frequentist statistical properties are derived assume a fixed, *a priori*, known sample size. Familiar properties then follow, such as, for example, the consistency, asymptotic normality, and efficiency of the sample average for the mean parameter, under a wide range of conditions.

Nevertheless, there is a variety of settings where sample size is random: 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 with its induced lack of balance; purely random sample sizes; time-to-event data; joint modeling of time-to-event data and longitudinal sequences; and random cluster sizes. Molenberghs *et al.* (2014) provide an overview of such situations. In all of these settings, the sample size, the length of the longitudinal sequence, or the censoring time for survival data, is itself a random variable that may depend on the data being collected. Further, the rule governing this may be deterministic or probabilistic. It is well known that specific issues can arise when evaluating the properties of statistical procedures under such sampling schemes, and much has been written about specific areas.

1.1 Random 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. It can occur for any outcome type, including continuous, binary, categorical, count, and event time.

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 missingness 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 variety 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; Hofman, 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 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 proportional to the population size. Litter sizes in pregnant rodents will truly be stochastic. When stochastic, the mechanism is 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 random are often termed informative.

1.2 Why Inference for Unequal Cluster Sizes?

Even when the cluster size contains no information about the scientific parameters, there are issues resulting from this that need further investigation. So this thesis does not focus on informative cluster sizes. In particular, the joint modelling of outcomes and cluster size is not considered. Attention is confined to the case where cluster size is unequal, but independent of both observed and unobserved outcomes. In doing so this research work distinguishes 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. This thesis focuses 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.

1.3 Aims and Scopes: Thesis Overview

In Part I of the thesis, **Chapters 2** and **3** consist of all necessary background material. In Chapter 2, notation, definitions and fundamental concepts and theory are given. A brief description of some hierarchical data sets used to demonstrate the proposed methodology in this thesis are outlined in Chapter 3.

Our own contributions can be read in Part II. The issues when doing likelihood inference and evaluating the properties of the estimator(s) under such sampling schemes, are the lack of completeness of the sufficient statistics and maximum-likelihood estimation may

be computationally prohibitive.

Completeness implies that any function of a sufficient statistic that has zero expectation for every value of the parameter indexing the parametric models class is the zero function almost everywhere. Although the definition of a *complete* sufficient statistic (Casella and Berger, 2001, pp. 285–286) is clear, its constructive verification in a given situation often involves tedious algebra. This is especially true in for example sequential trials, except for the simplest situation of two possible sample sizes only; such calculations are, quite literally, convoluted. Likewise, when completeness does not hold, the construction of counterexamples may or may not be straightforward. Nevertheless, a clear, simple, and easily verifiable criterion for completeness, of a constructive rather than an existential nature, would be welcome. For example, in a normal univariate sample with fixed sample size, a minimal sufficient statistic for the population mean is the sample sum, in contrast to the random sample case for which it is the sample sum *and* the realized (random) sample size. The parameter remains one-dimensional, but the minimal sufficient statistic is two-dimensional and incomplete. A general criterion is formulated in **Chapter 4** that starts from, but moves beyond, the length of a vector.

The relevance of complete sufficient statistics has been established through two theorems, Lehman-Scheffé (Casella and Berger, 2001) and Basu's (Basu, 1955) theorem. Completeness, combined with regularity conditions, provides a basis for estimators with desirable properties, such as unbiasedness and optimality. However, these properties are lost when dealing with incomplete sufficient statistics. What is more, incompleteness holds when the cluster size is non-constant for whatever reason. But as shown in Molenberghs *et al.* (2014) and Milanzi *et al.* (2016, 2015), this does not need to be a serious problem in practice and does not preclude the existence of estimators with very good properties. For example, it is very well-known that, when data are missing, likelihood and Bayesian inferences can be based on the observed-data likelihood, without any correction for the variable cluster size, i.e., without any correction for the missing-data mechanism. Importantly, though, such methods cannot, in general, by default be claimed to be optimal, given that the Lehman-Scheffé theorem (Casella and Berger, 2001) does not apply.

For medium to large sample sizes, full maximum likelihood or Bayesian inferences are statistically optimal and computationally feasible. 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. Small samples refers here to a small number of clusters; the clusters themselves may consist of smaller or larger numbers of within-cluster replication.

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. However, both of these methods are based on repeated sampling and will come at computational cost for high-dimensional data (Sikorska *et al.*, 2013). Therefore, in this thesis, the focus is on entirely non-iterative methods, bringing together the advantages of balanced data and simple averaging methodology.

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. 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.

Pseudo-likelihood has received considerable attention (Varin, Reid, and Firth, 2011; Molenberghs and Verbeke, 2005, Ch. 9, 12, 21, 24, 25; Aerts *et al.*, 2002, Ch. 6, 7). In this thesis, sample-splitting is used to propose a (near) optimal weighted estimator for these kind of data settings. It is a way to replace iterative optimization of a likelihood that does not admit an analytical solution, with closed-form calculations. As a simple, yet non-trivial, clustering paradigm, the compound-symmetry model, CS, and first-order autoregressive, AR(1), are considered. In **Chapter 5** a general split-sample approach for the CS model is provided. **Chapter 6** explores this further for the AR(1) model.

This idea could be extended to other grouped data settings. Specifically, the Mantel-Haenszel (MH) (Mantel and Haenszel, 1959) methodology for analysing the associations between binary variables involving stratification. Here group-specific odds ratios are combined using weights. In **Chapter 7** the MH estimator is contrasted with the optimal estimator, whose existence is demonstrated in spite of the absence of completeness. The MH estimator does not follow from optimality considerations. By comparing both esti-

mators, insight in specific nature the estimator and unique and interesting properties of the data settings for which it was developed, can be retrieved.

Next, the focus is on modeling hierarchical binary outcome data when the vector of planned measurements contains missing values. The process behind the missingness, as well as its impact on inference, need to be addressed.

The choice of inferential framework for analyzing incomplete data will depend largely upon the nature of missingness. Conventionally, the process driving the latter is classified according to the terminology of Little and Rubin (2002, Chap. 6). When missingness is independent of both the observed and unobserved outcomes, it is called *missing completely at random* (MCAR), while when the missingness is independent of the unobserved measurements, conditional on the observed ones, the process is said to be *missing at random* (MAR). When neither MCAR nor MAR holds, missingness is termed *missing not at random* (MNAR).

Often, direct likelihood is used as the basis for analyzing correlated outcomes under MAR. The unified modeling framework provided by the linear mixed model, yielding both random-effects as well as marginally interpretable regression parameters, is the dominant choice for Gaussian outcomes, while generalized linear mixed models remain popular for non-Gaussian outcomes, though marginalization is not always straightforward. Other likelihood-based options for marginal inference exist, such as the Bahadur (1961) model and the multivariate Dale or global odds ratio model (Molenberghs & Lesaffre, 1994, 1999) for binary data, but these involve complex likelihoods, can be computationally prohibitive in moderate to large studies, and are vulnerable to misspecification.

These issues have motivated the development of alternatives to likelihood, perhaps the most popular of which being generalized estimating equations or GEE (Liang and Zeger, 1986; Diggle et al., 2002; Molenberghs and Verbeke, 2005), along with variations or extensions such as GEE2 (Liang, Zeger, and Qaqish, 1992) and alternating logistic regressions (Carey, Zeger, & Diggle, 1993), when association parameters are also of scientific interest. Standard GEE is valid only under MCAR, but a weighted version (WGEE; Robins, Rotnitzky, and Zhao, 1995) has been developed, using Horvitz-Thompson ideas (Cochran, 1977), to allow valid use of GEE under MAR. The WGEE approach, however, tends to be biased when the model for the weights is misspecified (Beunckens, Sotto & Molenberghs, 2008; Molenberghs and Kenward, 2007). To this end, doubly robust approaches (Scharfstein, Rotnitzky, and Robins, 1999; Van der Laan & Robins, 2003; Bang & Robins, 2005; Rotnitzky, 2009; Birhanu et al., 2011), which further supplement the use of weights with a predictive model for the unobserved responses, given the observed ones, have been constructed. This not only removes or at least alleviates bias, but also increases efficiency.

Pseudo-likelihood (PL) methods (le Cessie & van Houwelingen, 1991; Geys, Molenberghs, and Lipsitz, 1998; Geys, Molenberghs, & Ryan, 1999; Aerts *et al.*, 2002) comprise yet another alternative to full likelihood. This is in contrast to GEE methods, where the score equations are replaced with alternative, simpler functions.

Pseudo-likelihood is different to full likelihood and is therefore not guaranteed to be valid under MAR. Rubin (1976) provided conditions for ignorability that are sufficient but not always necessary. Yi, Zeng, and Cook (2011) provide an example, using a pairwise (pseudo-)likelihood method for incomplete longitudinal binary data, that is ignorable under MAR, even though it is not a full likelihood approach. Molenberghs et al. (2011), on the other hand, propose a suite of corrections to pseudo-likelihood in its standard form, also to ensure its validity under MAR. These corrections hold for pseudo-likelihood in general and follow both single and double robustness ideas. They showed that, in contrast to the GEE case and in particular for both robust versions, PL-based estimating equations admit very convenient simplifications.

Molenberghs et al. (2011) applied the methodology to multivariate Gaussian responses and to a conditional model for clustered binary data. They provided a general outline with predominantly illustrative examples using normal and binary data. However, the marginal modeling of longitudinal binary data is very common in practice. Molenberghs et al. (2011) only sketched the methodology using a marginal Bahadur model for the binary responses; they did not pursue it in detail. The further development of doubly robust pseudo-likelihood for incomplete hierarchical binary data under MAR is investigated more in detail. The theoretical part, estimating equations and precision estimators, are calculated and reported for the first time. All can be found in **Chapter 8** of this thesis.

Finally, in Part III and so in the last **Chapter 9**, conclusions, ramifications and recommendations for further research are presented.

Extensive derivations and accompanying software code are excluded from the main text of this thesis, but provided in the Appendix. See further references in the chapters.

Chapter 2

Background Information

In this thesis important issues related to, and model strategies for, hierarchical data settings with unequal sizes are introduced. That material is presented in part II of the dissertation. This chapter discusses the important background theory and methodology that will be used to build on.

2.1 Hierarchical Data: Notation

Suppose that there is a sample of N independent clusters, the random variable Y_{ij} denotes the response for the ith study subject at the jth occasion ($i = 1, \ldots, N$, $j = 1, \ldots, n_i$). Independence across subjects is assumed.

First, among the N independent clusters K different cluster sizes n_k ($k = 1, \ldots, K$) can be distinguished. Let the multiplicity of cluster size n_k be equal to c_k. Evidently, $N = \sum_{k=1}^{K} c_k$. Within a subsample of clusters of size n_k, the ith ($i = 1, \ldots, c_k$) replicate is $\boldsymbol{Y}_i^{(k)}$.

Second, in case of missing data, $\boldsymbol{Y}_i$ ($i = 1, \ldots, N$) is divided into its observed ($\boldsymbol{Y}_i^o$) and missing ($\boldsymbol{Y}_i^m$) components. We further define a vector of missingness indicators $\boldsymbol{R}_i = (R_{i1}, R_{i2}, \ldots, R_{in_i})'$, with $R_{ij} = 1$ if Y_{ij} is observed and 0 otherwise. In the specific case of dropout in longitudinal studies, the vector $\boldsymbol{R}_i$ can be replaced by the dropout indicator $D_i = 1 + \sum_{j=1}^{n_i} R_{ij}$, denoting the time at which subject i drops out.

2.2 Concepts and Models

2.2.1 Complete Sufficient Statistics

Definition 2.1 (Completeness). *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, pp. 285–286)*

The relevance of completeness rests principally on two theorems. First, 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. Second, the connection with ancillarity follows from Basu's theorem (Basu, 1955): a statistic that is both boundedly complete and sufficient is independent of any ancillary statistic. See also Casella and Berger (2001, p. 287). Note, the theorems are implications rather than equivalences.

2.2.2 Bahadur Model

The Bahadur model (Bahadur, 1961) is a marginal model for correlated binary data, accounting for the associations via marginal correlations. Following Aerts *et al.* (2002) the marginal distribution of Y_{ij} is a Bernoulli distribution with $\nu_{ij} = P(Y_{ij} = 1)$, with the pairwise probability as $\nu_{ijk} = P(Y_{ij} = 1, Y_{ik} = 1)$, and the conditional probability as $\nu_{ik|j} = P(Y_{ik} = 1|y_{ij} = \ell)(\ell = 0, 1)$. With ρ_{ijk} the marginal correlation coefficient, the pairwise Bahadur probabilities take the form

$$\nu_{ijk} = \nu_{ij}\nu_{ik}\left[1 + \rho_{ijk}\frac{1-\nu_{ij}}{\sqrt{\nu_{ij}(1-\nu_{ij})}}\frac{1-\nu_{ik}}{\sqrt{\nu_{ik}(1-\nu_{ik})}}\right]. \tag{2.1}$$

The Bahadur model gives a closed form expression for the full joint distribution $f(\boldsymbol{y})$. The multivariate Bahadur probabilities are $f(\boldsymbol{y}_i) = f_1(\boldsymbol{y}_i)c(\boldsymbol{y}_i)$, with:

$$f_1(\boldsymbol{y}_i) = \prod_{j=1}^{n_i} \nu_{ij}^{y_{ij}} (1-\nu_{ij})^{1-y_{ij}}, \tag{2.2}$$

$$\begin{aligned} c(\boldsymbol{y}_i) = {} & 1 + \sum_{j_1<j_2} \rho_{ij_1j_2} e_{ij_1} e_{ij_2} + \sum_{j_1<j_2<j_3} \rho_{ij_1j_2j_3} e_{ij_1} e_{ij_2} e_{ij_3} + \\ & \cdots + \rho_{ij_1j_2\ldots j_{n_i}} e_{ij_1} e_{ij_2} \cdots e_{ij_{n_i}}, \end{aligned} \tag{2.3}$$

where $e_{ij} = \dfrac{y_{ij} - \nu_{ij}}{\sqrt{\nu_{ij}\,(1-\nu_{ij})}}$. This function is the product of the independence model $f_1(\boldsymbol{y}_i)$ and the correction factor $c(\boldsymbol{y}_i)$. Fitting a Bahadur model however, is not always straightforward. The parameter space of the marginal parameters is known to be of a

very special shape. Fitting high order Bahadur models implies an increasing amount of complex restrictions on the parameters space.

2.3 Pseudo-likelihood

2.3.1 General Theory and Concept

A pseudo-likelihood function replaces a numerically challenging joint density by a simpler function assembled from suitable factors. The method achieves computational advantages, it does not affect model interpretation. Consider a sample of size N with repeated measures sequences of length n. Define S as the set of all $2^n - 1$ vectors of length n, consisting solely of zeros and ones, with each vector having at least one non-zero entry. Denote by $\boldsymbol{Y}_i^{(s)}$ the sub-vector of $\boldsymbol{Y}_i$ corresponding to the components of s that are non-zero. The associated joint density is $f_s(\boldsymbol{y}_i^{(s)}; \boldsymbol{\theta}_i)$. To define a pseudo-likelihood function, one chooses a set $\delta = \{\delta_s | s \in S\}$ of real numbers, with at least one non-zero component. The log of the pseudo-likelihood is then defined as

$$p\ell = \sum_{i=1}^{N} \sum_{s \in S} \delta_s \ln f_s(\boldsymbol{y}_i^{(s)}; \boldsymbol{\theta}_i). \tag{2.4}$$

The classical log-likelihood function is found by setting $\delta_s = 1$ if s is the vector consisting solely of ones, and 0 otherwise.

Maximization of Eq. (2.4) can be done, subject to adequate regularity conditions, by solving the pseudo-likelihood (score) equations, which can be obtained by differentiating the logarithmic pseudo-likelihood and equating the resulting derivative to zero. Suppose that $\boldsymbol{\theta}$ is the true parameter. Under suitable regularity conditions (Arnold and Strauss, 1991; Geys, Molenberghs, & Ryan, 1999; Aerts *et al.*, 2002), it can be shown that maximizing Eq. (2.4) produces a consistent and asymptotically normal estimator $\widetilde{\boldsymbol{\theta}}$ so that $\sqrt{N}(\widetilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta})$ converges in distribution to

$$N_p \left[\mathbf{0} \, , \; I_0(\boldsymbol{\theta})^{-1} I_1(\boldsymbol{\theta}) I_0(\boldsymbol{\theta})^{-1} \right]. \tag{2.5}$$

Details on various forms of pseudo-likelihood can be found in Molenberghs and Verbeke (2005, Ch. 9, 12, 21, 22, 24, and 25).

2.3.2 Pseudo-likelihood and General Split Sample Theory

Fieuws and Verbeke (2006) and Fieuws *et al.* (2006) used pseudo-likelihood to fit mixed models to high-dimensional multivariate longitudinal data. They supplemented the standard method with an additional device by first replacing a set of M longitudinal sequences

by the $M(M-1)/2$ longitudinal pairs. This in itself is a standard application of pseudo-likelihood. They then assumed that each pair has its own parameter vector. Symbolically, this can be written as:

$$p\ell(\boldsymbol{\theta}) \equiv p\ell(\boldsymbol{y}_{1i}, \boldsymbol{y}_{2i}, \ldots, \boldsymbol{y}_{Mi}|\boldsymbol{\theta}) = \sum_{r<s} \ell(\boldsymbol{y}_{ri}, \boldsymbol{y}_{si}|\boldsymbol{\theta}_{rs}), \tag{2.6}$$

where $\boldsymbol{Y}_{ri}$ is the rth sequence for subject i. In (2.6), $\boldsymbol{\theta}$ results from stacking all $M(M-1)/2$ pair-specific parameter vectors $\boldsymbol{\theta}_{rs}$. The actual parameter vector of interest is $\boldsymbol{\theta}^*$, the set of non-redundant parameters is $\boldsymbol{\theta}$.

To obtain $\boldsymbol{\theta}^*$, Fieuws and Verbeke (2006) take averages of all available estimates for that specific parameter, implying that $\widehat{\boldsymbol{\theta}}^* = A\widehat{\boldsymbol{\theta}}$ for an appropriate linear combination matrix A. Further, combining this step with general pseudo-likelihood inference, a sandwich estimator is used:

$$\sqrt{N}(\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*) = \sqrt{N}(A\widehat{\boldsymbol{\theta}} - A\boldsymbol{\theta}) \overset{\text{approx.}}{\sim} N(\boldsymbol{0}, AI_0^{-1}I_1I_0^{-1}A'), \tag{2.7}$$

where

$$I_0(\boldsymbol{\theta}) = E\left[\frac{\partial^2 p\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'\partial\boldsymbol{\theta}}\right], \qquad I_1(\boldsymbol{\theta}) = E\left[\left(\frac{\partial p\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)'.\frac{\partial p\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right]. \tag{2.8}$$

Molenberghs, Verbeke, and Iddi (2011) took a very similar route to partition a potentially large sample into (independent) sub-samples.

Here, Molenberghs, Verbeke, and Iddi (2011) chose

$$A = \frac{1}{K}(I, \ldots, I) \tag{2.9}$$

to pass from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}^*$. This is a sensible choice in the i.i.d. setting (e.g., when all clusters in a CS model have the same size) and with the same number of subjects per sub-sample. The estimator and precision estimator then become:

$$\widehat{\boldsymbol{\theta}}^* = \frac{1}{K}\sum_{k=1}^{K}\widehat{\boldsymbol{\theta}}_k, \tag{2.10}$$

$$\text{var}(\widehat{\boldsymbol{\theta}}^*) = \frac{1}{K}H_{\widehat{\theta}}^{-1}, \tag{2.11}$$

with $H_{\widehat{\theta}}^{-1} = -I_0(\boldsymbol{\theta}_k)$. In this special case, the expected information matrices are identical. Alternatively, one can use the observed information matrices, and then use instead:

$$\frac{1}{K^2}\sum_{k=1}^{K}\mathcal{H}_{\widehat{\theta},k}^{-1}. \tag{2.12}$$

where $\mathcal{H}_{\widehat{\theta},k}$ is the observed information for sub-sample k.

In this particular case, pseudo-likelihood produces the same estimator as full likelihood. This also stems from the fact that all subjects follow the same distribution.

2.3.3 Pairwise Pseudo-likelihood for Data Missing at Random

For pseudo-likelihood to be valid under MAR, proper modifications need to be done. Precise statements are provided in Molenberghs et al. (2011). The proposed corrections to the standard form of pseudo-likelihood follow single and double robustness ideas. With single robustness, inverse probability weighting is incorporated, i.e. the inverse probability of being observed, whereas with double robustness a predictive model for the unobserved responses is added.

Molenberghs et al. (2011) presented general expressions for the estimating equations and established their validity. This result also provides an easy way to consistently estimate the asymptotic covariance. The matrix I_0 arises from evaluating the second derivative of $p\ell$ in Eq. (2.4) at the PL estimate. The expectation in I_1 can be replaced by the cross-products of the observed scores. As discussed by Arnold and Strauss (1991), the Cramèr-Rao inequality implies that $I_0^{-1}I_1I_0^{-1}$ is greater than the inverse of I (the Fisher information matrix for the maximum likelihood case), in the sense that $I_0^{-1}I_1I_0^{-1} - I^{-1}$ is positive semi-definite. Strict inequality holds if the PL estimator fails to be a function of a minimal sufficient statistic. Geys, Molenberghs, & Ryan (1999) have shown that, in realistic clustered-data settings in toxicology experiments, efficiency loss is often negligible and is certainly justified in view of computational convenience and speed.

As stated earlier, marginal models for non-Gaussian data can become computationally prohibitive when subjected to full maximum likelihood inference, especially with large within-unit replication. le Cessie & van Houwelingen (1991) and Geys, Molenberghs, and Lipsitz (1998) replace the true contribution of a vector of correlated binary data to the full likelihood, written as $f(y_{i1}, \ldots, y_{in_i})$, by the product of all pairwise contributions $f(y_{ij}, y_{ik}),\ 1 \leq j < k \leq n_i$, to obtain a pseudo-likelihood function. Also the term composite likelihood is encountered in this context, but in this thesis 'pseudo-likelihood' is used throughout. Renard, Molenberghs, and Geys (2004) refer to this particular instance of pseudo-likelihood as *pairwise likelihood*. The contribution of the ith subject or cluster to the log pseudo-likelihood then specializes to

$$p\ell_i = \sum_{j<k} \ln f(y_{ij}, y_{ik}), \tag{2.13}$$

if it contains more than one observation. Otherwise, $p\ell_i = f(y_{i1})$. Extension to three-way and higher-order pseudo-likelihood is straightforward, all of which are special cases of Eq. (2.4).

Table 2.1: *Contingency table for stratum i ($i = 1, \ldots, N$)*

	Exposure +	**Exposure -**	Total
Case +	a_i	b_i	n_{1i}
Control -	c_i	d_i	n_{2i}
Total	m_{1i}	m_{2i}	n_i

2.4 Mantel-Haenszel Estimator

The Mantel-Haenszel estimator (Mantel and Haenszel, 1959, MH) can be used in various contexts. It serves as estimator for the odds ratio in a series of 2×2 tables or in matched designs (Agresti, 2002, pp. 231). The simplest form of matching is 1:1, i.e., one control per case. This is effective when both are sufficiently prevalent in the population. Often, cases are more scarce than controls, e.g., with rare diseases. It then makes sense to select more controls per case. When a case is individually matched to a set of controls, having similar values for some confounding variables, the most extreme form of stratified design is created. Each case and corresponding control(s) can be seen as one stratum. In general, a setting with stratified 2×2 tables already exists when subpopulations within the overall population vary. In the case of several, N say, 2×2 tables or strata, the i^{th} $(i = 1, \ldots, N)$ stratum takes the form as presented in Table 2.1. The overall sample size is defined as $n = \sum_{i=1}^{N} n_i$.

The MH is a useful and convenient estimator for obtaining a common odds ratio, when there are one or more confounders. The MH is not fully parametric and typically strata are of varying sizes, naturally necessitating the use of weights. Mantel and Haenszel (1959) proposed several weighting schemes to estimate a common odds ratio. It is important to realize that these weighting schemes, while very effective, do not follow from formal optimality criteria. The most widely accepted estimator of the common odds ratio is:

$$\tilde{\psi}_{MH} = \frac{\sum_{i=1}^{N} \frac{a_i d_i}{n_i}}{\sum_{i=1}^{N} \frac{b_i c_i}{n_i}} = \frac{\sum_{i=1}^{N} w_i \frac{a_i d_i}{b_i c_i}}{\sum_{i=1}^{N} w_i}, \tag{2.14}$$

with $w_i = \frac{b_i c_i}{n_i}$.

The strata do not need to be of the same size and even if some cell counts are small or even zero, the estimator remains well-defined, an important asset. Also, when $b_i c_i$ equals zero a stratum is omitted in the calculation of the common odds ratio as the weight becomes zero as well. This estimator is very practical to use.

Interestingly, no expression for the variance was available. With time, others investigated the estimator and many extensions have been developed. Kuritz *et al.* (1988)

reviewed the MH and its variance formulas. At first, Hauck (1979) proposed an estimator for the variance using a product binomial model, appropriate for large stratum samples. Woolf (1955) used a logarithmic transformation of the odds ratio estimator, which makes the sampling variance simple and easy to use as weights. From another point of view, Flanders (1985) proposed a variance estimator based on a series of Monte Carlo experiments, leading to more accurate confidence intervals. Robins *et al.* (1986a,b) proposed a new robust variance estimator based on the unconditional distribution of the data. These last two are very similar and even identical for matched designs. Either is applicable in both sparse data and large-strata limiting models and are easily computed. None of these estimators had been formally shown to be "best", but the latter are preferred. Nowadays the variance formula of Robins *et al.* (1986b) is commonly used, particularly in statistical software. It will be used later in this thesis and takes the following form:

$$\begin{aligned} v_R &= \mathrm{var}(\log \hat{\psi}_{MH}) \\ &= \frac{\sum_{i=1}^{N} \frac{(a_i+d_i)a_i d_i}{n_i^2}}{2\left(\sum_{i=1}^{N} \frac{a_i d_i}{n_i}\right)^2} + \frac{\sum_{i=1}^{N} \frac{(a_i+d_i)b_i c_i+(b_i+c_i)a_i d_i}{n_i^2}}{2\left(\sum_{i=1}^{N} \frac{a_i d_i}{n_i}\right)\left(\sum_{i=1}^{N} \frac{b_i c_i}{n_i}\right)} \\ &+ \frac{\sum_{i=1}^{N} \frac{(b_i+c_i)b_i c_i}{n_i^2}}{2\left(\sum_{i=1}^{N} \frac{b_i c_i}{n_i}\right)^2}. \end{aligned} \tag{2.15}$$

Chapter 3

Case Studies

This chapter introduces the four data sets used to illustrate the methodology developed in Part II. All of these data sets have an unbalanced hierarchical structure with continuous or binary responses.

3.1 Developmental Toxicity Study Sets

These data sets were set up by the Research Triangle Institute under contract to the National Toxicology Program of the U.S.A. (NTP data). 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 3.1. 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.

3.2 Clinical Trials in Schizophrenia

These data were collected from five double-blind randomized clinical trials to compare the effects of different 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

Table 3.1: *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

each subject over time. The number of patients ranges from 9 to 128 per country with a total of 2039. The positive and negative syndrome scale (PANSS) was used to asses the global condition of a patient. This scale is constructed from 30 items, each taking values between 1 and 7, giving an overall range of 30 to 210. PANSS provides an operationalized, drug-sensitive instrument, which is useful for both typological and dimensional assessment of schizophrenia. Depending on the trial, treatment was administered for a duration of 48 weeks with at most 12 monthly measurements. For analysis we included patients with at least one follow-up measurement. Table 3.2 shows the number of patients participating in each trial for all different time patterns in receiving the treatments. Because not all subjects received treatment at the same time points and, not the same amount, there are 26 different time patterns, therefore, the final dataset is unbalanced.

3.3 Intego: Large General Practice Dataset

Intego is a Belgian general practice-based morbidity registration network at the Department of General Practice of the University of Leuven, Belgium. They built a large database as a result of continual recording of data in general practices since 1994. It holds over 4 million diagnoses, 44 million laboratory results and 17 million medication prescriptions and 700,000 vaccination data.

Intego procedures were approved by the ethical review board of the Medical School of the University of Leuven (ML 1723) and by the Belgian Privacy Commission (SC-SZG/13/079). Many general practices applied for inclusion in the registry. Before approval, the registration performance was checked using algorithms between all participants. Only those with an optimal performance were included. All participating general practices need to routinely record all new diagnoses, drug prescriptions, laboratory results and patient information. They use universal codes; diagnoses are classified using ICPC 2 codes (International Classification of Primary Care) and the WHO's Anatomical Therapeutic Chemical (ATC) classification system for drugs. See also Truyers *et al.* (2014) and the Intego website (http://www.intego.be).

Table 3.2: *PANSS data. Number of clusters in each trial for each cluster pattern. The pattern consists of the numbers representing the months after starting point for which a PANSS score is available.*

		Trial					
n	Pattern	FIN-1	FRA-3	INT-2	INT-3	INT-7	Total
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
	$(0,1,2,6)$	0	0	3	0	0	3
	$(0,1,2,3)$	0	4	1	0	0	5
	$(0,1,3,6)$	0	1	0	0	0	1
	$(0,2,6,8)$	0	0	0	0	1	1
5	$(0,1,2,4,6)$	58	0	85	35	6	184
	$(0,1,2,4,8)$	0	0	8	0	1	9
	$(0,1,4,6,8)$	0	0	6	0	0	6
	$(0,1,2,6,8)$	0	0	8	0	0	8
	$(0,2,4,6,8)$	0	0	3	0	2	5
	$(0,2,4,8,12)$	0	0	1	0	0	1
	$(0,1,2,3,4)$	0	44	0	0	0	44
	$(0,1,3,4,5)$	0	1	0	0	0	1
6	$(0,1,2,4,6,8)$	0	0	986	240	74	1300
	$(0,1,4,6,8,10)$	0	0	1	0	0	1
	$(0,1,2,6,8,12)$	0	0	1	0	0	1
	$(0,1,2,4,6,10)$	0	0	1	0	0	1
	$(0,1,2,4,5,6)$	0	0	2	0	0	2

Table 3.3: *The Analgesic Trial. Absolute and relative frequencies of the five GSA categories for each of the four follow-up times.*

GSA	Month 3		Month 6		Month 9		Month 12	
1	55	14.3%	38	12.6%	40	17.6%	30	13.5%
2	112	29.1%	84	27.8%	67	29.5%	66	29.6%
3	151	39.2%	115	38.1%	76	33.5%	97	43.5%
4	52	13.5%	51	16.9%	33	14.5%	27	12.1%
5	15	3.9%	14	4.6%	11	4.9%	3	1.4%
Total	385		302		227		223	

3.4 The Analgesic Trial

The analgesic trial was a single-arm clinical trial involving 395 patients who were given analgesic treatment for pain caused by chronic non-malignant disease. Treatment was to be administered for 12 months and assessed by means of a five-point 'Global Satisfaction Assessment' (GSA) scale: (1) very good; (2) good; (3) indifferent; (4) bad; (5) very bad. As it is frequently of interest to physicians to classify a patient's status as either improving or worsening, some analyses have considered a dichotomized version, GSABIN, which is 1 if GSA ≤ 3 and 0 otherwise; this outcome will be adopted for the analysis as well. Apart from the outcome of interest, a number of covariates are available, such as age, sex, weight, duration of pain in years prior to the start of the study, type of pain, physical functioning, psychiatric condition, respiratory problems, etc.

GSA was rated by each person four times during the trial: at months 3, 6, 9, and 12. An overview of the frequencies per follow-up time is given in Table 3.3. Inspection of Table 3.3 reveals varying totals per column, due to missingness. At three months, 10 subjects lack a measure, with these numbers being 93, 168, and 172 at subsequent times.

An overview of the extent of missingness (Table 3.4) indicates that only around 40% of the subjects have a complete data sequence. Both dropout and intermittent patterns of missingness occur – the former amounting to roughly 40%, with less than 20% for the latter.

Table 3.4: *The Analgesic Trial. Overview of missingness patterns and the frequencies with which they occur. 'O' indicates observed and 'M' indicates missing.*

	Measurement Occasion				N	%
	Month 3	Month 6	Month 9	Month 12		
Completers	O	O	O	O	163	41.2
Dropouts	O	O	O	M	51	12.91
	O	O	M	M	51	12.91
	O	M	M	M	63	15.95
Non-Monotone Missingness	O	O	M	O	30	7.59
	O	M	O	O	7	1.77
	O	M	O	M	2	0.51
	O	M	M	O	18	4.56
	M	O	O	O	2	0.51
	M	O	O	M	1	0.25
	M	O	M	O	1	0.25
	M	O	M	M	3	0.76

PART II

Contributions

Chapter 4

A Characterization of Incompleteness

4.1 Introduction

One consequence of unequal sample sizes is that *complete* sufficient statistics may no longer exist. Completeness implies that any function of a sufficient statistic that has zero expectation for every value of the parameter indexing the parametric model class, is the zero function almost everywhere, definition see Section 2.2.1. The relevance of complete sufficient statistics has been established through two theorems, Lehman-Scheffé (Casella and Berger, 2001) and Basu's (Basu, 1955) theorem. Completeness, combined with regularity conditions, provides a basis for estimators with desirable properties, such as unbiasedness and optimality.

In sequential designs (Wald, 1945) one incorporates a data-driven rule to potentially stop the trial before reaching the maximal sample size. Such methods are well established in clinical trials (Armitage, 1975). While the statistical aspects of sequential methods have been carefully studied (Lehman and Stein, 1950), the lack of completeness has led to disagreement and confusion, regarding appropriate (point and interval) estimation following such trials, leading to many *ad hoc* proposals. Liu and Hall (1999) and Liu *et al.* (2006), building upon Emerson and Fleming (1990), explored this aspect. Molenberghs *et al.* (2014) and Milanzi *et al.* (2016, 2015) studied the issue in a wider framework, encompassing stochastic stopping rules and completely random sample sizes. They demonstrated that, somewhat contrary to intuition and in spite of incompleteness, the ordinary sample average remains a viable estimator (because of consistency, asymptotic normality, and

high efficiency), even though it no longer has all properties that it enjoys in the conventional, fixed sample size setting. We elaborate on this in Section 4.2. Another setting without complete sufficient statistics is that of clusters of unequal size. Such designs include longitudinal, multilevel, spatial, and multi-stage survey designs. A counterexample is a longitudinal study where each one of the subjects is measured exactly the same number of times, at an *a priori* fixed set of measurement occasions. Then, N, the number of subjects, and n, the number of measurements, are design constants. However, such "clean" designs are the exception rather than the norm. A variety of *ad hoc* methods has been proposed for the random cluster size setting. Other settings without complete sufficient statistics are missing data, censored time-to-event data, random visit times, and joint modeling of longitudinal and time-to-event data.

To ensure completeness of the minimal sufficient statistics, Lehmann (1981, pp. 142–143), Brown (1986, pp. 42–44) and Boos and Stefanski (2013, pp. 103–104) formulated theorems, based on appropriate restrictions placed on the canonical form of the exponential family. Brown (1986) proves incompleteness using complex analytic properties and refers to the unique determination of a standard family by its Laplace transform. Here, however, the latter is more explicitly used and a result, both general and easy to use, follows. Boos and Stefanski (2013) and Lehmann (1981) base their theorems on the fact that the family is minimal and the parameter space contains a rectangle, thereby requiring that the family is of full rank. The characterization of incompleteness given here is also related to a property of curved exponential models (Van Garderen, 1997; Keener, 2010). These have the property that the dimension of the minimal sufficient statistic is larger than the number of parameters in the model. Van Garderen (1997) establishes a theorem that allows a straightforward comparison between the dimension of the minimal sufficient statistic and the number of parameters to determine when a model is a curved exponential model. Keener (2010) points out that curved exponential models arise naturally with data from sequential experiments and in applications to contingency table analysis.

In this chapter a general criterion for completeness is formulated that starts from, but moves beyond, the length of a vector. In Section 4.2, two commonly encountered settings are presented, where minimal sufficient statistics are incomplete. Known results leading up to the characterization of complete sufficient statistics are briefly reviewed in Section 4.3. The key result is presented in Section 4.4. To highlight the ease of use of the criterion, it is applied and shown to work for two more complex data settings, i.e., clusters of random size and missing data. Section 4.5 illustrates and further clarifies the findings for clustered data. Section 4.6 considers partially unobserved contingency tables, extends these results to other missing-data settings and shows why seemingly unrelated settings, all have led to incomplete sufficient statistics.

4.2 Motivating Settings

4.2.1 Sequential Trials

Group sequential trials are in common use and have been well studied (e.g., Wald, 1945; Armitage, 1975; Whitehead, 1997; Jennison and Turnbull, 2000). The corresponding design and hypothesis testing machinery is well developed. Nevertheless, issues still surround estimation following a sequential trial (Siegmund, 1978; Hughes and Pocock, 1988; Todd, Whitehead, and Facey, 1996; Whitehead, 1999). Several authors have reported that standard estimators such as the sample average are biased. In response to this, various proposals have been made to remove or alleviate this bias and its consequences (Tsiatis, Rosner, and Mehta, 1984; Rosner and Tsiatis, 1988; Emerson and Fleming, 1990). An early suggestion was to use an estimator (Blackwell, 1947) that conditions on the stopping event.

The origin of the problem was understood at an early stage of the development. Lehman and Stein (1950) showed that it originates from *incompleteness* of the sufficient statistics, generally implying the non-existence of a minimum variance unbiased linear estimator. Liu and Hall (1999) and Liu *et al.* (2006) explored this incompleteness in group sequential trials, and Molenberghs *et al.* (2014) and Milanzi *et al.* (2016, 2015) embedded the problem in the broader class of random sample size, which also includes, missing data, completely random sample sizes, censored time-to-event data, and random cluster sizes. Their main findings were: (1) the sample average, although asymptotically unbiased has finite sample bias; (2) apart from the exponential distribution setting, there is no finite-sample optimal linear estimator, although the sample average is asymptotically optimal (i.e., uniform minimum variance unbiased); (3) the validity (i.e., consistency and asymptotic normality) of the sample average also follows from standard ignorable likelihood theory (Little and Rubin, 2002); we will return to ignorability in Section 4.6; (4) there exists a maximum likelihood estimator that conditions on the realized sample size, which is finite sample unbiased, but has slightly larger variance and mean square error.

4.2.2 Clusters of Unequal Size

Even when the cluster size contains no information about the scientific parameters, there are issues resulting from lack of a complete sufficient statistic. One family of approaches is based on restricted moment estimators obtained through the use of generalized estimating equations (Liang and Zeger, 1986; Liang, Zeger, and Qaqish, 1992). Pseudo-likelihood, or composite likelihood, estimators have also been proposed (Lindsay, 1988; Arnold and

Strauss, 1991; le Cessie and van Houwelingen, 1994; Geys, Molenberghs, and Lipsitz, 1998; Aerts *et al.*, 2002). In these, the full likelihood is simplified and replaced by a more manageable function (Geys, Molenberghs, and Lipsitz, 1998). Various authors have studied weighted and unweighted approaches, in contrast to (non-) informative cluster sizes (Williamson, Datta, and Satten, 2003; Benhin, Rao, and Scott, 2005; Hofman, Sen, and Weinberg, 2001; Cong, Yin, and Shen, 2007; Chiang and Lee, 2008; Wang, Kong, and Datta, 2011).

4.3 (In)complete Sufficient Statistics and Some Known Results

The property of central interest is that of *completeness* (Casella and Berger, 2001, pp. 285–286). The relevance of completeness rests on two follow-up theorems. First, the Lehman-Scheffé theorem (Casella and Berger, 2001) states that, if a statistic is unbiased, complete, and sufficient for a parameter θ, then it corresponds to the best mean-unbiased estimator for θ. Second, the connection with ancillarity follows from Basu's theorem (Basu, 1955): a statistic that is both bounded complete and sufficient is independent of any ancillary statistic (Casella and Berger, 2001, p. 287). The theorems are implications rather than equivalences. For example, in the sequential trial context there exist estimators with very good properties, despite lack of completeness (Molenberghs *et al.*, 2014).

Liu and Hall (1999) established incompleteness of the sufficient statistic for a clinical trial with a stopping rule, for the case of normally distributed outcomes. 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 stopping rule, encompassing the important case of a completely random sample size. In the latter case, even though sample size and data are unrelated, completeness no longer holds.

Tables 4.1 and 4.2 contain a number of illustrative examples where the sufficient statistics are found to be (in)complete. In Table 4.1, continuous and categorical outcomes are considered. Positive outcomes (continuous times and counts) are the subject of Table 4.2. Some of these models are based upon Chakraborty (2015). Precise formulations and details can be found in Appendix A.1. Examples 1 and 2, a univariate sample with either known or unknown variance, have complete sufficient statistics. Example 3, a univariate normal sample with coupled mean and variance, does not; here, unlike in the previous examples, the sufficient statistic is of higher dimension than the parameter. When the mean-variance coupling parameter τ^2 is unknown (Example 3a), the sufficient statistic and the parameter are again of the same dimension and completeness holds, unlike when τ^2 is known (Example 3b). If the statistic is restricted to either the sample sum or the

sum of squared sample units, then it is no longer sufficient. This last situation occurs also in Example 4, a sequential trial, where the sufficient statistic consist not only of the data collected, but also of the sample size realized, i.e., a one-dimensional parameter needs a two-dimensional sufficient statistic. These developments emphasize that the establishment of either completeness or its converse requires tedious, situation-specific calculations when using the definition. It is therefore convenient to derive a simple criterion based on the dimensions of the parameter vector and the sufficient statistic, to be established next.

4.4 A Characterization of Incompleteness

We turn to a general characterization of incompleteness, in the exponential family with a vector-valued parameter and minimal sufficient statistic. Group the outcomes Y_i into a vector $\boldsymbol{Y}$, with vector-valued parameter $\boldsymbol{\theta}$ and write the exponential family in the form

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \tilde{h}(\boldsymbol{y}) \exp\left\{\boldsymbol{\eta}(\boldsymbol{\theta})'\boldsymbol{k}(\boldsymbol{y}) - A(\boldsymbol{\theta})\right\}, \tag{4.1}$$

where the sufficient statistic $\boldsymbol{K} \equiv \boldsymbol{K}(\boldsymbol{Y})$. Consider first the situation where the function $\boldsymbol{\eta}$ is everywhere of full rank. Examples 1 and 2 fall into this category. Because $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are in 1-to-1 relationship, we can use $\boldsymbol{\theta}$, without loss of generality. The score equation corresponding to (4.1) is $S(\boldsymbol{\theta}) = \partial\boldsymbol{\eta}/\partial\boldsymbol{\theta} \cdot \boldsymbol{K} - \partial A/\partial\boldsymbol{\theta} = 0$. If the transformation, $\boldsymbol{\eta}(\boldsymbol{\theta})$, is of full rank, then it follows that

$$\boldsymbol{K} = \left(\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\theta}}\right)^{-1} \frac{\partial A}{\partial\boldsymbol{\theta}}. \tag{4.2}$$

Taking expectations, the right hand side of Equation (4.2) equals $E(\boldsymbol{K})$. In the above situation, the sufficient statistic is complete. To see this, assume that there is a function $g(\boldsymbol{k})$ with expectation zero for all values of $\boldsymbol{\theta}$. It then satisfies

$$\int g(\boldsymbol{k})h(\boldsymbol{k}) \exp\left\{\boldsymbol{\theta}'\boldsymbol{k} - A(\boldsymbol{\theta})\right\} d\boldsymbol{k} = 0, \tag{4.3}$$

with obvious notation, similar to Equation (4.1) but $\tilde{h}$ expressed as function of $\boldsymbol{k}$ rather than $\boldsymbol{y}$. Applying Fubini's theorem (Rudin, 1974), we can write Equation (4.3) as

$$0 = \int dk_p h(k_p) e^{\theta_p k_p} \int dk_{p-1} h(k_{p-1}|k_p) e^{\theta_{p-1} k_{p-1}} \ldots$$

$$\int dk_1 g(k_1, \ldots, k_p) h(k_1|k_2, \ldots, k_p) e^{\theta_1 k_1}.$$

Table 4.1: *Examples with complete and incomplete sufficient statistics (continuous and categorical outcomes)*

Ex.	Setting	Parameter(s)	Sufficient statistic(s)
	Settings with complete sufficient statistics		
1	$Y_i \sim N(\mu, \sigma^2), i = 1, \ldots, n$ with μ unknown and σ^2 known	μ	K_1
2	$Y_i \sim N(\mu, \sigma^2), i = 1, \ldots, n$ with μ and σ^2 unknown	(μ, σ^2)	(K_1, K_2)
3a	$Y_i \sim N(\mu, \tau^2\mu^2), i = 1, \ldots, n$ with μ and τ^2 unknown	(μ, τ^2)	(K_1, K_2)
6	$Y_i \sim N(\mu, \mu), i = 1, \ldots, n$	μ	K_2
7a	$Y_i \sim N(\mu, \mu^{2\lambda}), i = 1, \ldots, n$ and $\lambda = 0$ or $1/2$	μ	K_1 or K_2
8	$M_1 \times M_2$ contingency table with $\varphi(k_1 \mid k_2)$ and $\pi(k_2)$	$\varphi(k_1\|k_2), \pi(k_2)$	
15	Fully observed 2×1 contingency table	p	Z_{21}
	Settings with incomplete sufficient statistics		
3b	$Y_i \sim N(\mu, \tau^2\mu^2), i = 1, \ldots, n$ with μ unknown and τ^2 known	μ	(K_1, K_2)
4	Sequential trial with stochastic stopping rule	μ	(K_3, N)
5	Bivariate parameter, one of which known (cf. Ex. 2)	μ	(K_1, K_2)
7b	$Y_i \sim N(\mu, \mu^{2\lambda}), i = 1, \ldots, n$ and $\lambda \neq 0$ and $\neq 1/2$	μ	K_1, K_2
9	$Y_i \sim N(\mu, 1)$, sample size N, $1 \leq N \leq n$ with π_N	μ	(K_3, N)
10	$\boldsymbol{Y} \sim N(\mu 1_N, \sigma^2 I_N + \tau^2 J_N)$	(μ, σ^2, τ^2)	(K_3, K_4, K_5, N)
11	Vector-valued data and parameter, with completely random sample size	$\pi(N\|k)$	(K_3, N)
12	N clusters of completely random size		$[\boldsymbol{K} = \boldsymbol{K}\{(\boldsymbol{Y}_i)\}; \boldsymbol{N} = \boldsymbol{N}\{(N_i)\}]$
13	$\boldsymbol{Y}_i \sim N(\mu 1_{N_i}, \sigma^2 I_{N_i} + \tau^2 J_{N_i}), i = 1, \ldots, N$	(μ, σ^2, τ^2)	$(S_{1\ell}, S_2, S_{3\ell}, S_{4\ell})$
14	General clustered-data setting with random cluster sizes	θ	
16	Partially missing 2×1 contingency table	p	(Z_{21}, Z_1)
17	Partially missing 2×1 contingency table	p_{jk}	(Z_{2jk}, Z_{1j})

$K_1 = \sum_{i=1}^{n} Y_i$; $K_2 = \sum_{i=1}^{n} Y_i^2$; $K_3 = \sum_{i=1}^{N} Y_i$; $K_4 = \boldsymbol{Y}'\boldsymbol{Y}$; $K_5 = \boldsymbol{Y}' J_N \boldsymbol{Y}$.

$S_{1\ell} = \sum_{i=1}^{c_\ell} \sum_{j-1}^{n_\ell} Y_{ij}^{(\ell)}$; $S_2 = \sum_{\ell=1}^{L} \sum_{i=1}^{c_\ell} \sum_{j-1}^{n_\ell} \left(Y_{ij}^{(\ell)}\right)^2$;

$S_{3\ell} = \sum_{i=1}^{c_\ell} \left(\sum_{j-1}^{n_\ell} Y_{ij}^{(\ell)}\right)^2$; $S_{4\ell} = c_\ell$.

Table 4.2: *Examples with complete and incomplete sufficient statistics (outcomes on* $[0, +\infty[$*)*

Ex.	Setting	Parameter(s)	Sufficient statistic(s)
	Settings with complete sufficient statistics		
18	$Y_i \sim$ Poisson(λ)	λ	K_1
19	$Y_i \sim$ Exponential(λ)	λ	K_1
20	$Y_i \sim$ Integrated Exponential(λ)	λ	K_1
	Settings with incomplete sufficient statistics		
21	$Y_i \sim$ Integrated Weibull(λ, ρ)	(λ, ρ)	$Y_1, \ldots, Y_n$

$K_1 = \sum_{i=1}^{n} Y_i$.

This leads to a telescopic series of Laplace transforms:

$$F_{\theta_1}(k_2, \ldots, k_p) = \mathcal{L}_{\theta_1}\left\{g(k_1, \ldots, k_p)h(k_1|k_2, \ldots, k_p)\right\}, \tag{4.4}$$

$$F_{\theta_2}(k_3, \ldots, k_p) = \mathcal{L}_{\theta_2}\left\{F_{\theta_1}(k_2, \ldots, k_p)h(k_2|k_3, \ldots, k_p)\right\}, \tag{4.5}$$

$$\vdots$$

$$F_{\theta_p} = \mathcal{L}_{\theta_p}\left\{F_{\theta_{p-1}}(k_p)h(k_p)\right\} = 0, \tag{4.6}$$

with obvious notation. Moving step-by-step from Equations (4.6) to (4.4), the sequence of F_{θ_j} is zero a.e. for j running down from $p-1$ to 1 and then eventually $g(k_1, \ldots, k_p) = 0$ a.e., establishing completeness. Looking to the bivariate Examples 2 and 5, the same is seen for Example 2, but a different result occurs for Example 5 where one of the two parameters is known. Then, the sufficient statistic is incomplete.

Proposition 4.1. *(Characterization of a complete sufficient statistic.) Provided the parameter space is rectangular, a sufficient statistic* $\boldsymbol{k}$ *is complete for a parameter* $\boldsymbol{\theta}$ *in a exponential family model if and only if* $\boldsymbol{\theta}$ *cannot be transformed to a parameterization* $\boldsymbol{\eta}$ *with a proper subset* $\boldsymbol{\eta}_1$ *such that* $\boldsymbol{\eta} = (\boldsymbol{\eta}_1', \boldsymbol{\eta}_2(\boldsymbol{\eta}_1)')'$.

Proof. The proof is based upon a more general version of Example 5.

Let $\boldsymbol{\eta}(\boldsymbol{\theta})$ be a function to match the minimal sufficient statistic $\boldsymbol{K}$ that is not of full rank. This can be decomposed, using the implicit function theorem, assuming the functions involved are continuously differentiable, and then mapped as follows:

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2(\boldsymbol{\eta}_1) \end{pmatrix} \qquad \longleftrightarrow \qquad \boldsymbol{K} = \begin{pmatrix} \boldsymbol{K}_1 \\ \boldsymbol{K}_2 \end{pmatrix}. \tag{4.7}$$

Incompleteness of the sufficient statistic would follow if a function $g(\boldsymbol{k}_1, \boldsymbol{k}_2)$ could be

found satisfying:

$$0 = \int d\boldsymbol{k}_2 h_2(\boldsymbol{k}_2) e^{\boldsymbol{\eta}_2(\boldsymbol{\eta}_1)'\boldsymbol{k}_2} \int d\boldsymbol{k}_1 h_1(\boldsymbol{k}_1|\boldsymbol{k}_2) g(\boldsymbol{k}_1, \boldsymbol{k}_2) e^{\boldsymbol{\eta}_1 \boldsymbol{k}_1}, \tag{4.8}$$

$$0 = \int d\boldsymbol{k}_2 h_2(\boldsymbol{k}_2) e^{\boldsymbol{\eta}_2(\boldsymbol{\eta}_1)'\boldsymbol{k}_2} \mathcal{L}_{\boldsymbol{\eta}_1} \left\{ h_1(\boldsymbol{k}_1|\boldsymbol{k}_2) g(\boldsymbol{k}_1, \boldsymbol{k}_2) \right\}, \tag{4.9}$$

$$0 = \int h_2(\boldsymbol{k}_2) F(\boldsymbol{k}_2, \boldsymbol{\eta}_1) e^{\boldsymbol{\eta}_2(\boldsymbol{\eta}_1)'\boldsymbol{k}_2} d\boldsymbol{k}_2. \tag{4.10}$$

Now, (4.10) is not a Laplace transform. Therefore, we can choose a function $F(\boldsymbol{k}_2, \boldsymbol{\eta}_1)$ that satisfies the equation and then use the inverse Laplace transform to derive $g(\boldsymbol{k}_1, \boldsymbol{k}_2)$. To see that such a function can easily be found, choose:

$$F(\boldsymbol{k}_2, \boldsymbol{\eta}_1) = e^{-\boldsymbol{\eta}_2(\boldsymbol{\eta}_1)'\boldsymbol{k}_2} \widetilde{F}(\boldsymbol{k}_2).$$

With this choice, condition (4.10) simplifies to:

$$0 = \int h_2(\boldsymbol{k}_2) \widetilde{F}(\boldsymbol{k}_2) d\boldsymbol{k}_2.$$

In other words, we need a function $\widetilde{F}(\boldsymbol{k}_2) \perp h_2(\boldsymbol{k}_2)$. □

Notice the similarity between the characterization and earlier work of Lehmann (1981), Brown (1986) and Boos and Stefanski (2013). However, our characterization leads to a more general result and an easy to use criterion. Also Van Garderen (1997) and Keener (2010) have already pointed out this relationship between the dimension of the sufficient statistic and the number of parameters for curved exponential models. Evidently, their focus is different from ours. With this characterization all examples of Tables 1 and 2 can be verified solely by counting the dimensions of the parameter vectors and sufficient statistic. The proposition explains why Examples 1 and 2 have complete sufficient statistics. This is trivial in Example 1 because the parameter and sufficient statistic are scalar. In Example 2, the parameter $\boldsymbol{\theta} = (\mu, \sigma^2)'$ consists of two functionally independent components. Example 3 has a bivariate sufficient statistic, like Example 2, and a bivariate parameter $\boldsymbol{\theta} = (\mu, \tau^2\mu^2)'$. Write $\eta_1 = \mu$ and $\eta_2(\eta_1) = \tau^2\eta_1^2$, which explains why this is an incomplete case when τ^2 is known. For Example 4, consider two sample sizes n and $2n$. The minimal sufficient statistic is (K_3, N), and both are governed by a distribution with sole parameter μ, trivially establishing incompleteness. This result relates to Shao (1999, p. 110). In their Proposition 2.1, they consider the exponential family case, of full rank, for a sufficient statistic that is complete and sufficient. Their proof is in terms of the positive and negative parts of the normalizing function, rather than Laplace transforms.

Corollary 4.1. *(Non-linearity of the function $\boldsymbol{\eta}_2(\boldsymbol{\eta}_1)$.) For complete minimal sufficient statistics, the function $\boldsymbol{\eta}_2(\boldsymbol{\eta}_1)$ cannot be linear.*

Proof. To see this, assume there is such a linear function. The correspondence becomes:

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\eta}_1 \\ L\boldsymbol{\eta}_1 \end{pmatrix} \qquad \longleftrightarrow \qquad \boldsymbol{K} = \begin{pmatrix} \boldsymbol{K}_1 \\ \boldsymbol{K}_2 \end{pmatrix}.$$

The inner product of these vectors is $\boldsymbol{\eta}_1' \{\boldsymbol{K}_1 + L'\boldsymbol{K}_2\}$, implying that $\boldsymbol{K}_1 + L'\boldsymbol{K}_2$ is a minimal sufficient statistic of a smaller dimension, which is impossible. This establishes that $\boldsymbol{\eta}_2(\boldsymbol{\eta}_1)$ must be non-linear. □

This corollary explains why in Example 3 the parameter is $\boldsymbol{\theta} = (\mu, \tau^2\mu^2)'$ and not, for example $(\mu, \mu)'$. The latter case is studied in Example 6, a univariate normal sample with mean equal to the variance. The convenience of Proposition 4.1 is seen by generalizing Examples 3 and 6 to Examples 7 and 8, respectively. The first one is a univariate normal with coupled mean and variance, for which it is very difficult to find a function $g(k)$ that establishes incompleteness, with the use of the criterion is straightforward. Example 8 is a 2×2 contingency table with unconstrained parameterisation, leading to a complete sufficient statistic. Similar logic will be used in the next section, to illustrate a simple yet generic clustered-data setting. Details can be found in Appendix A.

4.5 Illustration: Clusters Following a Compound-symmetry Model

First, consider univariate outcomes with random sample size. Example 9 is a univariate normal with unknown mean, unit variance, and random sample size. The sufficient statistic is then incomplete. In Example 10 this is extended to normal compound-symmetry vectors, where incompleteness evidently also applies. In Example 11 the same is seen to be true for the entire exponential family. In Example 12 the clusters are further allowed to be of variable size. The sufficient statistic is then still incomplete. In Example 13, this general result is applied to normal compound-symmetry data with clusters of unequal size, not allowing for a complete sufficient statistic. In Example 14, we allow for both random cluster sizes and a general exponential model formulation.

The use of Proposition 4.1 is trivial in this context. There are three model parameters, $\boldsymbol{\theta} = (\mu, \sigma^2, \tau^2)'$, but the sufficient statistic is necessarily of higher dimension, as soon as there are at least two different cluster sizes. This route is easier than the explicit construction of a function (B.1). Even though this was still practicable, the computations for Example 4 are much more complex. This is because in Example 4 the stopping rule depends on the data, in contrast to in our most recent Examples 9–13, where the cluster sizes are completely random.

In summary, because $\boldsymbol{\eta}(\boldsymbol{\theta})$ will generally be such that the dimension of $\boldsymbol{\eta}$ is higher than that of $\boldsymbol{\theta}$, Proposition 4.1 applies. The qualification 'generally' is needed, because

there are obvious (trivial) counterexamples. In Example 13 the sufficient statistic for σ^2 is one-dimensional, as an exception to the rule. When this would hold for all parameters, then completeness would hold. Such an example may be difficult to construct though. Another situation is when the cluster members are independent. Then the cluster sizes become irrelevant. If, in this special case, further $\sum_i N_i$, the overall sample size, would be constant, then such a clustered-data example reduces to a conventional univariate sample with fixed sample size, and completeness follows, establishing a counterexample. Apart from such pathological cases, virtually all practically relevant clustered data applications have incomplete minimal sufficient statistics.

4.6 Missing Data in Contingency Tables and Beyond

First consider the simple yet generic setting of missing data in contingency tables. We then turn to general missing data settings and end this section by bringing out commonality between seemingly disparate settings, considered earlier in this chapter, that all lead to incomplete sufficient statistics.

In Example 15 a fully observed 2×1 contingency table is considered, which allows for a complete sufficient statistic. When data are partially missing (Example 16), this is no longer true. Example 16 and function (A.30) are reminiscent of Example 3, where function (A.7) exists because τ^2 is known. In spite of the similarity, there is an important difference as well: q is an *unknown* constant that nevertheless does not need to be estimated, because of ignorability. Admittedly, Examples 15–16 are very simple and therefore it is hard to see the generality of the result. Thus consider a 2×2 contingency table with supplemental margins as well (Example 17), then no complete sufficient statistic exists.

In the above examples, there is nothing particular about the use of contingency tables, nor about the parameterization used for the counts, leading to the following proposition.

Proposition 4.2. *(Incomplete sufficient statistics with ignorable likelihood.) Let an exponential family model $f(\boldsymbol{Y}|\boldsymbol{\theta})$ admit a complete sufficient statistic when data are fully observed, then the same model does not admit a complete sufficient statistic under ignorable likelihood when data are partially missing.*

It is interesting to reflect upon the nature of this result. When data are partially missing, the data are effectively stratified, with one stratum grouping the fully observed trials and the other stratum the remaining trials. Still, the parameters of p-type (Example 16–17) describe both strata simultaneously. Because of ignorability, it is sensible to formulate a model where the parameter vector is of the same length as it would be when data were complete, but the stratification nevertheless implies that the length of the vector of sufficient statistics increases. This leads to the conclusion that this same

phenomenon also occurs in other settings, including many non-missing-data settings. In Example 13, the strata are defined by the different cluster sizes occurring in the data. In that example, completeness could be restored by assuming that for every one of the cluster sizes n_ℓ occurring, there is a separate parameter vector $(\mu_l, \sigma_\ell^2, \tau_\ell^2)$, together with a multinomial vector $(\pi_1, \ldots, \pi_L)$ describing the probabilities with which the various cluster sizes occur.

Other examples can now be reconsidered. In Example 9, completeness would be established by estimating a separate parameter for each of the cluster sizes that can occur. The parameter would then be $(\mu_1, \ldots, \mu_n; \pi_1, \ldots, \pi_n)$. Obviously, in this particular example, this consideration is of theoretical interest only, for two reasons. First, the parameters μ_N may not be of direct scientific value. Second, from a given experiment, we can estimate only one of them, and which one it will be is random in itself. This is different in Example 13, where typically more than one cluster size is observed in a given experiment. The fact that the parameter depends on the cluster size is then not a theoretical consideration, but a well studied problem often indicated by the term informative cluster size (Chiang and Lee, 2008; Aerts *et al.*, 2011). It is different, too, in the missing-data examples: allowing for a different parameter in different strata (also called patterns of missingness), brings us to the so-called pattern-mixture model (Molenberghs and Kenward, 2007).

While an informal statement only, it is useful to see that many estimands do not allow complete sufficient statistics because the corresponding parameters are estimated from data where this same parameter describes two or more natural strata simultaneously. By 'natural strata,' we mean strata that lead to separate sufficient statistics for the same parameter, without the opportunity to combine these into a single one. Looking at this from a different angle, it provides a basis for the following, existing, procedure. First, estimate separate copies of the parameter for every one of the strata. Second, combine these using appropriate weights. This procedure was studied by Hermans *et al.* (2018), based upon work by Molenberghs, Verbeke, and Iddi (2011).

4.7 Concluding Remarks

In this chapter, building upon the work reported in Liu and Hall (1999), Liu *et al.* (2006), Molenberghs *et al.* (2014), and Milanzi *et al.* (2016, 2015), we have provided an easy-to-use criterion for incompleteness of minimal sufficient statistics in univariate and multivariate exponential family models. Earlier work has typically studied incompleteness directly by means of the definition. This either implies that the existence of a non-trivial zero-expectation function needs to be falsified, or that such a function needs to be constructed.

Our result essentially requires checking the dimension of a minimal sufficient statistic relative to the length of the parameter vector. This turns the assessment of incompleteness into a feasible task, whereas the definition can be daunting to use and requires ad hoc construction of distributions of minimal sufficient statistics.

We have shown that clustered data designs with non-constant cluster sizes (random or otherwise) do not admit complete sufficient statistics. The term 'clustered data' has to be understood in the broadest sense; it encompasses longitudinal studies, multilevel designs, etc. On the one hand, longitudinal studies can have variable cluster sizes by design, while on the other, their cluster sizes can vary because of missing data.

The incompleteness of minimal sufficient statistics leads to the loss of some desirable properties, such as unbiasedness and optimality. But as shown in Molenberghs *et al.* (2014) and Milanzi *et al.* (2016, 2015), this does not need to be a serious problem in practice. For example, it is very well-known that, when data are missing, likelihood and Bayesian inferences can be based on the observed-data likelihood, without any correction for the variable cluster size, i.e., without any correction for the missing-data mechanism. Importantly, though, such methods cannot, in general, by default be claimed to be optimal, given that the Lehman-Scheffé theorem (Casella and Berger, 2001) does not apply. The consequences for the case of random cluster sizes, in particular informative cluster sizes, are not widely understood. When cluster sizes follow a random mechanism (in the sense of missing at random), it is thus possible to simply use the observed-data likelihood without *ad hoc* corrections. However, one cannot claim that such an approach is 'uniformly better' than any of the dedicated corrections. Arguably, it is prudent to investigate candidate methods' operational characteristics in settings relevant for the application at hand.

In the absence of complete sufficient statistics, some interesting philosophical issues appear. As discussed in Milanzi *et al.* (2015), some estimators will depend on the fact that more data could have been collected or that some data are available that, with certain probability, might not have been collected. They illustrated this using so-called *generalized sample averages* in sequential studies. When excluding such esoteric estimators, often only intuitively appealing estimators, such as the ordinary sample average, remain, even though there is no complete sufficient statistic, and in spite of some small-sample bias. Note, this shows great similarity with earlier work of Liu and Hall (1999) and Liu *et al.* (2006).

Our focus has been on characterizing incompleteness and, in particular, its consequences for point estimators. There obviously are important implications for hypothesis testing and interval estimation as well. An early reference is Anscombe (1949) and the topic, especially in the context of sequential designs, has received thorough treatment in Govindarajulu (1981), Barndorff-Nielsen and Cox (1984), and Barndorff-Nielsen and Cox (1994). More recently, members of the author team have studied the impact of incom-

plete sufficient statistics on estimation and hypothesis testing (Milanzi *et al.*, 2015), and the implications thereof for sequential designs (Milanzi *et al.*, 2016).

Chapter 5

Optimal Weighted Estimation for Hierarchical Models With Unequal Cluster Sizes: Compound-Symmetry Covariance

5.1 Introduction

In this chapter we consider the normal compound-symmetry structure to model hierarchical (or clustered) data with unequal cluster sizes. Molenberghs, Verbeke, and Iddi (2011) introduced the so-called split-sample methodology. The focus is on entirely non-iterative methods, bringing together the advantages of balanced data and simple averaging methodology. A consequence of this approach is the need for applying weights when combining results from the K strata. This chapter establishes how results on incomplete sufficient statistics in the context of weighted averages (Molenberghs *et al.*, 2014; Hermans *et al.*, 2018) imply that there may be no optimal set of weights. Given this, pragmatically attractive weights are proposed, in terms of efficiency, bias, and computational ease.

The remainder of the chapter is organized as follows. The compound-symmetry model is introduced in Section 5.2 and incompleteness results are reviewed, together with implications for likelihood-based estimation. Background from the pseudo-likelihood-based split-sample method was presented earlier in Section 2.3.2. A general split-sample approach to the CS model is provided in Section 5.3 and a number of specific but practically

relevant cases are considered. Details about the specifics for the CS case are presented in Section 5.4. Section 5.5 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 before in Section 3.2 are analyzed in Section 5.6. Ramifications and recommendations for practice are offered in Section 5.7.

5.2 The Compound-symmetry Model

The general notation is outlined in Section 2.1. Let $\boldsymbol{Y}$ follow the compound-symmetry normal law $\boldsymbol{Y} \sim N(\mu 1_n, \sigma^2 I_n + dJ_n)$. This is a three-parameter multivariate normal model with a common mean μ, a common variance $\sigma^2 + d$, and common covariance d. First incompleteness of the sufficient statistic is shown, and then continued with likelihood estimation. For both, we start from the log-likelihood function.

5.2.1 Incompleteness

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

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{i=1}^{c_k} -\frac{1}{2}\left(\boldsymbol{Y}_i^{(k)} - \mu 1_{n_k}\right)' \left(\sigma^2 I_{n_k} + dJ_{n_k}\right)^{-1} \left(\boldsymbol{Y}_i^{(k)} - \mu 1_{n_k}\right) \\
&= \sum_{k=1}^{K}\sum_{i=1}^{c_k} -\frac{1}{2}\left(\boldsymbol{Y}_i^{(k)} - \mu 1_{n_k}\right)' \left(I_{n_k} - \frac{d}{\sigma^2 + n_k d} J_{n_k}\right) \left(\boldsymbol{Y}_i^{(k)} - \mu 1_{n_k}\right) \\
&= \sum_{k=1}^{K}\sum_{i=1}^{c_k} \frac{\mu}{\sigma^2 + n_k d}\left(\sum_{j=1}^{n_k} Y_{ij}^{(k)}\right) - \frac{1}{2\sigma^2}\left(\sum_{k=1}^{K}\sum_{i=1}^{c_k}\sum_{j=1}^{n_k} Y_{ij}^{(k)2}\right) \\
&+ \sum_{k=1}^{K}\sum_{i=1}^{c_k} \frac{d}{2\sigma^2(\sigma^2 + n_k d)}\left(\sum_{j=1}^{n_k} Y_{ij}^{(k)}\right)^2 . \qquad (5.1)
\end{aligned}$$

The three terms in (5.1) 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. The sufficient

statistics are:

$$W_{1k} = \sum_{i=1}^{c_k}\sum_{j=1}^{n_k} Y_{ij}^{(k)}, \tag{5.2}$$

$$W_2 = \sum_{k=1}^{L}\sum_{i=1}^{c_k}\sum_{j=1}^{n_k}\left(Y_{ij}^{(k)}\right)^2, \tag{5.3}$$

$$W_{3k} = \sum_{i=1}^{c_k}\left(\sum_{j=1}^{n_k} Y_{ij}^{(k)}\right)^2, \tag{5.4}$$

$$W_{4k} = c_k. \tag{5.5}$$

Using the characterisation of Chapter 4, 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. The proof using the definition can be found in Appendix B.1.

5.2.2 Likelihood-based Estimation of the CS Model

Similar in spirit to (5.1), but now using all terms, the log-likelihood can be written as

$$\ell(\mu,\sigma^2,d) = \sum_{k=1}^{K}\ell_k(\mu,\sigma^2,d), \tag{5.6}$$

with the cluster size specific log-likelihood term:

$$\begin{aligned}\ell_k(\mu,\sigma^2,d) = -\frac{1}{2}\sum_{i=1}^{c_k}\Bigg\{&\ln\left[\sigma^{2n_k}+n_k\sigma^{2(n_k-1)}d\right]\\ +\quad&(\boldsymbol{Y}_i^{(k)}-\mu\mathbf{1}_{n_k})'\frac{1}{\sigma^2}\left(I_{n_k}-\frac{d}{\sigma^2+n_kd}J_{n_k}\right)(\boldsymbol{Y}_i^{(k)}-\mu\mathbf{1}_{n_k})\Bigg\}.\end{aligned} \tag{5.7}$$

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*. This means, 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:

$$\begin{aligned}\ell\left(\{\mu_k\}_k,\{\sigma_k^2\}_k,\{d_k\}_k\right) \propto\ &-\frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{c_k}\Big\{\ln|\Sigma_{n_k}|\\ &+\quad\left(\boldsymbol{y}_i^{(k)}-\boldsymbol{\mu}_{n_k}\right)'\Sigma_{n_k}^{-1}\left(\boldsymbol{y}_i^{(k)}-\boldsymbol{\mu}_{n_k}\right)\Big\},\end{aligned} \tag{5.8}$$

where $\boldsymbol{\mu}_k = \mu_k 1_N$ and $\Sigma_{n_k} = \sigma_k^2 I_{n_k} + d_k J_{n_k}$. The score functions are presented in Appendix B.2. Solving these score functions (B.2)–(B.4) leads to:

$$\widehat{\mu}_k = \frac{1}{c_k n_k} \sum_{i=1}^{c_k} \sum_{j=1}^{n_k} Y_{ij}^{(k)}, \tag{5.9}$$

$$\widehat{\sigma}_k^2 = \frac{1}{c_k n_k (n_k - 1)} \left(n_k \sum_{i=1}^{c_k} \boldsymbol{Z}_i^{(k)'} \boldsymbol{Z}_i^{(k)} - \sum_{i=1}^{c_k} \boldsymbol{Z}_i^{(k)'} J_{n_k} \boldsymbol{Z}_i^{(k)} \right), \tag{5.10}$$

$$\widehat{d}_k = \frac{1}{c_k n_k (n_k - 1)} \left(\sum_{i=1}^{c_k} \boldsymbol{Z}_i^{(k)'} J_{n_k} \boldsymbol{Z}_i^{(k)} - \sum_{i=1}^{c_k} \boldsymbol{Z}_i^{(k)'} \boldsymbol{Z}_i^{(k)} \right), \tag{5.11}$$

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

When the cluster size is constant, the compound-symmetry model has closed form ML estimators, given by (5.9)–(5.11). 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:

$$\text{var}(\widehat{\mu_k}) = \frac{\sigma_k^2 + n_k d_k}{c_k n_k}. \tag{5.12}$$

The expressions for the variance-covariance structure of $(\widehat{\sigma_k^2}, \widehat{d_k})$ is:

$$\text{var} \begin{pmatrix} \widehat{\sigma_k^2} \\ \widehat{d_k} \end{pmatrix} = \frac{2\sigma_k^4}{c_k n_k (n_k - 1)} \begin{pmatrix} n_k & -1 \\ -1 & \frac{\sigma_k^4 + 2(n_k-1) d_k \sigma_k^2 + n_k (n_k - 1) d_k^2}{\sigma_k^4} \end{pmatrix}. \tag{5.13}$$

The mean parameter is independent of the variance components.

The above 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 that will be 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.

All of these results are lost when $K \geq 2$. We briefly sketch the lack of closed-form solutions in this case in Appendix B.2.2.

The lack of a closed form is well known, but we highlight a few relevant features here. More detail is given in Appendix B.3, where the following identity is derived:

$$\widehat{\mu} = \frac{\sum_{k=1}^{K} \frac{n_k c_k}{\sigma^2 + n_k d} \widehat{\mu_k}}{\sum_{k=1}^{K} \frac{n_k c_k}{\sigma^2 + n_k d}}. \tag{5.14}$$

Examining (5.14) suggests weighted averages:

$$\widetilde{\mu} = \sum_{k=1}^{K} a_k \widehat{\mu_k}, \qquad \widetilde{\sigma^2} = \sum_{k=1}^{K} b_k \widehat{\sigma_k^2}, \qquad \widetilde{d} = \sum_{k=1}^{K} g_k \widehat{d_k}. \tag{5.15}$$

This idea is very similar to that in Molenberghs, Verbeke, and Iddi (2011), who split a sample in sub-samples, analyzed each of these separately, and then combined the result in an overall estimator. 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 will modify these ideas to the case of unequal cluster sizes, with a variable number of clusters per split.

5.3 Split-sample Methods for Clusters of Variable Size

The derivations are based on general pseudo-likelihood principles, reviewed in Section 2.3.1.

To fix ideas, consider log-likelihood (5.8). When used as an instrument to estimate a single vector (μ, σ^2, d), this function can be viewed as a pseudo-likelihood. This setting can be generalized by assuming that a dataset, consisting of repeated measures per subject, is divided into K subgroups, each containing c_k independent replicates. Consider the pseudo-likelihood:

$$p\ell(\boldsymbol{\theta}) = \sum_{k=1}^{K} \ell(\boldsymbol{\theta}_k | \boldsymbol{y}_1^{(k)}, \ldots, \boldsymbol{y}_{c_k}^{(k)}). \tag{5.16}$$

While the underlying principle is similar to (2.6), it is not identical. The similarities are: (1) all $\boldsymbol{\theta}_k$ are assumed to be different, allowing for separate, even parallel, estimation; (2) $\boldsymbol{\theta}$ stacks all vectors $\boldsymbol{\theta}_k$; (3) the parameter of interest $\boldsymbol{\theta}^*$, is found from an appropriate combination of the $\boldsymbol{\theta}_k$. Parallel estimation was also followed by Scott *et al.* (2013) and Neiswanger, Wang, and Xing (2013).

There are important differences, however. Here, and in the remainder of the chapter, we assume that $\ell(\boldsymbol{\theta}_k | \boldsymbol{y}_1^{(k)}, \ldots, \boldsymbol{y}_{c_k}^{(k)})$ is the likelihood that we would have, should group k be the only one in the data. That is, the individual likelihood contributions are not altered, rather the data are partitioned. This is similar to the independent partitioning done by Molenberghs, Verbeke, and Iddi (2011). In line with their derivations, (2.7) can also be used here. Given that $\ell_k(\boldsymbol{\theta}_k)$ is a genuine likelihood, its contributions to $I_0(\boldsymbol{\theta})$ and $I_1(\boldsymbol{\theta})$ are identical, up to the sign. As a result, $I_0(\boldsymbol{\theta})^{-1} I_1(\boldsymbol{\theta}) I_0(\boldsymbol{\theta})^{-1} = -I_0(\boldsymbol{\theta})^{-1}$, a block-diagonal matrix with blocks of the form $I_0(\boldsymbol{\theta}_k)$.

We now go further with pseudo-likelihood for split samples as discussed in Section 2.3.2. In general this produces the same estimator as full likelihood when all subjects follow the same distribution. This is in contrast with the CS settings outlined here.

Note that subjects in different sub-samples are allowed to have the same distribution, but that 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 these of size 1) contribute to all three parameters.

Consider pseudo-likelihood in this general case [see also Eq. (5.16)]. 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:

$$\widetilde{\boldsymbol{\theta}}^* = \sum_{k=1}^{K} A_k \widehat{\boldsymbol{\theta}}_k, \tag{5.17}$$

$$\text{var}(\widetilde{\boldsymbol{\theta}}^*) = \sum_{k=1}^{K} A_k V_k A_k', \tag{5.18}$$

with $V_k = I_0(\boldsymbol{\theta}_k)^{-1}$. We use the symbol $\widetilde{\boldsymbol{\theta}}^*$ to emphasize that this is not necessarily the maximum likelihood estimator even though, in our formalism, $\widehat{\boldsymbol{\theta}}_k$ is the maximum likelihood estimator when restricting attention to sub-sample k. Equation (5.18) 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

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

whence $I_p = \sum_{k=1}^{K} A_k$. Note that this requirement is satisfied for (2.9). 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 obvious choice when all subjects are i.i.d. and partitioning is in sub-samples of equal size. Proportional weights are obvious in the i.i.d. case, but with sub-samples of varying size.

The informal 'obvious' can be formalized by considering optimal weights. For this, define 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, $\Lambda = 2\left(\sum_{m=1}^K V_m^{-1}\right)^{-1}$ and finally:

Optimal weights. These take the form:

$$A_k^{\mathsf{opt}} = \left(\sum_{m=1}^K V_m^{-1}\right)^{-1} V_k^{-1}. \tag{5.19}$$

With this choice, (5.17)–(5.18) become:

$$\widetilde{\boldsymbol{\theta}}^* = \widehat{\boldsymbol{\theta}}^* = \left(\sum_{k=1}^K V_k^{-1}\right)^{-1} \sum_{k=1}^K V_k^{-1}\widehat{\boldsymbol{\theta}}_k, \tag{5.20}$$

$$\mathsf{var}(\widetilde{\boldsymbol{\theta}}^*) = V = \left(\sum_{k=1}^K V_k^{-1}\right)^{-1}. \tag{5.21}$$

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

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

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

where then formally $A_k = \mathsf{diag}(a_{k,1}, \ldots, a_{k,p})$. 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 will not necessarily correspond to the maximum likelihood solution. The rationale for choosing this route is computational convenience, and its advantages will vary from problem to problem.

Iterated optimal weights. The following iterative scheme can be followed:

1. Estimate $\widehat{\boldsymbol{\theta}}_k$.
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:

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

5. Repeat steps 3–4 until convergence.

This scheme can always be followed and it has the advantage that the data need only be analyzed once, to yield $\widehat{\boldsymbol{\theta}}_k$. 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 $V_k(\widetilde{\boldsymbol{\theta}_k})$ in (5.20). Here, $\widetilde{\boldsymbol{\theta}_k}$ could be, for example: (a) the sub-sample specific estimator $\widehat{\boldsymbol{\theta}_k}$; (b) $\widetilde{\boldsymbol{\theta}^*}$ 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 (5.18) 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, first note that

$$\frac{\partial}{\partial \boldsymbol{\theta}_k}(A_k \boldsymbol{\theta}_k) = A_k + \left(\frac{\partial A_k}{\partial \theta_{k1}} \boldsymbol{\theta}_k, \cdots, \frac{\partial A_k}{\partial \theta_{kp}} \boldsymbol{\theta}_k \right), \tag{5.22}$$

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,

$$\frac{\partial A_k}{\partial \theta_{kj}} = W^{-1} \frac{\partial W_k}{\partial \theta_{kj}} (I_p - W^{-1} W_k), \tag{5.23}$$

with I_p the p-dimensional identity matrix. Plugging (5.23) into (5.22), the proper delta-method approximation to the variance is:

$$\text{var}(\widetilde{\boldsymbol{\theta}}^*) \simeq \sum_{k=1}^{K} (A_k + B_k) V_k (A_k + B_k)', \tag{5.24}$$

with

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

and $\otimes$ signifying Kronecker product.

5.4 Partitioned-sample Analysis for the Compound-symmetry Model

For the normal compound-symmetry model, introduced in Section 5.2, a variety of options exists. We will sketch these here, and then consider some in greater detail.

Consider first 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 will do so here, in Section 5.4.2, both for sub-samples of equal as well as for unequal size.

Turning to the case of variable cluster size, we know from Section 5.2.2 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 5.3 to this case, and contrast it with an *ad hoc* moment-based set of estimators.

5.4.1 Variable Cluster Size

5.4.1.1 Optimal Weights

As we will see in Section 5.4.1.3, 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 (6.19) to the mean produces:

$$\widetilde{\mu} = \left(\sum_{k=1}^{K} \frac{c_k n_k}{\widehat{\sigma_k^2} + n_k \widehat{d_k}} \right)^{-1} \sum_{k=1}^{K} \frac{c_k n_k \widehat{\mu_k}}{\widehat{\sigma_k^2} + n_k \widehat{d_k}}. \tag{5.25}$$

The corresponding estimators for the variance components, specific to a cluster size, are given by (5.10) and (5.11). Using these, and expression (5.13) for the variance, it

follows that the optimal weighted estimator satisfies:

$$\begin{pmatrix} \widetilde{\sigma^2} \\ \widetilde{d} \end{pmatrix} = \left(\sum_{k=1}^{K} V_k^{-1} \right)^{-1} \sum_{k=1}^{K} \begin{pmatrix} \frac{Q_k}{2\widehat{\sigma_k^2}} - \frac{d_k(2\widehat{\sigma_k^2}+n_k\widehat{d_k})}{2\widehat{\sigma_k^4}(\widehat{\sigma_k^2}+n_k\widehat{d_k})^2} R_k \\ \frac{R_k}{2(\sigma_k^2+n_k d_k)^2} \end{pmatrix}, \tag{5.26}$$

with Q_k and R_k as in (B.5) and (B.6), respectively.

5.4.1.2 Iterated and Approximate Optimal Weights

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

Replacing the variance components in (5.25) by their expectation leads to:

$$\widetilde{\mu} = \left(\sum_{k=1}^{K} \frac{c_k n_k}{\sigma^2 + n_k d} \right)^{-1} \sum_{k=1}^{K} \frac{c_k n_k \widehat{\mu_k}}{\sigma^2 + n_k d}. \tag{5.27}$$

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

$$\mu = \left(\sum_{k=1}^{K} \frac{c_k n_k}{\sigma^2 + n_k d} \right)^{-1} \sum_{k=1}^{K} \frac{c_k n_k \mu}{\sigma^2 + n_k d}. \tag{5.28}$$

Although (5.25) cannot directly be used, because of circularity, (5.27) and (5.28) are available to us.

Replacing the variance components on the right hand side of (5.26) by their expectations leads to:

$$\begin{pmatrix} \widetilde{\sigma^2} \\ \widetilde{d} \end{pmatrix} = \left(\sum_{k=1}^{K} V_k^{-1} \right)^{-1} \sum_{k=1}^{K} \begin{pmatrix} \frac{Q_k}{2\sigma^2} - \frac{d(2\sigma^2+n_k d)}{2\sigma^4(\sigma^2+n_k d)^2} R_k \\ \frac{R_k}{2(\sigma^2+n_k d)^2} \end{pmatrix}. \tag{5.29}$$

Using the explicit expressions for these, and using the fact that the expectation must be $(\sigma^2, d)'$, (5.26) leads to the following identity:

$$\begin{pmatrix} \sigma^2 \\ d \end{pmatrix} = V \sum_{k=1}^{K} \frac{c_k n_k}{2(\sigma^2 + n_k d)} \begin{pmatrix} \frac{\sigma^2+(n_k-1)d}{\sigma^2} \\ 1 \end{pmatrix}. \tag{5.30}$$

Expressions (5.25) and (5.26) 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 (5.27) and (5.29) can be used for iterated weighting, as discussed in the previous section. Note that 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, $\widetilde{\mu}$ on the one hand and $\widetilde{\sigma^2}$ and $\widetilde{d}$ on the other can be determined separately, provided the latter are done first.

Expressions (5.28) and (5.30) move beyond the previous schemes and exist by virtue of their explicit expressions. In (5.30) 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 (5.28) and the process repeated for μ, until convergence.

5.4.1.3 Scalar Weights

In this case, A_k equals diag(a_k, b_k, g_k), with the scalars as in (5.15). 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 (5.9)–(5.11) and their variance-covariance structure by (5.12)–(5.13). The objective function to find the optimum is

$$Q = \sum_{k=1}^K a_k^2 \mathsf{var}(\widehat{\mu_k}) - \lambda \left(\sum_{k=1}^K a_k \right).$$

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

$$a_k = \frac{\frac{c_k n_k}{\sigma^2 + n_k d}}{\sum_{m=1}^K \frac{c_m n_m}{\sigma^2 + n_m d}} = \frac{\frac{c_k n_k}{(1-\rho) + n_k \rho}}{\sum_{m=1}^K \frac{c_m n_m}{(1-\rho) + n_m \rho}}, \tag{5.31}$$

$$b_k = \frac{c_k (n_k - 1)}{\sum_{m=1}^K c_m (n_m - 1)}, \tag{5.32}$$

$$g_k = \frac{\frac{c_k n_k}{\frac{\sigma^4}{n_k - 1} + 2\sigma^2 d + n_k d^2}}{\sum_{m=1}^K \frac{c_m n_m}{\frac{\sigma^4}{n_m - 1} + 2\sigma^2 d + n_m d^2}} = \frac{\frac{c_k n_k (n_k - 1)}{(1-\rho)^2 + [2\rho(1-\rho) + n_k \rho^2](n_k - 1)}}{\sum_{m=1}^K \frac{c_m n_m (n_m - 1)}{(1-\rho)^2 + [2\rho(1-\rho) + n_m \rho^2](n_m - 1)}}, \tag{5.33}$$

where $\rho = d/(\sigma^2 + d)$. Several observations are worth making. First, 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, exactly like for the vector-valued weights. The added advantages are: (1) matrix computations simplify to scalar computations; for models with relatively few parameters, like the one here, this is a small advantage; (2) 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 (5.31), 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. However, they are different for the variance components.

5.4.1.4 Approximate Optimal Scalar Weights

To illustrate the logic of this method, consider (5.31)–(5.33) for the case where cluster sizes, for a good majority of the clusters, are sufficiently large. Taking limits for $n_k \to +\infty$ produces:

$$a_k^{\mathsf{app}} = g_k^{\mathsf{app}} = c_k/N. \tag{5.34}$$

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

$$b_k^{\mathsf{app}} = \frac{c_k n_k}{\sum_{m=1}^{K} c_m n_m}, \tag{5.35}$$

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

Figure 5.1 depicts optimal scalar weights (5.31)–(5.33), alongside the apparently simplistic proportional weights, for two of the five NTP datasets, chosen such as 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 (5.35), the proportional weights seem to offer a sensible choice. This issue will be examined further in the data-analytic Section 5.6.

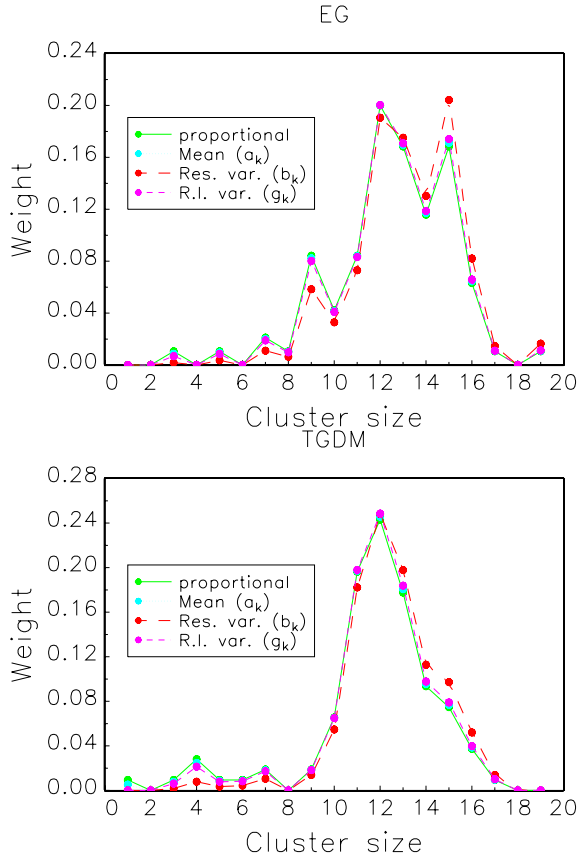

Figure 5.1: *NTP Data. Scalar weights: proportional and optimal scalar versions for EG and TGDM datasets. The optimal scalar weights are computed for* $\rho = d/(\sigma^2 + d) = 0.5$.

5.4.2 The Special Case of Common Cluster Size, Splits of (Un)equal Size

When $n_k \equiv n$ is constant, (5.31)–(5.33) reduce to:

$$a_k = b_k = g_k = c_k/N, \tag{5.36}$$

Hence, while $a_k = b_k$ reduce to proportional weights, this is not true 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 thus, evidently, equal weights follow. Both the similarities and subtle differences with the results from Section 5.4.1.4 are worth pointing out. Expressions (5.34) and (5.36) are identical except for the parameter σ^2.

5.4.3 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 all $c_k \equiv 1$. 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 Appendix Section B.4.1. In particular, we derive under what asymptotics such an estimator is consistent.

5.5 Simulation Study

A first, limited, simulation study was carried out to examine the behavior of the partitioning method. Details are given in Appendix Section B.5. 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 (i.e., proportional to c_k), size proportional (i.e., 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). Most importantly, all of these results are extremely 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 Appendix Section B.6. The most striking conclusion is that closed-form solutions are very much faster than their alternatives, while at the same time yielding the 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.

5.6 Application: NTP Data

The data, introduced in Section 3.1, are here analyzed in three ways. In Section 5.6.1 maximum likelihood estimators are presented, with split sampling, where splitting is by cluster size and using various weighting schemes. In Section 5.6.2, a dose effect is added to these. Finally, the cluster-by-cluster methodology of Section 5.4.3 is illustrated in Section 5.6.3.

5.6.1 Splitting by Cluster Size, No Dose Effect

Tables 5.1–5.3 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 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.

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 these, 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 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

Table 5.1: *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 (5.35) 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 [(B.25)(B.26)(B.27)]								
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.$(\widehat{\mu})$	0.01149	0.01155	0.01076	0.01360	0.01076	0.00766	0.00766	0.00766
s.e.$(\widehat{\sigma^2})$	0.00084	0.00084	0.00128	0.00199	0.00094	0.00092	0.00061	0.00138
s.e.$(\widehat{d})$	0.00196	0.00199	0.00210	0.00293	0.00210	0.00048	0.00045	0.00340
Two-stage [(B.25)(B.30)(B.31)]								
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.$(\widehat{\mu})$	0.01149	0.01155	0.01169	0.01496	0.01169	0.00901	0.00901	0.00901
s.e.$(\widehat{\sigma^2})$	0.00084	0.00084	0.00092	0.00127	0.00074	0.00072	0.00057	0.02404
s.e.$(\widehat{d})$	0.00196	0.00199	0.03045	0.02915	0.03045	0.02537	0.00087	0.27337
Unbiased two-stage [(B.25)(B.34)(B.35)]								
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.$(\widehat{\mu})$	0.01149	0.01155	0.01257	0.01679	0.01257	0.00958	0.00958	0.00958
s.e.$(\widehat{\sigma^2})$	0.00084	0.00084	0.00128	0.00199	0.00094	0.00092	0.00061	0.00172
s.e.$(\widehat{d})$	0.00196	0.00199	0.00291	0.00447	0.00291	0.00101	0.00102	0.00634

there is information loss because of using plug-in estimates in the weights, rather than the true ones.

Table 5.2: *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 (5.35) 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 [(B.25)(B.26)(B.27)]								
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.$(\widehat{\mu})$	0.01402	0.01410	0.01393	0.01485	0.01393	0.01346	0.01346	0.01346
s.e.$(\widehat{\sigma^2})$	0.00041	0.00041	0.00046	0.00051	0.00044	0.00044	0.00031	0.00328
s.e.$(\widehat{d})$	0.00265	0.00269	0.00264	0.00272	0.00264	0.00230	0.00230	0.00476
Two-stage [(B.25)(B.30)(B.31)]								
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.$(\widehat{\mu})$	0.01402	0.01410	0.01423	0.01522	0.01423	0.01379	0.01379	0.01379
s.e.$(\widehat{\sigma^2})$	0.00041	0.00041	0.00037	0.00041	0.00037	0.00037	0.00029	0.03410
s.e.$(\widehat{d})$	0.00265	0.00269	0.02814	0.02555	0.02814	0.02632	0.00243	0.05214
Unbiased two-stage [(B.25)(B.34)(B.35)]								
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.$(\widehat{\mu})$	0.01402	0.01410	0.01493	0.01665	0.01493	0.01452	0.01452	0.01452
s.e.$(\widehat{\sigma^2})$	0.00041	0.00041	0.00046	0.00051	0.00044	0.00044	0.00031	0.00363
s.e.$(\widehat{d})$	0.00265	0.00269	0.00302	0.00333	0.00302	0.00271	0.00271	0.00533

5.6.2 Splitting by Cluster Size, With Dose Effect

While the above results illustrate the explicit derivations in this thesis (i.e., with a constant mean), the data analysis in Section 5.6.1 does not do full justice to the actual design of the experiment, because the question of real scientific interest is the dose-response

Table 5.3: *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 (5.35) 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 [(B.25)(B.26)(B.27)]								
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.$(\widehat{\mu})$	0.01926	0.01936	0.01780	0.02502	0.01780	0.01257	0.01257	0.01257
s.e.$(\widehat{\sigma^2})$	0.00044	0.00044	0.00052	0.00079	0.00047	0.00046	0.00033	0.00308
s.e.$(\widehat{d})$	0.00529	0.00537	0.00570	0.01043	0.00570	0.00159	0.00159	0.00329
Two-stage [(B.25)(B.30)(B.31)]								
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.$(\widehat{\mu})$	0.01926	0.01936	0.01804	0.02535	0.01804	0.01297	0.01297	0.01297
s.e.$(\widehat{\sigma^2})$	0.00044	0.00044	0.00042	0.00059	0.00039	0.00039	0.00030	0.02568
s.e.$(\widehat{d})$	0.00529	0.00537	0.03433	0.03113	0.03433	0.02215	0.00173	0.04036
Unbiased two-stage [(B.25)(B.34)(B.35)]								
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.$(\widehat{\mu})$	0.01926	0.01936	0.01937	0.02989	0.01937	0.01390	0.01390	0.01390
s.e.$(\widehat{\sigma^2})$	0.00044	0.00044	0.00052	0.00079	0.00047	0.00046	0.00033	0.00353
s.e.$(\widehat{d})$	0.00529	0.00537	0.00718	0.01542	0.00718	0.00205	0.00205	0.00382

relationship. Let x_i be the dose administered to cluster i, taking one out of 4 to 5 values. Recall that the dose levels for the DEHP study are given in Table 3.1. The model then becomes $\boldsymbol{Y}_i \sim N\left(\{\beta_0 + \beta_1 x_i\}\, 1_n, \sigma^2 I_n + d J_n\right)$. 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 5.4. The results are comforting, showing that proportional and approximate scalar weights are a sensible choice. This consistent with theoretical considerations, the simulations results, and the analysis in Section 5.6.1.

5.6.3 Cluster-by-cluster Methods

Next, we illustrate the cluster-by-cluster methods. Results are presented in Tables 5.1–5.3, 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 cases, (B.25) is used for the mean. For the variance components, the pairs (B.26)–(B.27), (B.30)–(B.31), and (B.34)–(B.35) 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 5.1–5.3.

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. Which of the three versions is better is less clear: it differs somewhat from compound to compound and from parameter to parameter. All in all, all three show acceptable behaviour. 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 analysed in parallel and the results then combined.

5.7 Concluding Remarks

To study simple, computationally effective, and statistically sound methods of estimation, we considered the simple but insightful case of clustered data with a normal compound-symmetry structure and clusters of varying size. Because of this non-constant cluster size, there is no closed-form maximum likelihood estimator and maximization must proceed iteratively. This is not a problem in small and medium numbers of clusters and cluster sizes. However, when the number of clusters and/or the number of repetitions per cluster grow large to very large, computations may become challenging. Our simulations show

Table 5.4: *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 (5.35) 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.$(\widehat{\beta_0})$	0.01343	0.01357	0.01343	0.01609	0.01343
s.e.$(\widehat{\beta_1})$	0.00012	0.00012	0.00014	0.00018	0.00014
s.e.$(\widehat{\sigma^2})$	0.00084	0.00084	0.00128	0.00199	0.00094
s.e.$(\widehat{d})$	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.$(\widehat{\beta_0})$	0.01453	0.01470	0.01389	0.01406	0.01389
s.e.$(\widehat{\beta_1})$	0.00001	0.00001	0.00001	0.00001	0.00001
s.e.$(\widehat{\sigma^2})$	0.00041	0.00041	0.00044	0.00049	0.00042
s.e.$(\widehat{d})$	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.$(\widehat{\beta_0})$	0.01356	0.01370	0.01335	0.02000	0.01335
s.e.$(\widehat{\beta_1})$	0.00006	0.00006	0.00006	0.00007	0.00006
s.e.$(\widehat{\sigma^2})$	0.00044	0.00044	0.00052	0.00079	0.00047
s.e.$(\widehat{d})$	0.00126	0.00130	0.00110	0.00205	0.00110

that, in certain settings, computations can be made up to 30,000 times faster than with standard maximum likelihood.

Fundamentally, we observe that the sufficient statistic for this setting is incomplete,

implying that there is no uniform optimal unbiased estimator. The MLE is only locally optimal. This offers an interesting perspective. First note that, when the cluster size is constant, then there is a closed-form solution. Then 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 these, but the weights depend on the parameters. This suggests the consideration of approximations to these weights, as well as alternative weights. Based on theoretical results and simulations, as well as on real-data analysis, we found that equal weights and so-called approximate optimal weights do not perform well. Iterated optimal and proportional weights show excellent performance. While the former of these two are somewhat more computationally intensive, the latter 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 has been based on the simple, three-parameter compound-symmetry model, in the data analysis we considered a slightly expanded setting, where the mean takes the form of a regression function rather than a constant. This is encouraging towards 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 proportional weights or rather size proportional weights are to be preferred becomes difficult or impossible. 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 would then imply 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 have considered the important special case of a cluster-by-cluster analysis. Encouragingly, 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 chapter, but may be very sensible in a meta-analysis of clinical trials.

When clusters become very large, it may become 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 Chapter 1. While work has been done in this area, it is of interest to combine the ideas developed in this chapter with a model for cluster size.

Chapter 6

Optimal Weighted Estimation for Hierarchical Models With Unequal Cluster Sizes: AR(1) Covariance

6.1 Introduction

In the previous chapter the normal CS-model was considered to model hierarchical (or clustered) data with unequal cluster sizes. 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 chapter, 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.

The chapter is organised as follows. In Section 6.2 the model formulation is given. In Section 6.3 the estimators for a single constant cluster size are presented. The (in)completeness property is outlined in Section 6.4, and in Section 6.5 various weighting schemes for clusters of unequal size are explored. In Section 6.6, a simulation study is described for the investigation of the performance of the suggested weights and the data are analysed in Section 6.7. The closing discussion is presented in Section 6.8.

6.2 Model Formulation

The general notation is outlined in Section 2.1. All models considered in this chapter will be versions of the following general linear mixed model:

$$\boldsymbol{Y}_i^{(k)}|b_i^{(k)} \sim N(X_i^{(k)}\boldsymbol{\beta} + Z_i^{(k)}b_i^{(k)}, \Sigma_i^{(k)}), \tag{6.1}$$

$$b_i^{(k)} \sim N(0, D), \tag{6.2}$$

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 $Z_i^{(k)}b_i^{(k)}$ drops from (6.1), while $\Sigma_i^{(k)} = \sigma^2 C_{n_k}$, with entry (r, s) equal to $\rho^{|r-s|}$. For ease of exposition, the mean structure will often be taken to be $\mu\mathbf{1}_{n_k}$, with $\mathbf{1}_{n_k}$ an n_k column vector of ones.

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

6.3 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 6.5.

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 6.5. The AR(1) model of Section 6.2 can then be written as:

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

Because $C \equiv C(\rho)$, the parameter vector is $\boldsymbol{\theta} = (\boldsymbol{\beta}', \rho, \sigma^2)$. When the mean is constant $\boldsymbol{\mu}_i = X_i\boldsymbol{\beta} = \mu\mathbf{1}$. It is often stated that the MLE for the AR(1) model, with a constant or more elaborate mean structure, requires numerical iteration. This is certainly the case when not all clusters are of the same size. However, in the constant cluster size case considered here, there is a closed-form solution. Our development follows, in part, Kenward (1981).

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

$$\ell \propto -\frac{c}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{c}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)'\Sigma^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i). \tag{6.3}$$

The score equation for the mean produces, as usual:

$$\widehat{\boldsymbol{\beta}} = \frac{1}{c}\sum_{i=1}^{c}(X_i'\Sigma^{-1}X_i)^{-1}(X_i'\Sigma^{-1}\boldsymbol{Y}_i). \tag{6.4}$$

Consider (6.4) for the case of a constant mean. If Σ corresponds to independence or compound-symmetry, the MLE for μ is the ordinary sample average, it does not depend on covariance parameters. For a general design $\boldsymbol{\beta}$ is estimated by the OLS estimator. However, in our AR(1) case, solving the score equations leads to:

$$\widehat{\mu} = \frac{1}{c[(n-2)(1-\rho)+2]}\sum_{i=1}^{c}\left(\sum_{j=1}^{n}Y_{ij} - \rho\sum_{j=2}^{n-1}Y_{ij}\right). \tag{6.5}$$

Not only does (6.5) depend on ρ (hence the MLE for ρ needs to be plugged in), it differs from the OLS:

$$\widetilde{\mu} = \frac{1}{cn}\sum_{i=1}^{c}\sum_{j=1}^{n}Y_{ij}. \tag{6.6}$$

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

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

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

$$\sigma^2 = \frac{1}{cn}\sum_{i=1}^{c}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)'C^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i). \tag{6.7}$$

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

$$\sigma^2\frac{2\rho}{1-\rho^2} = \frac{1}{c(n-1)}\sum_{i=1}^{c}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)'F(\boldsymbol{y}_i - \boldsymbol{\mu}_i), \tag{6.8}$$

with

$$\begin{aligned}
F &= \frac{\partial C^{-1}}{\partial\rho}\\
&= \frac{1}{(1-\rho^2)^2}\text{tridiag}\left\{[2\rho, 4\rho, \ldots, 4\rho, 2\rho]'; \left[-(1+\rho^2), \ldots, -(1+\rho^2)\right]'\right\}
\end{aligned} \tag{6.9}$$

and with tridiag$(\boldsymbol{v}_1, \boldsymbol{v}_2)$ a tri-diagonal matrix with $\boldsymbol{v}_1$ along the main diagonal and $\boldsymbol{v}_2$ on the adjacent diagonals. Both (6.7) and (6.8) contain a summation that can be rewritten as tr$(S \cdot Q)$, with

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

and Q either C^{-1} or F, as in (6.9), respectively. Using this formulation, and some straightforward but tedious algebra, produces:

$$f(\rho) = (n-1)S_2\rho^3 - (n-2)R\rho^2 - (nS_2 + S_1)\rho + nR = 0, \tag{6.10}$$

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

$$S_1 = \sum_{j=1}^{n} s_{jj}, \qquad S_2 = \sum_{j=2}^{n-1} s_{jj}, \qquad R = \sum_{j=1}^{n-1} s_{j,j+1}. \tag{6.11}$$

These can be plugged into (6.5) to obtain $\widehat{\mu}$ and into:

$$\widehat{\sigma}^2 = \frac{1}{c} \cdot \frac{1}{1-\widehat{\rho}^2} \left(S_1 + \widehat{\rho}^2 S_2 - 2\widehat{\rho} R \right), \tag{6.12}$$

to obtain the MLE for σ^2.

It is easy to see that $f(\rho)$ has a single root in $[-1, 1]$. Indeed, $f(-\infty) = -\infty$, $f(+\infty) = +\infty$, $f(-1) > 0$, and $f(1) < 0$. The other two real roots are therefore in $]-\infty, -1]$ and $[1, +\infty[$. The general solution of a third-degree polynomial follows from Cardano's method. The polynomial under study was examined by Kenward (1981) who, using results of Koopmans (1942), derived an expression for the solution inside $[-1, 1]$. Alternatively, the method of Shelbey (1975) can be used. It takes the following form. Write the polynomial symbolically as $f(\rho) = a\rho^3 + b\rho^2 + c\rho + d$, define

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

and further

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

For three real roots $t_0 \leq t_1 \leq t_2$, it follows that $t_0 = C(p, q)$, $t_2 = -C(p, -q)$, and $t_1 = -t_0 - t_2$. Finally, $\widehat{\rho} = t_1 - b/(3a)$. While not as simple as the other explicit expressions for estimators, the key point is that it has a closed-form which, in turn, can be used to obtain a closed form solution for the mean and the variance, using (6.5) and (6.12), 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 C.2, lead to:

$$\mathcal{I}_{\sigma^2\rho,\sigma^2\rho} = c \begin{pmatrix} \frac{n}{2(\sigma^2)^2} & -\frac{n-1}{\sigma^2}\cdot\frac{\rho}{1-\rho^2} \\ -\frac{n-1}{\sigma^2}\cdot\frac{\rho}{1-\rho^2} & \frac{(n-1)(1+\rho^2)}{(1-\rho^2)^2} \end{pmatrix}. \tag{6.13}$$

It is convenient to slightly change (6.13) to

$$\widetilde{\mathcal{I}}_{\sigma^2\rho,\sigma^2\rho} = c \begin{pmatrix} \frac{n-1}{2(\sigma^2)^2} & -\frac{n-1}{\sigma^2}\cdot\frac{\rho}{1-\rho^2} \\ -\frac{n-1}{\sigma^2}\cdot\frac{\rho}{1-\rho^2} & \frac{(n-1)(1+\rho^2)}{(1-\rho^2)^2} \end{pmatrix}, \tag{6.14}$$

yielding a very simple inverse:

$$\widetilde{\mathcal{I}}^{-1}_{\sigma^2\rho,\sigma^2\rho} = \frac{1}{c(n-1)} \begin{pmatrix} \frac{2(\sigma^2)^2(1+\rho^2)}{1-\rho^2} & 2\sigma^2\rho \\ 2\sigma^2\rho & 1-\rho^2 \end{pmatrix}. \tag{6.15}$$

6.4 Complete and Incomplete Sufficient Statistics

In what follows, we will establish completeness for the balanced conditional independence model, with the reverse holding for AR(1) model. The criteria set out in Chapter 4 will be used. 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.

6.4.1 Balanced Conditionally Independent Model.

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

6.4.2 AR(1) Model

The mean estimator (6.5) consists of two sufficient statistics:

$$K_1 = \sum_{i=1}^{c}\sum_{j=1}^{n} Y_{ij}, \qquad K_2 = \sum_{i=1}^{c}\sum_{j=2}^{n-1} Y_{ij}, \tag{6.16}$$

with the sufficient statistics for σ^2 and ρ spelt out in (6.11). In other words, the three-component vector $\boldsymbol{\theta} = (\mu, \sigma^2, \rho)'$ has a minimal sufficient statistic (K_1, K_2, S_1, S_2, R) of dimension 5, establishing incompleteness.

Even though the AR(1) model has not been studied before from the perspective of incomplete sufficient statistics, its ramifications have been mentioned in the literature. For example, as described by Martin (2006), Papadakis proposed, as early as 1937, a correction to the least-squares estimator for correlated observations arising in such settings as adjoining plots designs (Papadakis, 1937; Bartlett, 1938, 1976, 1978). The topic was also touched upon by Cochran and Bliss (1948), in the context of discriminant analysis combined with analysis of covariance. Clearly, the opportunity for such an *ad hoc* correction arises from the incompleteness. Martin (2006) and earlier authors discussing Papadakis' method refer to the somewhat unusual dependence of the mean estimator on the variance components. This parallels the property of the MLE for the mean in the AR(1) case, as in (6.5). Indeed, because ρ is estimated from solving a third-degree polynomial with coefficients that are functions of the sums of squares and cross-products matrix, it too is a function of such deviations. Of course, the ρ in our case is more complex than Papadakis' correction, which was more of an *ad hoc* nature, while our estimator is the solution to the likelihood equations. In essence, Papadakis' method builds a covariate from deviations observed from adjacent plots. Especially when the plots are arranged as a linear array, the connection with AR(1) is strong. Both non-iterative and iterative versions were proposed by Papadakis. In the iterative case, the covariate is re-built after every iteration, using the current value of the parameters. In more general settings, the data have a spatial layout.

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

Cochran and Bliss (1948, p. 172) noted that the relative efficiency of the estimators with or without the use of covariance is not uniformly larger or smaller than one, but that for sufficiently large sample sizes the difference between them is small. This is entirely consistent with our findings for the AR(1) case. For Papadakis' method, the impact on bias and efficiency is described by Martin (2006). We refer to our simulations in Section 6.6.

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 6.1. *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 C.2.4. This results offers the opportunity to consider estimators, based on weighting that, while not statistically fully efficient, have computational advantages such as stability (e.g., by being entirely non-iterative) and speed.

Proposition 6.2. *The result of Proposition 6.1 easily generalizes to a mean of the form $\boldsymbol{\mu} = X\beta$, when the design is constant among clusters.*

6.5 Clusters Of Variable Size

Various weighting schemes were studied in Section 5.4 for clusters of unequal size in the compound-symmetry case. The work was rooted in the pseudo-likelihood and split-sample methods of Fieuws and Verbeke (2006) and Molenberghs, Verbeke, and Iddi (2011). We will not reproduce their entire argument here, it suffices to focus on the following two-stage procedure:

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

$$\widetilde{\boldsymbol{\theta}}^* = \sum_{k=1}^{K} A_k \widehat{\boldsymbol{\theta}}_k, \tag{6.17}$$

$$\text{var}(\widetilde{\boldsymbol{\theta}}^*) = \sum_{k=1}^{K} A_k V_k A_k'. \tag{6.18}$$

We showed that the sum of the weight matrices should be the identity matrix, an obvious result, and considered, among others, the optimal expression:

$$A_k^{\text{opt}} = \left(\sum_{m=1}^{K} V_m^{-1} \right)^{-1} V_k^{-1}. \tag{6.19}$$

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 β coefficients can then be estimated by:

$$\widetilde{\beta} = \left(\sum_{k=1}^{K} \sum_{i=1}^{c_k} X_i^{(k)'} C_k^{-1} X_i^{(k)} \right)^{-1} \left(\sum_{k=1}^{K} \sum_{i=1}^{c_k} X_i^{(k)'} C_k^{-1} Y_i^{(k)} \right). \tag{6.20}$$

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

$$\text{var}(\widehat{\mu}_k) = v_k = \frac{\sigma^2(1+\rho)}{c_k} \cdot \frac{1}{[n_k - (n_k - 2)\rho]}. \tag{6.21}$$

The optimal weight is then

$$a_k = \frac{c_k \left[n_k - (n_k - 2)\rho \right]}{\sum_{m=1}^{K} c_m \left[n_m - (n_m - 2)\rho \right]}. \tag{6.22}$$

It is insightful to consider (6.22) in a few special cases:

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

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

How well the approximation works is seen in a few special cases. When $\rho = 0.5$, $a_k \propto c_k(n_k + 2)$; for $\rho = 0.9$ this becomes $a_k \propto c_k(n_k + 18)$; finally for $\rho = 0.99$, we find $a_k \propto c_k(n_k + 198)$. Thus, for larger correlations, the size-proportionally matches clusters of sizes much larger than actually observed. But again, in practice, it is convenient and reasonable to operate under size-proportionality.

When estimating the variance of

$$\widetilde{\mu} = \sum_{k=1}^{K} a_k \mu_k, \tag{6.23}$$

using (6.22), the fact that the weights depend on ρ_k needs to be taken into account. Applying the delta method to (6.23), and using the variance expressions in both (6.21)

and (6.15), we find:

$$\begin{aligned}\mathsf{var}(\widetilde{\mu}) &= \frac{\sum_{k=1}^{K} a_k' \sigma_k^2 (1+\rho_k)}{\left(\sum_{k=1}^{K} a_k'\right)^2} \\ &+ \frac{\sum_{k=1}^{K} \left[c_k(n_k-2)\sum_{m=1}^{K} a_m'(\mu_k-\mu_m)\right]^2 \frac{1-\rho_k^2}{c_k(n_k-1)}}{\left(\sum_{k=1}^{K} a_k'\right)^4}. \end{aligned} \tag{6.24}$$

We can plug in the stratum-specific $\widehat{\rho}_k$ and $\widehat{\sigma}_k^2$, or instead use the overall $\widehat{\rho}$ and $\widehat{\sigma}^2$. In the latter case, (6.24) becomes:

$$\begin{aligned}\mathsf{var}(\widetilde{\mu}) &= \sigma^2(1+\rho)\left\{\frac{1}{\sum_{k=1}^{K} a_k'}\right\} \\ &+ (1-\rho^2)\left\{\frac{\sum_{k=1}^{K} \frac{c_k(n_k-2)^2}{(n_k-1)}\left[\sum_{m=1}^{K} a_m'(\mu_k-\mu_m)\right]^2}{\left(\sum_{k=1}^{K} a_k'\right)^4}\right\}. \end{aligned} \tag{6.25}$$

Turning to the variance components, we start from (6.14), and use $V_k^{-1} c_k(n_k-1)P$ with

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

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

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

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

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

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

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

Algebraic details on formulas can be found in separate Appendix C.2.5 and C.2.6.

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 as in Chapter 5.

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

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

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

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

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

A second simulation study was conducted to compare computation time for closed form solutions to numerical solutions. Using closed form solutions reduces computation time significantly. Details can be found in the separate Appendix C.3.

6.7 Application: Clinical Trials in Schizophrenia

The data, introduced in Section 3.2, are analysed here. The active treatments are: risperidone, haloperidol, perphenazine, and zuclopernthixol. We included for analysis patients with at least one follow-up measurement. Table 6.1 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:

$$Y_{ij} = \mu + \alpha_i + \beta t_{ij} + (\alpha\beta)_{ij} + \epsilon_{ij},\ i = 1, \ldots, 4,\ j = 1, \ldots, n,\ \epsilon_{ij} \sim N_n(0, R), \quad (6.26)$$

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

Table 6.2 shows the treatment levels which appear in the different splits. Not all the treatments are present in each split. In other words, not all the splits are contributing to the estimation of every parameter. This fact should be taken into account for constructing the weights. For example, for estimating levomepromazine effect, just the first two splits are contributing, therefore, we have $(c_1 = 142, n_1 = 2)$ and $(c_2 = 143, n_2 = 3)$, which give proportional weights as $(0.498, 0.502)$, and the size-proportional weights as $(0.398, 0.602)$.

Table 6.3 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 5 trials. It might be useful to include the trial and its interaction with the variables already in the model (6.26) to control for the trial effect:

$$\begin{aligned} Y_{ijk} = \mu + \tau_i + \alpha_j + \beta t_{ij} + (\tau\alpha)_{ij} + (\tau\beta)_{ik} + (\alpha\beta)_{jk} + (\tau\alpha\beta)_{ijk} + \epsilon_{ijk}, \\ i = 1, \ldots, 5,\ j = 1, \ldots, 4,\ k = 1, \ldots, n,\ \epsilon_{ijk} \sim N_n(0, R), \quad (6.27) \end{aligned}$$

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

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

Table 6.4 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 C.4.

Table 6.2: *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	✓	✓	✓	✓	✓

6.8 Concluding Remarks

As an extension to the normal-compound symmetry model, discussed in Hermans *et al.* (2018) i.e. Chapter 5, the normal AR(1) model was studied in the light of computationally effective estimation for clustered data with unequal cluster sizes.

For constant cluster size, there are closed-form solutions but no complete minimal sufficient statistics. However the MLE is shown to be optimal, with weights depending on ρ for the mean. Returning to unequal cluster sizes, there are, in general, no closed form solutions. But again estimators have been obtained using a two-stage procedure. Estimators are calculated separately within each stratum (typically defined by cluster size) and combined in an overall estimator. Both theoretical and simulation results show excellent performance of the size-proportional weights, that is through weighting according to the number of measurements in a cluster ($c_k \cdot n_k$), rather than the number of clusters c_k in a subsample, that is, proportional weights. By contrast, the latter are a good choice for the compound-symmetry structure. Under AR(1) they are worse than equal weights. Approximate optimal weights can also be used, but this leads to an estimate of ρ with a large sample variance. In practice, it is convenient and appropriate to use size-proportional weights; these are parameter free and simple to use. Simulations show, that in certain large settings, computation time can be 1000 times faster than with standard maximum likelihood.

Table 6.3: *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 (6.26).*

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)

There are missing observations in the PANSS data set. One might therefore consider possible dependencies between cluster size and the outcomes themselves. To handle such informative cluster sizes it might be of interest to extend the current methodology of this chapter 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.

Table 6.4: *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 (6.27).*

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)

Chapter 7

Optimal Weighted Estimation Versus Cochran-Mantel-Haenszel

7.1 Introduction

Categorical variables take on an *a priori* fixed and finite number of possible values, in particular, two values for binary data. Investigators often want to examine associations between categorical variables, for example, in a case-control study. In the past, the need grew to analyse the relation between binary variables incorporating the classification due to other relevant confounders. Mantel and Haenszel (1959) published their view on the analysis of case-control studies more than half a century ago. Their methodology has become ubiquitous in epidemiology and beyond. The general methodology and notation were outlined in Section 2.4. All statistical procedures mentioned in this chapter are appropriate for data settings with binary responses, involving stratification, grouping, or matching based on confounding variables.

Our goal is to place the MH against the background of estimators that do follow from optimality considerations. The split-sample based approach, constructed in Chapters 5 and 6, is now extended to the described grouped data settings. The basic ideas, combining subsample-specific results using appropriate weights, are similar, but there are computational differences. By contrasting both methods the nature of the Mantel-Haenszel estimator becomes more clear, as well as its unique properties.

This chapter is organised as follows. In Section 7.2, pseudo-likelihood methodology is applied and weighting schemes are explored. In Section 7.3, a simulation study is described to compare the performance of the MH estimator with one following from

optimal weighting considerations. An example is discussed in Section 7.4. The discussion and recommendations for practice are offered in Section 7.5.

7.2 Optimal Weighted Estimation

Breslow and Day (1980) examined likelihood estimation to obtain an estimator for the common odds ratios. However, they showed that there is no closed-form solution, except for the case where there are only two strata. Numerical estimation techniques are necessary in pursuing an estimate. In comparison with this, the MH is much simpler to use. In this thesis we use weighted estimation for data settings with unequal cluster sizes. Based on the pseudo-likelihood split-sample approach of Molenberghs, Verbeke, and Iddi (2011), the sample is divided into subsamples, containing clusters of equal size. For each subsample, maximum likelihood estimators are calculated and subsample-specific results then combined using weights. The entire argument will not be reproduced, but these same ideas can be used for the data settings discussed in this chapter. The subsamples considered here are naturally the various strata in the sample. The subsample or stratum-specific estimator is the odds ratio, ψ, calculated as:

$$\psi_i = \frac{a_i \cdot d_i}{b_i \cdot c_i}. \tag{7.1}$$

These can be combined into a common odds ratio using weights α_i:

$$\tilde{\psi} = \sum_{i=1}^{N} \alpha_i \psi_i, \tag{7.2}$$

with $\sum_{i=1}^{N} \alpha_i = 1$. It is natural and well known that optimal weights are inversely proportional to a measure of variance (see the Appendix for a brief sketch of the argument). We will now investigate this further, against the background of the lack of complete sufficient statistics.

7.2.1 (In)Complete Sufficient Statistics

Consider the weighted odds ratio estimator as in (7.2), and assume $E[\psi_i] = \psi$, then

$$E[\tilde{\psi}] = \sum_{i=1}^{N} \alpha_i E[\psi_i] = \psi \sum_{i=1}^{N} \alpha_i = \psi.$$

Suppose that there is a non-zero function $g((\psi_i)_i) = \sum_i \beta_i \psi_i$, such that $E[g((\psi_i)_i)] = \sum_i \beta_i \psi = \psi \sum_i \beta_i = 0$. This is satisfied for all β_i's where $\sum_i \beta_i = 0$. By this counterexample, incompleteness holds. As a consequence, it is *a priori* not guaranteed that there is a uniformly optimal estimator. However, it should be noted that the existence of a uniform optimum, while not guaranteed by the theorem, is not necessarily excluded.

7.2.2 Optimal Weights

To obtain (potentially local) optimal weights, we seek to minimize the variance, $\text{var}(\tilde{\psi}) = \sum_{i=1}^{N} \alpha_i^2 \text{var}(\psi_i)$, under the constraint that $\sum_{i=1}^{N} \alpha_i = 1$ using the method of Lagrange multipliers. The calculations, which are standard and applicable in a wide variety of settings, are briefly reviewed in Appendix D.1. The weights are:

$$\alpha_i = \frac{v_i^{-1}}{\sum_j v_j^{-1}}. \tag{7.3}$$

In the next step, the variance of a stratum-specific odds ratio will be expressed explicitly. When taking the natural logarithm of the odds ratio, the variance here equals $\text{var}(\log(\psi_i)) = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}$. By using the delta method we find $\text{var}(\log(\psi_i)) \cong \frac{1}{\psi_i^2} \cdot \text{var}(\psi_i)$ and now

$$\begin{aligned} \text{var}(\psi_i) &= \psi_i^2 \text{var}(\log(\psi_i)) \\ &= \psi_i^2 \left(\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i} \right) \\ &\overset{\text{pop. value}}{\cong} \frac{\psi^2 q}{n_i}, \end{aligned} \tag{7.4}$$

with $q = \frac{1}{p_{11}} + \frac{1}{p_{10}} + \frac{1}{p_{01}} + \frac{1}{p_{00}}$, and p_{11}, p_{10}, p_{01} and p_{00} the 2×2 cell probabilities.

Note that a transition from population to estimated variances is made. By doing this, the stratum-specific variance and its inverse equal:

$$\text{var}(\psi_i) \cong \frac{\psi^2 q}{n_i} = v_i \Rightarrow v_i^{-1} = \frac{n_i}{\psi^2 q}, \tag{7.5}$$

resulting in the following formula for the weights:

$$\alpha_i = \frac{\frac{n_i}{\psi^2 q}}{\sum_j \frac{n_j}{\psi^2 q}} = \frac{n_i}{n}, \tag{7.6}$$

with $n = \sum_{i=1}^{N} n_i$. If all $(n_i)_i$ would be fixed by design, this is a uniform minimal solution, in spite of incompleteness. Using the above expressions, it follows that the uniformly optimal weighted estimator satisfies:

$$\tilde{\psi} = \sum_{i=1}^{N} \frac{n_i}{n} \psi_i. \tag{7.7}$$

Equation (D.4) yields an expression for the overall variance of this common odds ratio:

$$\text{var}(\tilde{\psi}) = \sum_{i=1}^{N} \frac{\sum_{i=1}^{N} v_i^{-2} v_i}{(\sum_{j=1}^{N} v_j^{-1})^2} = \frac{1}{\sum_{i=1}^{N} v_i^{-1}}. \tag{7.8}$$

This scheme is more principled and enjoys minimum variance properties, in comparison to MH. However, a disadvantage is that the weighted estimator does not yield a well-defined estimate as soon as there is a zero cell in one contingency table. It suggests that MH may well be superior in small samples. Also, because MH does not take the form of a conventional weighted estimator, with weights differing from the optimal ones, also the behavior in large to very large samples should be investigated.

We can also use the observed rather than expected variances. Then, an alternative form emerges:

$$\tilde{\tilde{\psi}} = \sum_{i=1}^{N} \alpha_i \psi_i = \frac{\sum_{i=1}^{M} v_i^{-1} \psi_i}{\sum_{i=1}^{M} v_i^{-1}}. \tag{7.9}$$

Now,

$$\begin{aligned} v_i &= \psi_i^2 \left(\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i} \right) \\ v_i^{-1} &= \psi_i^{-2} h_i, \end{aligned} \tag{7.10}$$

with

$$h_i = \left(\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i} \right)^{-1} = \frac{a_i b_i c_i d_i}{e_i}, \tag{7.11}$$

with $e_i = a_i b_i c_i + a_i b_i d_i + a_i c_i d_i + b_i c_i d_i$. This leads to:

$$\tilde{\tilde{\psi}} = \frac{\sum_i \frac{(c_i b_i)^2}{e_i}}{\sum_i \frac{1}{\psi_i} \frac{(c_i b_i)^2}{e_i}}. \tag{7.12}$$

Yet another estimator would be:

$$\tilde{\tilde{\psi}}' = \frac{\sum_i a_i d_i \frac{(c_i b_i)^2}{e_i}}{\sum_i c_i b_i \frac{(c_i b_i)^2}{e_i}}. \tag{7.13}$$

7.3 Simulation Study

In the following simulation study, carried out in R, the performance of the derived estimators is compared with the conventional MH.

To start, a sample with five strata and stratum sizes 900, 500, 200, 600, and 300 is considered. For each stratum, random numbers following a uniform distribution are generated, as many as the stratum size requires. These numbers are compared to the cumulative probabilities (0, 0.2, 0.3, 0.7, 1) from the original contingency table (Table 7.1) to determine the frequencies for a new stratum-specific 2×2 table. For example, the sampled number 0.4 will contribute to the count in cell (2,1) as it lies between 0.3 and

Table 7.1: *Simulation study. The original* 2×2 *table used in the simulation study.*

0.2	0.1
0.4	0.3

0.7. This results in five new 2×2 tables. To these tables, the MH and estimators (7.7), (7.12), and (7.13) are applied. In this scenario there were 5 strata with mean stratum size 500. In further samples, both the number of strata and mean stratum size are multiplied by a factor of ten. All values for numbers of strata (5; 50; 500; 5,000; and 50,000) and mean stratum sizes (500; 5,000; 50,000; 500,000; and 5,000,000) are combined into 25 scenarios. Each scenario was sampled 500 times, leading to 500 estimates of the common odds ratio for each estimator. By doing this we can explore the estimators' performance under varying designs. As an aside, when the mean stratum size and number of strata get large, the simulation is time-consuming. As the performance of the estimators is already proven in the middle of the table, some of the bottom cells in the upcoming tables are left blank.

7.3.1 Relative Efficiency

As an important evaluation criterion, we compare the relative efficiency of the proposed estimators with the MH. First, we calculate the empirical variances of our estimators over the sample of 500 odds ratios. Second, the model-based variances are used. For the Mantel-Haenszel estimator the variance formula (2.15) is used, and (7.8) is used as our estimator.

The results based on the empirical variances are presented in Tables 7.2–7.4. The results based on the model-based variances are reproduced in Table 7.5. The relative efficiencies are presented as percentages.

Considering Tables 7.2–7.4, the MH is slightly more efficient for large samples but not for huge samples. For extremely large overall sample sizes, the alternative estimators (7.7) and (7.12) are performing equally well as the MH. The opposite is found in Table 7.5, where variance estimator (7.8) turns out to be smaller, and so more precise, than the asymptotic variance of Robins *et al.* (1986b); at the same time it is easier to compute. Pattanayak *et al.* (2012) also showed in their simulation study that the variance estimator of Robins *et al.* (1986b) is conservative, producing unnecessarily wide confidence intervals. This simulation study confirms this. However, when the mean stratum sizes become very large, this issue goes away.

Table 7.2: *Simulation study. Relative efficiency of estimator (7.7) w.r.t. Mantel-Haenzel estimator. (Empirical)*

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	90.51	98.62	100.12	100.02	100.00
50	90.42	98.71	99.52	100.03	99.96
500	89.61	99.12	99.51	99.96	100.02
5 000	90.93	99.19	100.17	99.92	
50 000	92.41	99.38	99.79		

Table 7.3: *Simulation study. Relative efficiency of estimator (7.12) w.r.t. Mantel-Haenszel estimator. (Empirical)*

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	97.65	101.44	99.57	99.92	99.99
50	94.08	100.35	100.86	99.86	100.07
500	90.30	99.17	100.77	100.06	100.01
5 000	90.14	98.97	99.25	100.20	
50 000	90.95	98.44	99.94		

Table 7.4: *Simulation study. Relative efficiency of estimator (7.13) w.r.t. Mantel-Haenszel estimator. (Empirical)*

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	69.65	75.31	65.65	72.73	73.28
50	66.07	71.80	64.36	68.85	72.04
500	67.04	73.18	72.28	64.97	69.57
5 000	70.98	66.53	70.02	66.33	
50 000	71.02	68.67	67.58		

Table 7.5: *Simulation study. Relative efficiency w.r.t. Mantel-Haenszel estimator. (Model based)*

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	106.58	100.65	100.06	100.01	100.00
50	108.50	100.80	100.08	100.01	100.00
500	108.63	100.83	100.08	100.01	100.00
5 000	107.66	100.83	100.08	100.01	
50 000	107.66	100.83	100.08		

7.3.2 Coverage Probabilities, Bias, and Mean Squared Error

Next, the coverage probabilities, bias, and mean squared error (MSE) of the estimators are examined. Table 7.6 shows the coverage probabilities for estimator (7.7), where the 95% confidence interval is calculated using the proposed variance estimator (7.8) and normal quantiles. For comparison, Tables 7.8 and 7.9 give the coverage probabilities for the MH. Where Table 7.9 is constructed in line with Table 7.6, in Table 7.8 the confidence intervals are calculated starting from the log odds ratio and using an exponential transformation. These tables suggest that the proposed estimator is performing badly in the left bottom corner of the tables, as the coverage probabilities become small and even zero. The same phenomena can be observed in Tables 7.7 and 7.10 where the confidence intervals are calculated with the sample variance. The latter is done to make sure that there is no over-coverage, which does not seem to be the case. Tables 7.11–7.14 show, respectively, the bias and MSE of the alternative estimator and the MH. Here, the bias and MSE are larger in the same part of the tables.

The reason for this occurrence cannot be assigned to the proposed weights of the alternative estimator. These are of the same magnitude in each row, and working perfectly well in a part of the settings. It appears that, when the number of strata becomes equal to or larger than the mean stratum size, the bias and the uncertainty become too large to obtain a proper estimate.

7.4 Application: Intego Data

The Intego data (see Section 3.3) set analysed here contains information about 338,581 patients, listing their gender, year of birth, the general practices, and diagnosis recorded with the correct ICPC Code. We chose to make several 2×2 tables with the binary

Table 7.6: *Simulation study. Coverage Probabilities for estimator (7.7).*

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	0.940	0.950	0.950	0.930	0.950
50	*0.816*	0.932	0.946	0.936	0.942
500	*0.186*	*0.846*	0.940	0.964	0.932
5 000	*0.000*	*0.204*	*0.868*	0.948	
50 000	*0.000*	*0.000*	*0.204*		

Table 7.7: *Simulation study. Coverage Probabilities for estimator (7.7) with sample variance.*

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	0.938	0.940	0.946	0.938	0.956
50	*0.862*	0.946	0.944	0.940	0.954
500	*0.198*	*0.848*	0.944	0.956	0.930
5 000	*0.000*	*0.212*	*0.832*	0.948	
50 000	*0.000*	*0.000*	*0.200*		

Table 7.8: *Simulation study. Coverage Probabilities for Mantel-Haenszel estimator (log oddsratio).*

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	0.962	0.950	0.950	0.930	0.950
50	0.948	0.942	0.946	0.936	0.942
500	0.960	0.958	0.944	0.960	0.932
5 000	0.954	0.954	0.966	0.942	
50 000	0.954	0.940	0.958		

Table 7.9: *Simulation study. Coverage Probabilities for Mantel-Haenszel estimator.*

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	0.952	0.952	0.952	0.930	0.950
50	0.958	0.940	0.946	0.938	0.942
500	0.960	0.958	0.944	0.960	0.932
5 000	0.954	0.954	0.966	0.942	
50 000	0.954	0.940	0.958		

Table 7.10: *Simulation study. Coverage Probabilities for Mantel-Haenszel estimator with sample variance.*

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	0.954	0.948	0.950	0.938	0.956
50	0.956	0.954	0.946	0.942	0.958
500	0.956	0.956	0.948	0.956	0.932
5 000	0.948	0.954	0.952	0.942	
50 000	0.956	0.940	0.954		

Table 7.11: *Simulation study. Bias for estimator (7.7).*

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	0.045 487	0.006 383	-0.000 194	-0.000 174	0.000 067
50	*0.039 827*	0.004 114	0.000 356	-0.000 001	0.000 018
500	*0.040 240*	*0.003 951*	0.000 404	0.000 019	9.580e-07
5 000	*0.039 646*	*0.003 817*	*0.000 400*	0.000 025	
50 000	*0.039 552*	*0.003 840*	*0.000 386*		

Table 7.12: *Simulation study. MSE for estimator (7.7).*

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	0.022 471	0.001 935	0.000 190	0.000 020	0.000 002
50	*0.003 755*	0.000 225	0.000 019	0.000 002	2.016e-07
500	*0.001 819*	*0.000 034*	0.000 002	1.762-e07	1.851e-08
5 000	*0.001 591*	*0.000 016*	*3.248e-07*	1.939e-08	
50 000	*0.001 566*	*0.000 015*	*1.674e-07*		

Table 7.13: *Simulation study. Bias for Mantel-Haenszel estimator.*

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	0.013 178	0.003 413	-0.000 491	-0.000 204	0.000 064
50	0.001 020	0.000 444	-0.000 012	-0.000 038	0.000 015
500	0.000 849	0.000 181	0.000 028	-0.000 018	-0.000 003
5 000	0.000 180	0.000 053	0.000 025	-0.000 013	
50 000	0.000 104	0.000 072	0.000 010		

Table 7.14: *Simulation study. MSE for Mantel-Haenszel estimator.*

Number	**Mean stratum size**				
of strata	500	5 000	50 000	500 000	5 000 000
5	0.018 639	0.001 880	0.000 191	0.000 020	0.000 002
50	0.001 962	0.000 205	0.000 018	0.000 002	2.014e-07
500	0.000 180	0.000 019	0.000 002	1.762e-07	1.852-e08
5 000	0.000 017	0.000 002	1.657e-07	1.894-e08	
50 000	0.000 002	1.929e-07	1.864e-08		

Table 7.15: *Intego Data. General* 2×2 *table.*

	Diabetes +	**Diabetes -**	Total
Female	a_i	b_i	n_{1i}
Male	c_i	d_i	n_{2i}
Total	m_{1i}	m_{2i}	n_i

Table 7.16: *Intego Data. Common odds ratio and variance estimates: (a) ψ_{MH}: Mantel-Haenszel odds ratio estimate; (b) $\tilde{\psi}$: odds ratio estimate with (7.7); (c) $\tilde{\tilde{\psi}}$: odds ratio estimate with (7.12); (d) $\tilde{\tilde{\psi}}'$: odds ratio estimate with (7.13); (e) v_R: variance estimate according to Robins* et al. *(1986b) (2.15); (f) v: variance estimate with (7.8).*

Strat.	# strata	ψ_{MH}	$\tilde{\psi}$	$\tilde{\tilde{\psi}}$	$\tilde{\tilde{\psi}}'$	v_R	v
GP	75	0.967	0.937	0.904	0.944	2.330×10^{-4}	2.524×10^{-4}
YB[1]	2	0.921	0.954	0.916	0.951	2.303×10^{-4}	2.286×10^{-4}
YB[2]	4	0.930	1.108	0.907	0.895	2.294×10^{-4}	2.367×10^{-4}
YB[3]	8	0.928	-	-	-	2.305×10^{-4}	-

[1] split by 1950

[2] split by 1920, 1950, 1980

[3] split by 1900, 1920, 1935, 1950, 1965, 1980, 2000

variables diabetes and gender (Table 7.15). General practices and the year of birth can serve as stratification variables. There are 75 different general practices. The years of birth vary from 1898 to 2015. Table 7.16 presents the estimates for the common odds ratio and variance according to different stratification variables. In the first row general practices (GP) divided the sample in different strata. For the second row, year of birth (YB) was split into two groups, those who where born before 1950 and those born in 1950 and later. In further rows more splits were made.

As the simulation study already suggested, all estimates are very close to each other. All estimators are equally easily computed, no matter the number of strata. In the last line of the table, the stratum with people born before 1900 has one zero in its 2×2 table. As we stated earlier and now can see here: the MH is well-defined and gives an estimate, however due to a zero weight this specific stratum is omitted in the calculation. On the other hand, the formally optimal estimator does not yield a well-defined estimate.

7.5 Concluding Remarks

To study associations between binary variables, incorporating classification, the common odds ratio is often used. The odds ratio estimator of Mantel and Haenszel (1959) is well established in epidemiology. It is a weighted estimator, combining information of the different strata, that do not need to be of the same size. It follows neither from the likelihood principle, nor from optimally weighted considerations. We therefore used pseudo-likelihood split-sample methodology (Molenberghs, Verbeke, and Iddi, 2011), to

study it against the background of an optimal weighted estimator and two variations of this. We noted that, while there is no complete sufficient statistic, there nevertheless is a uniformly optimal weighted estimator. Unlike a full likelihood estimator, both the MH as well as the optimal estimator, and the variance thereof, are easy to calculate.

The MH, in spite of its somewhat *ad hoc* nature is very efficient for large datasets, and in more cases well-defined when cell counts are small or even zero. It is fair to say the optimal estimator has a somewhat easier and more intuitive variance expression but, as we have already seen, both are easy to compute.

In some cases when datasets are huge, the optimal estimator is somewhat more efficient. However, we should bring forward two reservations. First, this is not the case when the number of strata becomes equal to or larger than the mean stratum size; here, the estimator fails. Second, with really huge samples, a slight amount of efficiency loss is usually not an issue.

Categorical variables may take on more than 2 levels and also here associations can be of scientific interest. This suggests that one might also consider a possible extension to $I \times J \times K$ tables as is done for the MH (Agresti, 2002, pp. 295).

We conclude that the MH retains its practical and theoretical attraction, even when compared with formally optimal estimators.

Chapter 8

Doubly Robust Pseudo-likelihood for Incomplete Hierarchical Binary Data

8.1 Introduction

Incomplete data has become an important concern for applied statisticians, especially in longitudinal and otherwise hierarchical outcome data. When there is missingness in the data, the process behind these, as well as its impact on inference, needs to be addressed. Very commonly, direct likelihood is used for analyzing correlated data under MAR. Linear mixed models and generalized linear mixed models are popular choices, though marginalization is not always straightforward. Other likelihood-based options, e.g. the Bahadur (1961) and the multivariate Dale or global odds ratio model (Molenberghs & Lesaffre, 1994, 1999), can involve complex likelihoods, can be computationally prohibitive in moderate to large studies, and are vulnerable to misspecification. The most popular alternative is generalized estimating equations or GEE (Liang and Zeger, 1986; Diggle et al., 2002; Molenberghs and Verbeke, 2005). Standard GEE is valid only under MCAR, but a weighted version (WGEE; Robins, Rotnitzky, and Zhao, 1995) has been developed, using Horvitz-Thompson ideas (Cochran, 1977), to allow valid use of GEE under MAR. Doubly robust approaches (Scharfstein, Rotnitzky, and Robins, 1999; Van der Laan & Robins, 2003; Bang & Robins, 2005; Rotnitzky, 2009; Birhanu et al., 2011), which further supplement the use of weights with a predictive model for the unobserved responses, given the observed ones, have been constructed. This not only removes or at least alleviates bias, but also increases efficiency.

Pseudo-likelihood (PL) methods (le Cessie & van Houwelingen, 1991; Geys, Molen-

berghs, and Lipsitz, 1998; Geys, Molenberghs, & Ryan, 1999; Aerts *et al.*, 2002) comprise yet another alternative to full likelihood. Molenberghs et al. (2011) proposed corrections, following single and double robustness ideas, to the standard form of pseudo-likelihood, to ensure the validity under MAR. This is reviewed in Section 2.3.3. Molenberghs et al. (2011) applied the methodology to multivariate Gaussian responses and to a conditional model for clustered binary data. They provided a general outline with predominantly illustrative examples using normal and binary data. However, the marginal modeling of longitudinal binary data is very common in practice. Molenberghs et al. (2011) only sketched the methodology using a marginal Bahadur model for the binary responses; they did not pursue it in detail. The further development of doubly robust pseudo-likelihood for incomplete hierarchical binary data under MAR is the central theme of this chapter.

The theoretical part, estimating equations and precision estimators, are calculated and reported for the first time. Application is shown through a case study and easy-to-use SAS code is provided.

It should be clear that we are not fitting the full Bahadur model. In fact, we use its first and second moments only, because this allows us to describe the marginal mean function, whilst providing the vehicle to take correlations and incompleteness into account. Note that there is a similar connection between standard and weighted GEE for binary data on the one hand and the Bahadur model on the other. The latter connection was studied in detail by Molenberghs and Kenward (2010). Note that apart from very simple settings, the Bahadur model is prohibitive to fit (Aerts *et al.*, 2002).

For further background, Section 2.3.3 reviews Pseudo-likelihood for data MAR and 2.2.2 describes the full Bahadur model. In this chapters, the contribution, i.e., pseudo-likelihood based on the Bahadur model, is the subject of Section 8.2. Analysis of the case study can be found in Section 8.3

8.2 Pseudo-likelihood for Incomplete Binary Data

8.2.1 General Formulation

Molenberghs et al. (2011) considered three classes of estimating equations for pairwise likelihood, respectively naive, singly robust ('sr'), and doubly robust ('dr'). For each of these three, the original authors further considered: complete cases (CC; using only subjects will all planned measurements observed), complete pairs (CP; where all complete pairs from incomplete sequences are also added), and available cases (AC; where additionally single observations from incomplete pairs are used), leading to nine sets of estimating equations. The word 'naive' refers to the fact that these estimating equations would generally lead to biased estimators under MAR. Here only the response is modelled

with a Bahadur model. For the single robust setting a weight model is introduced, using a logistic structure. For the double robust version the model was further extended with a predictive model for the unobserved outcomes using again a Bahadur model. All these estimating equations are presented in Table 8.1.

In this table, $\widetilde{R}_i = 1$ if subject i is fully observed and 0 otherwise. In the robust cases, the probability for subject i to be completely observed and to be observed up to and including occasion j are respectively denoted as

$$\pi_i = \prod_{\ell=2}^{n_i}(1-p_{i\ell}) \qquad \text{and} \qquad \pi_{ij} = \prod_{\ell=2}^{j}(1-p_{i\ell}),$$

where $p_{i\ell} = P(D_i = \ell | D_i \geq \ell, \boldsymbol{y}_{i\bar{\ell}}, \boldsymbol{x}_{i\bar{\ell}})$ are the component probabilities of dropping out at occasion ℓ, given the subject is still in the study, the covariate history $\boldsymbol{x}_{i\bar{\ell}}$ and the outcome history $\boldsymbol{y}_{i\bar{\ell}}$. $p_{i\ell}$ can be modeled using a logistic regression. Further, R_{ijk} and π_{ijk} are the indicator and probability, respectively, for observing both Y_{ij} and Y_{ik}. Note that for the case of dropout, whenever $j < k$,

$$R_{ijk} \equiv R_{ik} \qquad \text{and} \qquad \pi_{ijk} \equiv \pi_{ik} = \prod_{\ell=2}^{k}(1-p_{i\ell}),$$

in which case, e.g. the single robust version of the CP estimating equation can be re-expressed as:

$$\boldsymbol{U}_{\mathsf{CP,sr}} = \sum_{i=1}^{N} \sum_{j<k<d_i} \frac{R_{ik}}{\pi_{ik}} \boldsymbol{U}_i(y_{ij}, y_{ik}).$$

An important result is that all three doubly robust versions coincide (Molenberghs et al., 2011), i.e.,

$$\begin{aligned} \boldsymbol{U}_{\mathsf{CC,dr}} &= \boldsymbol{U}_{\mathsf{CP,dr}} = \boldsymbol{U}_{\mathsf{AC,dr}} = \\ &= \sum_{i=1}^{N} \left\{ \sum_{j<k<d_i} \boldsymbol{U}_i(y_{ij}, y_{ik}) + \sum_{j=1}^{d_i-1} (n_i - d_i + 1) \boldsymbol{U}_i(y_{ij}) \right. \\ &\qquad \left. + \sum_{j<d_i\leq k} E\left[\boldsymbol{U}_i(y_{ik}|y_{ij})\right] + \sum_{d_i\leq j<k} E\left[\boldsymbol{U}_i(y_{ij}, y_{ik})\right] \right\}. \end{aligned} \tag{8.1}$$

It is thus not necessary to explicitly model the missing-data mechanism. Further, under exchangeability, Molenberghs et al. (2011) showed that the expectations in $\boldsymbol{U}_{\mathsf{AC,dr}}$ vanish, making Eq. (8.1) essentially equivalent to $\boldsymbol{U}_{\mathsf{AC,naive}}$, which is very convenient for implementation, as this reduces to an observed data analysis. More information on this can be found in Appendix E.1.

Table 8.1: *Estimating equations for pairwise pseudo-likelihood. Abbreviations used: CC: complete cases; CP: complete pairs; AC: available pairs; sr: singly robust; dr: doubly robust.*

type	$\boldsymbol{U}_{*,\mathsf{naive}}$	$\boldsymbol{U}_{*,\mathsf{sr}}$	$\boldsymbol{U}_{*,\mathsf{dr}}$
$\boldsymbol{U}_{\mathsf{CC},*}$	$\sum_{i=1}^{N} \widetilde{R}_i \sum_{j<k} \boldsymbol{U}_i(y_{ij}, y_{ik})$	$\sum_{i=1}^{N} \frac{\widetilde{R}_i}{\pi_i} \sum_{j<k} \boldsymbol{U}_i(y_{ij}, y_{ik})$	$\sum_{i=1}^{N} \sum_{j<k} \left[\frac{\widetilde{R}_i}{\pi_i} \boldsymbol{U}_i(y_{ij}, y_{ik}) + \left(1 - \frac{\widetilde{R}_i}{\pi_i}\right) E_{\boldsymbol{Y}^m \mid \boldsymbol{y}^o} \boldsymbol{U}_i(y_{ij}, y_{ik}) \right]$
$\boldsymbol{U}_{\mathsf{CP},*}$	$\sum_{i=1}^{N} \sum_{j<k<d_i} \boldsymbol{U}_i(y_{ij}, y_{ik})$	$\sum_{i=1}^{N} \sum_{j<k<d_i} \frac{R_{ijk}}{\pi_{ijk}} \boldsymbol{U}_i(y_{ij}, y_{ik})$	$\sum_{i=1}^{N} \sum_{j<k<n_i} \left[\frac{R_{ijk}}{\pi_{ijk}} \boldsymbol{U}_i(y_{ij}, y_{ik}) + \left(1 - \frac{R_{ijk}}{\pi_{ijk}}\right) E_{\boldsymbol{Y}^m \mid \boldsymbol{y}^o} \boldsymbol{U}_i(y_{ij}, y_{ik}) \right]$
$\boldsymbol{U}_{\mathsf{AC},*}$	$\sum_{i=1}^{N} \left[\sum_{j<k<d_i} \boldsymbol{U}_i(y_{ij}, y_{ik}) + \sum_{j=1}^{d_i-1} (n_i - d_i + 1) \boldsymbol{U}_i(y_{ij}) \right]$	$\sum_{i=1}^{N} \left[\sum_{j=1}^{d_i-1} \frac{R_{ij}}{\pi_{ij}} \boldsymbol{U}_i(y_{ij}) + \sum_{j<k} \frac{R_{ik}}{\pi_{ik}} \boldsymbol{U}_i(y_{ik} \mid y_{ij}) \right]$	$\sum_{i=1}^{N} \left[\sum_{j<k} \frac{R_{ik}}{\pi_{ik}} \boldsymbol{U}_i(y_{ik} \mid y_{ij}) + \sum_{j=1}^{d_i-1} \frac{R_{ij}}{\pi_{ij}} \boldsymbol{U}_i(y_{ij}) + \sum_{j<k} \left(1 - \frac{R_{ik}}{\pi_{ik}}\right) E_{\boldsymbol{Y}^m \mid \boldsymbol{y}^o} \boldsymbol{U}_i(y_{ik} \mid y_{ij}) + \sum_{j=1}^{d_i-1} \left(1 - \frac{R_{ij}}{\pi_{ij}}\right) E_{\boldsymbol{Y}^m \mid \boldsymbol{y}^o} \boldsymbol{U}_i(y_{ij}) \right]$

8.2.2 Pairwise Bahadur Model for the Outcome

Based on the definitions made in Section 2.2.2 the log-likelihood terms from a pairwise Bahadur model take the following form:

$$\begin{aligned} p\ell_{ijk} &= y_{ij}y_{ik}\ln\nu_{ijk} + y_{ij}(1-y_{ik})\ln(\nu_{ij}-\nu_{ijk}) + (1-y_{ij})y_{ik}\ln(\nu_{ik}-\nu_{ijk}) \\ &+ (1-y_{ij})(1-y_{ik})\ln(1-\nu_{ij}-\nu_{ik}+\nu_{ijk}). \end{aligned} \tag{8.2}$$

Starting from pseudo-likelihood contribution (8.2), pairwise and conditional contributions to the score equation take the form as follows

$$\begin{aligned} \boldsymbol{U}_{ijk} &= \frac{y_{ij}y_{ik}}{\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\theta}}\nu_{ijk} + \frac{y_{ij}(1-y_{ik})}{\nu_{ij}-\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\theta}}(\nu_{ij}-\nu_{ijk}) + \frac{(1-y_{ij})y_{ik}}{\nu_{ik}-\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\theta}}(\nu_{ik}-\nu_{ijk}) \\ &+ \frac{(1-y_{ij})(1-y_{ik})}{1-\nu_{ij}-\nu_{ik}+\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\theta}}(1-\nu_{ij}-\nu_{ik}+\nu_{ijk}), \end{aligned} \tag{8.3}$$

$$\begin{aligned} \boldsymbol{U}_{ik|j} &= \frac{y_{ij}y_{ik}\nu_{ij}}{\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{\nu_{ijk}}{\nu_{ij}}\right) + \frac{y_{ij}(1-y_{ik})\nu_{ij}}{\nu_{ij}-\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{\nu_{ij}-\nu_{ijk}}{\nu_{ij}}\right) \\ &+ \frac{(1-y_{ij})y_{ik}(1-\nu_{ij})}{\nu_{ik}-\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{\nu_{ik}-\nu_{ijk}}{1-\nu_{ij}}\right) \\ &+ \frac{(1-y_{ij})(1-y_{ik})(1-\nu_{ij})}{1-\nu_{ij}-\nu_{ik}+\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{1-\nu_{ij}-\nu_{ik}+\nu_{ijk}}{1-\nu_{ij}}\right), \end{aligned} \tag{8.4}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$, and $\nu_{ij} = \nu_{ij}(\boldsymbol{\beta})$ and the association parameters are functions of $\boldsymbol{\alpha}$. Hence, $\nu_{ijk} = \nu_{ijk}(\boldsymbol{\beta}, \boldsymbol{\alpha})$.

The expectations of these over the conditional distribution of the unobserved outcomes given the observed ones are further required. Evidently, because Eqs. (8.3)–(8.4) are linear in the triplet y_{ij}, y_{ik} and $y_{ij}y_{ik}$, it suffices to calculate the expectations over these. Their corresponding probabilities are

$$\nu_{ij|\overline{d}} = \frac{\nu_{i\overline{d}j}}{\nu_{i\overline{d}}} \qquad \text{and} \qquad \nu_{ijk|\overline{d}} = \frac{\nu_{i\overline{d}jk}}{\nu_{i\overline{d}}}, \tag{8.5}$$

where $\overline{d}$ refers to the set of indices $(1, 2, \ldots, d-1)$, corresponding to the observed portion of $\boldsymbol{y}$.

Combining Eqs. (8.3) and (8.4) with Eq. (8.5) leads to

$$\begin{aligned} E(\boldsymbol{U}_{ijk}) &= \frac{\nu_{i\overline{d}jk}}{\nu_{i\overline{d}}\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\theta}}\nu_{ijk} + \frac{\nu_{i\overline{d}j}-\nu_{i\overline{d}jk}}{\nu_{i\overline{d}}(\nu_{ij}-\nu_{ijk})}\frac{\partial}{\partial\boldsymbol{\theta}}(\nu_{ij}-\nu_{ijk}) \\ &\quad + \frac{\nu_{i\overline{d}k}-\nu_{i\overline{d}jk}}{\nu_{i\overline{d}}(\nu_{ik}-\nu_{ijk})}\frac{\partial}{\partial\boldsymbol{\theta}}(\nu_{ik}-\nu_{ijk}) \\ &\quad + \frac{\nu_{i\overline{d}}-\nu_{i\overline{d}j}-\nu_{i\overline{d}k}+\nu_{i\overline{d}jk}}{\nu_{i\overline{d}}(1-\nu_{ij}-\nu_{ik}+\nu_{ijk})}\frac{\partial}{\partial\boldsymbol{\theta}}(1-\nu_{ij}-\nu_{ik}+\nu_{ijk}) \qquad (8.6) \end{aligned}$$

and

$$\begin{aligned} E(\boldsymbol{U}_{ik|j}) &= \frac{y_{ij}\nu_{i\overline{d}k}\nu_{ij}}{\nu_{i\overline{d}}\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{\nu_{ijk}}{\nu_{ij}}\right) + \frac{y_{ij}(\nu_{i\overline{d}}-\nu_{i\overline{d}k})\nu_{ij}}{\nu_{i\overline{d}}(\nu_{ij}-\nu_{ijk})}\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{\nu_{ij}-\nu_{ijk}}{\nu_{ij}}\right) \\ &\quad + \frac{(1-y_{ij})\nu_{i\overline{d}k}(1-\nu_{ij})}{\nu_{i\overline{d}}(\nu_{ik}-\nu_{ijk})}\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{\nu_{ik}-\nu_{ijk}}{1-\nu_{ij}}\right) \\ &\quad + \frac{(1-y_{ij})(\nu_{i\overline{d}}-\nu_{i\overline{d}k})(1-\nu_{ij})}{\nu_{i\overline{d}}(1-\nu_{ij}-\nu_{ik}+\nu_{ijk})}\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{1-\nu_{ij}-\nu_{ik}+\nu_{ijk}}{1-\nu_{ij}}\right). \qquad (8.7) \end{aligned}$$

8.2.3 Predictive Bahadur model in the Doubly Robust Estimating Equations

Many of the probabilities in the predictive model, i.e., the ones involving $\overline{d}$, in (8.6)–(8.7) are of dimension 3 or higher. The calculation of the probabilities in the multivariate Bahadur model is cumbersome because of the very constrained parameter space. Pairwise PL is used exactly to circumvent this problem. In the spirit of, among others, Bang & Robins (2005), we follow a more pragmatic route and propose a convenient and sufficiently rich predictive model. An attractive option is the pairwise Bahadur model, pertaining to response at occasions j and k, but where the history, corresponding to $\overline{d}$, is included as a set of predictor variables. This amounts to using

$$\begin{aligned} E(\boldsymbol{U}_{ijk}) &\equiv E\left[\boldsymbol{U}_i(y_{ij},y_{ik})\right] \\ &= \sum_{y_{ij}=0}^{1}\sum_{y_{ik}=0}^{1}\boldsymbol{U}_i(y_{ij},y_{ik})q(y_{ij},y_{ik}), \qquad (8.8) \end{aligned}$$

$$E(\boldsymbol{U}_{ik|j}) \equiv E\left[\boldsymbol{U}_i(y_{ik}|y_{ij})\right] = \sum_{y_{ik}=0}^{1}\boldsymbol{U}_i(y_{ik}|y_{ij})q(y_{ik}|y_{ij}), \qquad (8.9)$$

where $q(y_{ij},y_{ik}) = P(Y_{ij}=y_{ij},Y_{ik}=y_{ik}|Y_{i\overline{d}}=y_{i\overline{d}})$ and $\boldsymbol{U}_i(y_{ij},y_{ik})$ and $\boldsymbol{U}_i(y_{ik}|y_{ij})$ are as defined in Eqs. (8.3) and (8.4). Evidently, modeling the $q(\cdot)$ terms, will imply the need for an additional parameter vector, $\boldsymbol{\phi}$, say.

8.2.4 Precision Estimation

In the naive case, uncertainty stems from the $\boldsymbol{\theta}$ parameter only. The asymptotic variance-covariance matrix in Eq. (2.5) can then be consistently estimated by $\widehat{I}_0^{-1}\widehat{I}_1\widehat{I}_0^{-1}$, with

$$I_0 = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\theta}} \quad \text{and} \quad I_1 = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{S}_i(\widehat{\boldsymbol{\theta}})\boldsymbol{S}_i'(\widehat{\boldsymbol{\theta}}), \tag{8.10}$$

where $\boldsymbol{U} = \sum_{i=1}^{N}\boldsymbol{V}_i(\boldsymbol{\theta})$ and $\boldsymbol{S}_i(\widehat{\boldsymbol{\theta}}) = \boldsymbol{V}_i$ is the corresponding estimating function, i.e., shorthand notation for the formulas in Table 8.1.

In the singly robust case, we must also take into account uncertainty coming from estimating the $\boldsymbol{\psi}$ parameters in the weight model. The entire score for subject i is $\boldsymbol{S}_i = (\boldsymbol{V}_i', \boldsymbol{W}_i')'$, with $\boldsymbol{W} = \sum_{i=1}^{N}\boldsymbol{W}_i(\boldsymbol{\psi})$ the estimating equations coming from the weight model, and the asymptotic variance-covariance is based on the following matrices:

$$I_0 = \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} \dfrac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\theta}} & \dfrac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\psi}} \\ \\ \mathbf{0} & \dfrac{\partial \boldsymbol{W}_i}{\partial \boldsymbol{\psi}} \end{pmatrix} \quad \text{and} \quad I_1 = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{S}_i(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}})\boldsymbol{S}_i'(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}). \tag{8.11}$$

In the doubly robust case, for the general expression, the weight model is complemented with a predictive model. The score function for this conditional Bahadur model is $\boldsymbol{T}(\boldsymbol{\phi})$, with an extra set of parameters $\boldsymbol{\phi}$. The precision of the parameters can be estimated using the matrices as follows:

$$I_0 = \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} \dfrac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\theta}} & \dfrac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\psi}} & \dfrac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\phi}} \\ \\ \mathbf{0} & \dfrac{\partial \boldsymbol{W}_i}{\partial \boldsymbol{\psi}} & \mathbf{0} \\ \\ \mathbf{0} & \mathbf{0} & \dfrac{\partial \boldsymbol{T}_i}{\partial \boldsymbol{\phi}} \end{pmatrix} \quad \text{and} \quad I_1 = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{S}_i(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\phi}})\boldsymbol{S}_i'(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\phi}}), \tag{8.12}$$

From Eq. (8.1), (8.12) can be simplified to the following expressions

$$I_0 = \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} \dfrac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\theta}} & \dfrac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\phi}} \\ \\ \mathbf{0} & \dfrac{\partial \boldsymbol{T}_i}{\partial \boldsymbol{\phi}} \end{pmatrix} \quad \text{and} \quad I_1 = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{S}_i(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\phi}})\boldsymbol{S}_i'(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\phi}}). \tag{8.13}$$

More detailed calculations and complete formulas can be found in Appendix E.2. See also Bang & Robins (2005) and Rotnitzky (2009).

8.3 Application: The Analgesic Trial

In this section, we apply the proposed methodology to data from a clinical trial designed to investigate an analgesic drug, see Section 3.4. All analyses have been performed with SAS (version 9.4). First, the Bahadur model, using three different estimating equations for CC, CP and AC, was fitted with an NLMIXED procedure. To make use of NLMIXED's functionality, an objective function is formulated of which the first derivative coincides with the estimating function under consideration. For optimization, the default Quasi-Newton technique was applied. Further, to estimate the precision, a sandwich-type robust variance estimator was used and, to perform the calculations, the IML procedure was implemented. The Bahadur model, based on the full likelihood, was again fitted in an NLMIXED procedure with similar settings. For more details, see Appendix E.3.

8.3.1 Results

For all ensuing analyses of the analgesic trial data, we consider only completers and dropouts, i.e., a subset of 328 patients from the original data set, and the dichotomized outcome (GSABIN). We first build a logistic regression for the dropout indicator, in terms of the previous outcome $(y_{i,j-1})$ and pain control assessment at baseline (x_i), i.e.,

$$\text{logit } P(D_i = j | D_i \geq j, x_i, y_{i,j-1}) = \psi_0 + \psi_x x_i + \psi_{prev} y_{i,j-1}.$$

The highly significant p-value $(p < .0001)$ for the parameter ψ_{prev} corresponding to the previous outcome provides evidence against MCAR in favor of MAR. Weights are then calculated based on predicted probabilities from this logistic model.

Preliminary analyses have indicated that, among a set of potential covariates, the linear and quadratic effects of time t_{ij}, as well as the effect of baseline pain control assessment (PCA_0, denoted x_i) are of importance. The marginal regression model for the dichotomized GSA score, GSABIN, denoted as Y, is thus specified as

$$\text{logit } P(Y_{ij} = 1 | t_{ij}, x_i) = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 x_i. \tag{8.14}$$

For the correlation across the within-subject outcomes, we posit a Toeplitz type correlation structure:

$$\begin{pmatrix} 1 & \rho^{(1)} & \rho^{(2)} & \rho^{(3)} \\ \rho^{(1)} & 1 & \rho^{(1)} & \rho^{(2)} \\ \rho^{(2)} & \rho^{(1)} & 1 & \rho^{(2)} \\ \rho^{(3)} & \rho^{(2)} & \rho^{(1)} & 1 \end{pmatrix}, \tag{8.15}$$

where $\rho^{(k)},\ k = 1, 2, 3$ denotes the correlation between outcomes that are k time points apart. Hence, the Bahadur density is $f(\boldsymbol{y}_i) = f_1(\boldsymbol{y}_i)c(\boldsymbol{y}_i)$, with $f_1(\boldsymbol{y}_i)$ as in Eq. (2.2)

with $n_i = 4$ and Eq. (2.3) taking the specific form:

$$\begin{aligned} c(\boldsymbol{y}_i) &= 1 + \sum_{j_1<j_2; j_2-j_1=k} \rho^{(k)}_{ij_1j_2} e_{ij_1} e_{ij_2}, \\ &= 1 + \rho^{(1)} \left(e_{i1}e_{i2} + e_{i2}e_{i3} + e_{i3}e_{i4}\right) + \rho^{(2)} \left(e_{i1}e_{i3} + e_{i2}e_{i4}\right) + \rho^{(3)} e_{i1}e_{i4}. \end{aligned}$$

The resulting parameter estimates, along with corresponding standard errors, for model specification Eq. (8.14), with a Toeplitz correlation structure (Eq. 8.15), using full likelihood and estimating equations from Table 8.1 are presented in Table 8.2. The variability of the estimated weights, or additionally the variability of the estimated parameters of the predictive model, is incorporated in the computation of the standard errors. The high degree of similarity with the results of full likelihood indicate that the extra variability induced by the weights, or additionally by the parameters of the predictive model, does not seem to have a large impact on either the estimates or their standard errors.

Similar results are observed throughout the whole table, but in particular for the parameter estimates under full likelihood, naive AC and the doubly robust cases. Moreover, substantial efficiency over full likelihood seems to be gained under the naive AC and doubly robust approaches. Whereas these observations are not surprising for the doubly robust case, precisely because of their property, the relatively good performance of the naive AC case seems counterintuitive. However, under exchangeability, as shown before, the naive AC can be seen as a doubly robust estimator, given that then the expectation in these estimation equations can be removed because observed and unobserved components from a subject's history are interchangeable. To this effect, we assessed the plausibility of the Toeplitz correlation structure of the analgesic trial data, using full likelihood (approximate F-test in NLMIXED), and determined that the three correlation parameters $\rho^{(k)},\ k = 1, 2, 3$, were not significantly different $(p = 0.9078)$, which implies that the underlying correlation structure might very well be exchangeable. This explains the excellent behaviour of the naive AC estimator.

Next, we consider the CP versions, both single and doubly robust. The estimates for the parameters seem reasonably close to those under full likelihood. In addition, the standard errors under the singly robust case seem comparable, but those of the doubly robust case are generally larger than those from full likelihood, a result that could be attributed to the fact of single robustness. The estimates for the β parameters from the CC cases are somewhat higher, whereas those from the AC cases are lower than those for full likelihood. The CP results seem to fall in between the CC and AC results, suggesting a comprise between the latter two. This can be inferred from the incremental nature of the contributions in expressions in Table 8.1. However, as AC case uses more information than the CP case, this one is generally to be preferred.

Table 8.2: *The Analgesic Trial. Parameter estimates (empirically-corrected standard errors) for naive, singly and doubly robust pairwise likelihood and for full likelihood.*

Effect	Par.	$\boldsymbol{U}_{\text{CC,naive}}$	$\boldsymbol{U}_{\text{CP,naive}}$	$\boldsymbol{U}_{\text{AC,naive}}$	$\boldsymbol{U}_{\text{full.lik.}}$
Inter.	β_0	3.131 (0.678)	2.962 (0.563)	2.590 (0.493)	2.626 (0.509)
Time	β_1	-0.913 (0.492)	-0.908 (0.401)	-0.675 (0.354)	-0.602 (0.362)
Time^2	β_2	0.170 (0.096)	0.177 (0.081)	0.151 (0.074)	0.120 (0.076)
PCA_0	β_3	-0.130 (0.132)	-0.125 (0.113)	-0.186 (0.099)	-0.209 (0.106)
corr_1	$\rho^{(1)}$	0.217 (0.069)	0.244 (0.055)	0.259 (0.057)	0.297 (0.063)
corr_2	$\rho^{(2)}$	0.199 (0.075)	0.234 (0.068)	0.250 (0.069)	0.293 (0.074)
corr_3	$\rho^{(3)}$	0.224 (0.102)	0.232 (0.104)	0.240 (0.104)	0.337 (0.117)
Effect	Par.	$\boldsymbol{U}_{\text{CC,sr}}$	$\boldsymbol{U}_{\text{CP,sr}}$	$\boldsymbol{U}_{\text{AC,sr}}$	
Inter.	β_0	3.090 (0.637)	2.712 (0.552)	1.718 (0.560)	
Time	β_1	-0.997 (0.468)	-0.775 (0.389)	-0.280 (0.347)	
Time^2	β_2	0.193 (0.090)	0.151 (0.078)	0.092 (0.070)	
PCA_0	β_3	-0.195 (0.133)	-0.167 (0.113)	-0.196 (0.115)	
corr_1	$\rho^{(1)}$	0.263 (0.079)	0.295 (0.062)	0.333 (0.064)	
corr_2	$\rho^{(2)}$	0.257 (0.086)	0.273 (0.076)	0.303 (0.076)	
corr_3	$\rho^{(3)}$	0.295 (0.115)	0.298 (0.112)	0.299 (0.108)	
Effect	Par.	$\boldsymbol{U}_{\text{CC,dr}}$	$\boldsymbol{U}_{\text{CP,dr}}$	$\boldsymbol{U}_{\text{AC,dr}}$	
Inter.	β_0	3.577 (1.136)	2.736 (0.874)	1.533 (0.692)	
Time	β_1	-1.333 (0.851)	-0.785 (0.647)	-0.104 (0.480)	
Time^2	β_2	0.241 (0.164)	0.149 (0.132)	0.052 (0.108)	
PCA_0	β_3	-0.196 (0.220)	-0.153 (0.193)	-0.197 (0.147)	
corr_1	$\rho^{(1)}$	0.255 (0.118)	0.305 (0.088)	0.366 (0.108)	
corr_2	$\rho^{(2)}$	0.247 (0.165)	0.281 (0.139)	0.338 (0.158)	
corr_3	$\rho^{(3)}$	0.305 (0.276)	0.329 (0.275)	0.350 (0.243)	

8.4 Concluding Remarks

Pseudo-likelihood approaches have become a practical alternative to full likelihood methods, particularly for applications involving complex likelihood forms. In view of the various issues arising from marginally modelling incomplete non-Gaussian longitudinal data, we move away from conditional pseudo-likelihood, and focus on *marginal pseudo-likelihood*, considering the specific case of incomplete longitudinal binary data, as proposed in Molenberghs et al. (2011). While the numerical and computational issues accompanying the likelihood expressions of the marginal model for the binary longitudinal responses are circumvented by means of substituting pairwise pseudo-likelihood expressions for their full likelihood counterparts, the incompleteness in the data is addressed using concepts of inverse probability weighting and predictive terms in the form of expectations, thereby yielding singly and doubly robust estimators. This expands the set of tools available for fitting marginal models to incomplete non-Gaussian longitudinal data.

In this chapter, we assessed the performance of pseudo-likelihood approaches proposed in Molenberghs et al. (2011), in order to provide practical insight into alternative strategies for marginal models for non-Gaussian incomplete longitudinal data. The analysis of the case study demonstrates the feasibility and adequacy of the proposed methodology. Singly robust estimators with correctly specified dropout model and our doubly robust estimators were found to be at least as efficient as direct likelihood methods. Moreover, under full or near exchangeability, the naive available case version is as efficient as the doubly robust estimators, allowing one to invoke double robustness without having to use weights or expectations.

While the situation examined in this chapter focuses on dropout, in principle, the general methodology applies for non-monotone missingness as well; one then has to pay particular attention to the construction of both weights and predictions, and some non-trivial algebraic challenges will emerge. Other possibilities include imputing all missing cases or imputing only non-monotone missing cases to render the missingness monotone and subsequently using pseudo-likelihood on the imputed data sets. Also, while multiple imputation approaches generally prescribe Gaussian type data, variations for non-Gaussian data can be utilized and seem reasonably stable even with model misspecification; see, for instance, Beunckens, Sotto & Molenberghs (2008).

PART III

Conclusion

Chapter 9

General Discussion and Conclusion

9.1 Conclusion

An easy-to-use criterion for incompleteness of minimal sufficient statistics in univariate and multivariate exponential family models was presented. Typically, incompleteness is studied directly by means of the definition, which means that the existence of a non-trivial zero-expectation function needs to be falsified, or that such a function needs to be constructed. The 'Characterization of incompleteness' requires checking the dimension of a minimal sufficient statistic relative to the length of the parameter vector. Whereas the definition can be daunting to use, this criterion turns the assessment of incompleteness into a feasible task.

Clustered data designs with non-constant cluster sizes (random or otherwise) do not admit complete sufficient statistics. Next to this, maximum likelihood estimation must proceed iteratively, as there are no closed-form solutions. This is not a problem in medium sized clusters and cluster sizes. However, when the number of clusters and/or the cluster sizes are small or very large, computations may become challenging.

Pseudo-likelihood and split-sampling (Molenberghs, Verbeke, and Iddi, 2011) were proposed as model strategy and its statistical properties were investigated. In a first attempt, the normal compound-symmetry model was considered. When the cluster size is constant, there is a closed-form solution. 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 these, but the weights depend on the parameters. This suggests the consideration of approximations to these weights, as well as alternative weights. Based on theoretical results and simulations, as well as on real-data analysis, equal weights and so-called approximate optimal weights were found to not perform well.

Iterated optimal and proportional weights show excellent performance. While the former of these two are somewhat more computationally intensive, the latter are simple and parameter-free. One refinement is that for the mean parameter μ 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, $c_k \cdot n_k$.

In the data analysis a slightly expanded setting was considered, where the mean takes the form of a regression function rather than a constant. This is encouraging towards the use of our results in more elaborate settings, as long as some form of exchangeability prevails.

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 would then imply 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, the important special case of a cluster-by-cluster analysis was considered. Encouragingly, 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 thesis, but may be very sensible in a meta-analysis or clinical trials.

Second, an AR(1) model is considered. In the case of AR(1) model with constant cluster size, there are closed-form solutions but no complete minimal sufficient statistics. Returning to unequal cluster sizes, there are, in general, no closed form solutions. Our results show excellent performance of the size-proportional weights, that is through weighting according to the number of measurements in a cluster ($c_k \cdot n_k$), rather than the number of clusters c_k in a subsample, that is, proportional weights. By contrast, the latter are a good choice for the compound-symmetry structure. Under AR(1) they are worse than equal weights. Approximate optimal weights can also be used, but this leads to an estimate of ρ with a large sample variance. In practice, it is convenient and appropriate to use size-proportional weights; these are parameter free and simple to use. Simulations show that this split-sample methodology, in certain (large) settings, computation time can be 1000 to 30,000 times faster than with standard maximum likelihood.

Further, to study associations between binary variables, incorporating classification, the common odds ratio is often used. The odds ratio estimator of Mantel and Haenszel (1959) is well established in epidemiology. It is a weighted estimator, combining information of the different strata, that do not need to be of the same size. It follows neither from the likelihood principle, nor from optimally weighted considerations. Therefore, also the pseudo-likelihood split-sample methodology (Molenberghs, Verbeke, and Iddi, 2011) was

used to compare it against an optimal weighted estimator. While there is no complete sufficient statistic, there nevertheless is a uniformly optimal weighted estimator. Unlike a full likelihood estimator, both the MH as well as the optimal estimator, and the variance thereof, are easy to calculate.

The MH estimator is very efficient for large datasets, and in more cases well-defined when cell counts are small or even zero. When datasets are huge, the optimal estimator is somewhat more efficient. However, two reservations should be mentioned. First, this is not the case when the number of strata becomes equal to or larger than the mean stratum size; here, the estimator fails. Second, with really huge samples, a slight amount of efficiency loss is usually not an issue. The MH retains its practical and theoretical attraction, even when compared with formally optimal estimators.

Last, alternative strategies for marginal models for non-Gaussian incomplete longitudinal data were the focus. Next to GEE, the performance of pseudo-likelihood approaches, proposed in Molenberghs et al. (2011), were assessed. This is a very feasible and adequate methodology. Singly robust estimators with correctly specified dropout model and our doubly robust estimators were found to be at least as efficient as direct likelihood methods. Especially, under full or near exchangeability, the naive available case version is as efficient as the doubly robust estimators, allowing one to invoke double robustness without having to use weights or expectations.

9.2 Further Research

In addition to the research presented in this thesis, a lot of extra interesting paths which can be followed, were identified.

When clusters become very large, it may become attractive to further sub-divide them in sub-clusters. Such a splitting method was also considered by Molenberghs, Verbeke, and Iddi (2011), i.e. splitting in dependent subsamples. Here, only independent subsamples were considered, so this would require further investigation.

In the analysis of the NTP data and PANSS data, the observed cluster size seemed related to the outcome of interest or another variable under investigation. This suggests that it is useful to consider, at the same time, the impact on the cluster size. This brings us back to the informative cluster sizes mentioned in the introduction, Chapter 1. While work has been done in this area, it is of interest to extend the ideas developed in this thesis to a joint model including cluster size.

As the focus in this work was on normal distributed outcomes, 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. An extension to the methodology could be investigated.

Finally, the MH also knows extensions to $I \times J \times K$ tables (Agresti, 2002, pp. 295). As categorical variables may take on more than 2 levels and also here associations can be of scientific interest, one might also consider here an extension of the investigation for optimal weights.

Bibliography

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

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

Agresti, A. (2002). *Categorical Data Analysis*. Second edition. New Jersey: Wiley series in probability and statistics.

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

Anscombe, F.J. (1949). Large-sample theory of sequential estimation. *Biometrika*, **36**, 455–458.

Armitage, P. (1975). *Sequential Medical Trials*. Oxford: Blackwell.

Arnold, B.C. and Strauss, D. (1991). Pseudolikelihood estimation: some examples. *Sankhya: the Indian Journal of Statistics - Series B*, **53**, 233–243.

Bahadur, R. R. (1961). A representation of the joint distribution of responses of n dichotomous items. In Solomon, H. (Ed.) *Studies in Item Analysis and Prediction*, Stanford University Press, California, 158–168.

Bang, H. & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, **61**, 962–972.

Barndorff-Nielsen, O.E. and Cox, D.R. (1984). The effect of sampling rules on likelihood statistics. *International Statistical Review*, **52**, 309–326.

Barndorff-Nielsen, O.E. and Cox, D.R. (1994). *Inference and Asymptotics.* London: Chapman&Hall.

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.

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.

Beunckens, C., Sotto, C., & Molenberghs, G. (2008). A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. *Computational Statistics and Data Analysis*, **52**, 1533–1548.

Birhanu, T., Molenberghs, G., Sotto, C., & Kenward, M. G. (2011). Doubly robust and multiple-imputation-based generalized estimating equations. *Journal of Biopharmaceutical Statistics*, **21**, 202–225.

Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation. *Annals of Mathematical Statistics*, **18**, 105–110.

Boos D.D. and Stefanksi L.A., (2013). *Essential Statistical Inference: Theory and Methods*. New-York: Springer.

Breslow, N.E. and Day, N.E. (1980). *Statistical Methods in Cancer Research*. Volume 1 - The analysis of case-control studies. Lyon: International Agency for Research on Cancer.

Brown L.D., (1986). *Fundamentals of Statistical Exponential Families: with Applications in Statistical Decision Theory*. Institute of Mathematical Statistics.

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

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

Carey, V. C., Zeger, S. L., & Diggle, P. J. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, **80**, 517–526.

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

Chakraborty, S. (2015). Generating discrete analogues of continuous probability distributions – A survey of methods and constructions. *Journal of Statistical Distributions and Applications*, **2**, 6.

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

Cochran, W. G. (1977). *Sampling Techniques*, John Wiley & Sons, New York.

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

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

Dale, J. R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics*, **42**, 909–917.

Diggle, P. J., Heagerty, P., Liang, K. Y., & Zeger, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed., Oxford University Press, New York.

Emerson, S.S. and Fleming, T.R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika*, **77**, 875–892.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd ed., Springer-Verlag, New York.

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.

Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (Eds.) (2009). *Longitudinal Data Analysis*, Chapman & Hall/CRC, Florida.

Flanders D.W. (1985). A new variance estimator for the Mantel-Haenszel odds ratio. *Biometrics*, **41**, 637–642.

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

Geys, H., Molenberghs, G., and Lipsitz, S.R. (1998). A note on the comparison of pseudo-likelihood and generalized estimating equations for marginal odds ratio models. *Journal of Statistical Computation and Simulation*, **62**, 45–72.

Geys, H., Molenberghs, G., & Ryan, L. (1999). Pseudo-likelihood modelling of multivariate outcomes in developmental toxicology. *Journal of the American Statistical Association*, **94**, 734–745.

Govindarajulu, Z. (1981). *The Sequential Statistical Analysis of Hypothesis Testing, Point and Interval Estimation, and Decision Theory*. Colombus, OH: American Sciences Press.

Hauck, W.W. (1979). The Large Sample Variance of the Mantel-Haenszel Estimator of a Common Odds Ratio. *Biometrics*, **35**, 817–819.

He, W. & Yi, G. Y. (2011). A pairwise likelihood method for correlated binary data with/without missing observations under generalized partially linear single-index models. *Statistica Sinica*, **21**, 207–229.

Hermans, L., Molenberghs G., Aerts, M., Kenward, M.G. and Verbeke, G. (2018). A Tutorial on the practical use and implication of complete sufficient statistics. *International Statistical Review*, **86(3)**, 403–414.

Hermans, L., Nassiri, V., Molenberghs, G., Kenward, M.G., Van der Elst, W., Aerts, M., and Verbeke, G. (2018). Clusters with unequal size: maximum likelihood versus weighted estimation in large samples. *Statistica Sinica*, **28(3)**, 1107-1132.

Hermans, L., Nassiri, V., Molenberghs, G., Kenward, M.G., Van der Elst, W., Aerts, M., and Verbeke, G. (2018). Fast, closed-form, and efficient estimators for hierarchical models with AR(1) covariance and unequal clusters sizes. *Communications in Statistics: Simulation and Computation*, **47(5)**, 1492–1505.

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

Hughes, M.D. and Pocock, S.J. (1988). Stopping rules and estimation problems in clinical trials. *Statistics in Medicine*, **7**, 1231–1242.

Intego. Department of general practice, KU Leuven. Intego-project. [Online]. 2011 [cited 2018 04 07]; Available from: URL:http://www.intego.be.

Jennison, C. and Turnbull, B.W (2000). *Group Sequential Methods With Applications to Clinical Trials*. London: Chapman & Hall/CRC.

Keener, R.W. (2010). *Theoretical Statistcis: Topics for a Core Course, Springer Texts in Statistics.* Springer 85–99.

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

Kenward, M.G. and Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science*, **13**, 236–247.

Kenward, M. G. & Molenberghs, G. (2009). Last observation carried forward: A crystal ball? *Journal of Biopharmaceutical Statististics*, **19**, 872–888.

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

Kuritz, S.J., Landis, J.R. and Koch, G.G. (1988). A general overview of Mantel-Haenszel methods: application recent developments. *Annual Reviews Public Health*, **9**, 123–160.

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

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

le Cessie, S. & van Houwelingen, J. C. (1991). A goodness-of-fit test for binary regression models, based on smoothing methods. *Biometrics*, **47**, 1267–1282.

le Cessie, S. and van Houwelingen, J.C. (1994). Logistic regression for correlated binary data. *Applied Statistics*, **43**, 95–108.

Lehmann, E.L., (1981). *Testing statistical hypothesis*. New York: John Wiley & Sons.

Lehmann, E.L. and Stein, C. (1950). Completeness in the sequential case. *Annals of Mathematical Statistics*, **21**, 376–385.

Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Liang, K.ælJř., Zeger, S.L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B*, **54**, 3–40.

Lindsay, B. (1988), Composite likelihood methods. *Contemporary Mathematics*, **80**, 220–239.

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.

Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

Mantel, N., and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, **22–4**, 719–748.

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

McCullagh, P., Nelder, J.A. (1989). *Generalized Linear Models.* London New York: Chapman and Hall.

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

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

Molenberghs, G. and Kenward, M.G. (2007). *Missing Data in Clinical Studies*. Chichester: John Wiley & Sons.

Molenberghs, G., Kenward, M. G. (2010). Semi-parametric marginal models for hierarchical data and their corresponding full models. *Computional Statistics & Data Analysis*, **54**, 585–597.

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., Kenward, M. G., Verbeke, G., & Birhanu, T. (2011a). Pseudo-likelihood estimation for incomplete data. *Statistica Sinica*, **21**, 187–206.

Molenberghs, G. & Lesaffre, E. (1994). Marginal modelling of correlated ordinal data using a multivariate Plackett distribution. *Journal of the American Statistical Association*, **89**, 633–644.

Molenberghs, G. & Lesaffre, E. (1999). Marginal modelling of multivariate categorical data. *Statistics in Medicine*, **18**, 2237–2255.

Molenberghs, G., Verbeke, G., and Iddi, S. (2011b). Pseudo-likelihood methodology for partitioned large and complex samples. *Statistics & 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

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

Parzen, M., Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J.G., & Troxel, A. (2006). Pseudo-likelihood methods for longitudinal binary data with non-ignorable missing responses and covariates. *Statistics in Medicine*, **25**, 2784–2796.

Pattanayak, C.W., Rubin, D.B and Zell, E.R. (2012). A potential outcomes, and typically more powerful, alternative to "Cochran-Mantel-Haenszel". *SSRN*.

Poularikas, A.D. and Seely, S. (2000). Laplace transforms. *The Transforms and Applications Handbook*. Boca Raton: Chapman & Hall/CRC Press

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.

Renard, D., Molenberghs, G., & Geys, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics and Data Analysis*, **44**, 649–667.

Robins, J., Breslow, N. and Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics*, **42(2)**, 311–232.

Robins, J., Greenland, S. and Breslow, N. (1986). A general estimator for the variance of the Mantel-Haenszel odds ratio. *American Journal of Epidemiology*, **124(5)**, 719–723.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**, 106–121.

Rosner, G.L. and Tsiatis, A.A. (1988). Exact confidence intervals following a group sequential trial: A comparison of methods. *Biometrika*, **75**, 723–729.

Rotnitzky, A. (2009). Inverse probability weighted methods. In Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (Eds.) *Longitudinal Data Analysis*, Chapman & Hall/CRC, Florida, 453–476.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

Rudin, W. (1974). *Real & Complex Analysis* (2nd ed.). New Delhi: McGraw Hill.

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semi-parametric nonresponse models. *Journal of the American Statistical Association*, **94**, 1096ÂŰ-1120 (with Rejoinder, 1135–1146).

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**.

Shao, T. (1999). *Mathematical Statistics (2nd ed.)*. New York: Springer.

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

Siegmund, D. (1978). Estimation following sequential tests. *Biometrika*, **64**, 191–199.

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.

Todd, S., Whitehead, J., and Facey, K.M. (1996). Point and interval estimation following a sequential clinical trial. *Biometrika*, **83**, 453–461.

Truyers, C., Goderis., G., Dewitte, H., vanden Akker, M. and Buntinx, F. (2014). The Intego database: background, methods and basic results of a Flemish general practice-based continuous morbidity registration project. *BMC Medical Informatics and Decision Making*. **14(1)**, 48.

Tsiatis, A.A., Rosner, G.L., and Mehta, C.R. (1984). Exact confidence intervals following a group sequential test. *Biometrics*, **40**, 797–803.

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.

Van der Laan, M. J. & Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*, Springer-Verlag, New York.

Van Garderen, K.J. (1997). Curved exponential models in econometrics. *Econometric Theory*, **1**, 771–790.

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(11)**, 1–17.

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

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

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, **16**, 117–186.

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.

Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials (2nd ed.)*. New York: John Wiley & Sons.

Whitehead, J. (1999). A unified theory for sequential clinical trials. *Statistics in Medicine*, **18**, 2271–2286.

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

Woolf, B. (1955). On estimating the relation between blood group and disease. *Annals of human genetics*, **19**, 251–253.

Yi, G. Y., Zeng, L., & Cook, R. J. (2011). A robust pairwise likelihood method for incomplete longitudinal binary data arising in clusters. *The Canadian Journal of Statistics*, **39**, 34–51.

Appendix

Appendix A

Appendix for Chapter 4

A.1 Examples

Example 1 (Univariate normal sample with known variance)**.** *Let* $Y_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$, *with* μ *unknown and* σ^2 *known. Then* $K_1 = \sum_{i=1}^{n} Y_i$ *is a complete sufficient statistic for* μ.

Clearly, $K_1 \sim N(n\mu, n\sigma^2)$. Suppose that there is a function $g(k_1)$ such that $E\{g(k_1)\} = 0$ for all values of μ. Then

$$\int g(k_1) \frac{1}{\sqrt{n\sigma^2}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{(k_1 - n\mu)^2}{n\sigma^2}\right\} dk_1 = 0,$$

implying that

$$\int g(k_1) \exp\left\{-\frac{1}{2}\frac{k_1^2\sigma^2}{n\sigma^4} + \frac{k_1}{\sigma^2}\mu\right\} d\left(\frac{k_1}{\sigma^2}\right) = 0.$$

With a simple change of variables, this can be written as

$$\int g(t\sigma^2) e^{-\frac{1}{2}\frac{t^2\sigma^2}{n}} e^{t\mu} dt = \mathcal{L}\left\{g(t\sigma^2) e^{-\frac{1}{2}\frac{t^2\sigma^2}{n}}\right\} = 0,$$

where $\mathcal{L}(\cdot)$ denotes the two-sided Laplace transform. This, in turn, implies that the argument must be equal to zero almost everywhere (a.e.). Because of the exponential factor, this forces $g(t\sigma^2) = 0$ a.e. Hence, $g(k_1) = 0$ a.e.

Example 2 (Univariate normal sample with unknown variance)**.** *Let* $Y_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$, *with* μ *and* σ^2 *unknown. Then* (K_1, K_2) *with* K_1 *as in Example 1 and* $K_2 = \sum_{i=1}^{n} Y_i^2$ *is a complete sufficient statistic for* (μ, σ^2).

The kernel of the log-likelihood, i.e., the terms of the log-likelihood that are functions of the parameters (Mccullagh and Nelder, 1989), is

$$\begin{aligned}\ell &= -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 \\ &= -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}y_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^{n}y_i - \frac{n\mu^2}{2\sigma^2}. \qquad \text{(A.1)}\end{aligned}$$

The sufficient statistic (K_1, K_2) immediately follows. K_1 is normally distributed as in Example 1 and K_2 has a non-central chi-squared distribution. K_1 and K_2 are independent. Assume a function $g(k_1, k_2)$ with zero expectation for all values of the pair (μ, σ^2).

Even though we have a bivariate statistic, we can start from the derivations in Example 1. Write the kernel of the density of K_m $(m = 1, 2)$ as $h_m(k_m)\exp(\theta_m k_m)$, then the condition on $g(k_1, k_2)$ is:

$$\begin{aligned}0 &= \int\int g(k_1,k_2)h_1(k_1)h_2(k_2)\exp(\theta_1k_1+\theta_2k_2)dk_1dk_2 \qquad \text{(A.2)}\\ &= \int dk_2 h_2(k_2)\exp(\theta_2k_2)\int dk_1 g(k_1,k_2)h_1(k_1)\exp(\theta_1k_1) \qquad \text{(A.3)}\\ &= \int dk_2 h_2(k_2)\exp(\theta_2k_2)\mathcal{L}_{\theta_1}\left\{g(k_1,k_2)h_1(k_1)\right\} \qquad \text{(A.4)}\\ &= \mathcal{L}_{\theta_2}\left[h_2(k_2)\mathcal{L}_{\theta_1}\left\{g(k_1,k_2)h_1(k_1)\right\}\right], \qquad \text{(A.5)}\end{aligned}$$

where $\mathcal{L}_{\theta_1}$ is a two-sided and $\mathcal{L}_{\theta_2}$ a one-sided Laplace transform. This implies that $h_2(k_2)\mathcal{L}_{\theta_1}\{g(k_1,k_2)h_1(k_1)\} = 0$ a.e. and thus, because $h_2(k_2) > 0$ over the support, that $\mathcal{L}_{\theta_1}\{g(k_1,k_2)h_1(k_1)\} = 0$ a.e. This, in turn, implies that $g(k_1,k_2)h_1(k_1) = 0$ a.e. For a reason similar to that used above, it follows that $g(k_1,k_2) = 0$ a.e. Note that a two-sided, or bilateral, Laplace transform is unique only, i.e., one-to-one onto its inverse, when not only the function but also the region of convergence is specified (Poularikas and Seely, 2000). However, in our case, because of the use of exponential family functions, the region of convergence is not restricted, hence this subtle issue does not apply here. In fact, an unrestricted region of convergence is a regularity condition: it is violated, for example, in the case of a deterministic stopping rule, but not when a stochastic stopping rule is used.

This derivation is quite general. Clearly, the argument can be extended to a vector of arbitrary length. Note however that we have used the fact that K_1 and K_2 are independent. While this is true for the mean and the variance related sufficient statistics for normal samples, it is not true in general. However, the extension to dependent sufficient statistics is almost trivial: we can replace $h_1(k_1)$ by $h_1(k_1|k_2)$ in (A.2)–(A.5). Furthermore, a univariate version of this argument generalizes Example 1.

The multivariate extension of this argument will be used in Section 4. It is not true however, that such an argument can cover all situations, for example, for the sequential trial case. A simple but instructive counterexample is provided next.

Example 3 (Univariate normal sample with coupled mean and variance)**.** *Let* $Y_i \sim N(\mu, \tau^2\mu^2)$*,* $i = 1, \ldots, n$*, with* μ *unknown. In the case that* τ^2 *is known, there is no complete sufficient statistic for* μ*. On the other hand, when* τ^2 *is unknown, there is a complete sufficient statistic for* (μ, τ^2)*.*

The kernel of the log-likelihood immediately follows from (A.1), upon equating $\sigma^2 = \tau^2\mu^2$:

$$\ell = -\frac{n}{2}(\ln \tau^2 + 2\ln \mu) - \frac{1}{2\tau^2\mu^2}\sum_{i=1}^{n} y_i^2 + \frac{1}{\tau^2\mu}\sum_{i=1}^{n} y_i - \frac{n}{2\tau^2}. \tag{A.6}$$

Assume τ^2 known and consider the function

$$g(k_1, k_2) = \frac{k_1^2}{\tau^2 + n} - \frac{k_2}{\tau^2 + 1}. \tag{A.7}$$

Because $E(K_1^2) = n^2\mu^2 + n\tau^2\mu^2 = n\mu^2(\tau^2 + n)$ and $E(K_2) = n\mu^2(\tau^2 + 1)$, it readily follows that the expectation of $g(K_1, K_2)$ is zero, while the function is non-trivial. Function $g(k_1, k_2)$ satisfies the definition of incompleteness only because τ^2 is a known constant.

The score equation for (A.6) can be written as:

$$n\mu^2 + \frac{K_1\mu}{\tau^2} - \frac{K_2}{\tau^2} = 0,$$

with solution

$$\widehat{\mu} = \frac{-K_1 \pm \sqrt{K_1 + 4n\tau^2}}{2n\tau^2}.$$

Clearly, $\widehat{\mu} + g(K_1, K_2)$ would provide another estimator with the same expectation, for every non-trivial function $g(k_1, k_2)$ with expectation zero. The derivations above show that this type of function exists. Adding such a function comes down to reweighing the amount of information taken from K_1 relative to that from K_2.

This counterexample is interesting because, at first sight, it is close to Examples 1 and 2. However, in both of these earlier examples, the sufficient statistic and the parameter are of the same dimension, while here, the statistic is by necessity two-dimensional. If it is restricted to either K_1 or K_2, then it is no longer sufficient.

But when τ^2 is unknown, the sufficient statistic and the parameter are again of the same dimension as in Example 2. The score for τ^2 leads to:

$$\tau^2 = \frac{2}{n}\left(\frac{K_2}{2\mu^2} - \frac{K_1}{\mu} + \frac{n}{2}\right)$$

and to solutions:

$$\widehat{\mu} = \frac{K_1}{n}, \widehat{\tau^2} = \frac{nK_2}{k_1^2} - 1.$$

This example, with τ^2 known, is similar to the sequential-trial case where the sufficient statistic consists not only of the data collected, but also of the sample size realized, i.e., a one-dimensional parameter needs a two-dimensional sufficient statistic. The following example sets this out in some generality. It is based on developments in Milanzi *et al.* (2015). In this, a group sequential trial was considered with an arbitrary number of looks L and exponential family distributed outcomes. It generalizes the results of Milanzi *et al.* (2016), who only considered a trial with two possible sample sizes, n and $2n$.

Example 4 (Sequential trial with stochastic stopping rule)**.** *Consider a sequential trial with L pre-specified looks, with sample sizes $n_1 < n_2 < \cdots < n_L$. Assume that there are n_j i.i.d. observations $Y_1, \ldots, Y_{n_j}$, from the jth look that follow an exponential family distribution with density*

$$f_\theta(y) = h(y) \exp\{\theta y - a(\theta)\}, \tag{A.8}$$

for θ the natural parameter, $a(\theta)$ the mean generating function, and $h(y)$ the normalizing constant. There is no complete sufficient statistic for the mean μ or, equivalently, for the natural parameter θ.

Subsequent developments are based on a generic data-dependent stochastic stopping rule, which we write as:

$$\pi(N = n_j | k_{n_j}) = F\left(k_{n_j} \middle| \boldsymbol{\psi}\right) = F\left(k_{n_j}\right), \tag{A.9}$$

where $k_{n_j} = \sum_{i=1}^{n_j} y_i$ is a realisation from an exponential family density:

$$f_{n_j}(k) = h_{n_j}(k) \exp\left\{\theta k_{n_j} - n_j a(\theta)\right\}. \tag{A.10}$$

While we do not need to provide an explicit expression for the stopping rule at this point, as our developments apply to a broad class, it is useful to note that Milanzi *et al.* (2016) studied in detail the behaviour of stopping rules taking the form $F\left(\alpha + \beta k_{n_j}/n_j^m\right)$, for some power m and some cumulative distribution function $F(\cdot)$. Our inferential target is the parameter θ, or a function thereof.

In a sequential setting, a convenient minimal sufficient statistic is (K_3, N), with $K_3 = \sum_{i=1}^{N} Y_i$. Following the developments in the above papers, the joint distribution for

(K_3, N) is:

$$p(K_3, N) = f_0(K_3, N)\, F(K_N), \tag{A.11}$$

$$f_0(k_{n_1}, n_1) = f_{n_1}(k_{n_1}), \tag{A.12}$$

$$f_0(k_{n_j}, n_j) = \int f_0(k_{n_{j-1}}, n_{j-1}) f_{n_j - n_{j-1}}(k_{n_j} - k_{n_{j-1}}) \left[1 - F(k_{n_{j-1}})\right] dk_{n_{j-1}}. \tag{A.13}$$

If (K_3, N) were complete, then there would exist a function $g(K_3, N)$ such that $E\left[g(K_3, N)\right] = 0$ if and only if $g(K_3, N) = 0$, implying that

$$\begin{aligned} 0 = & \int g(k_{n_1}, n_1) f_{n_1}(k_{n_1}) F(k_{n_1}) dk_{n_1} + \sum_{j=2}^{L-2} \int g(k_{n_j}, n_j) H(k_{n_j}) F(k_{n_j}) dk_{n_j} \\ & + \int g(k_{n_L}, n_L) H(k_{n_L}) F(k_{n_L}) dk_{n_L}, \end{aligned} \tag{A.14}$$

with

$$H(k_{n_j}) = \underbrace{\int \dots \int}_{j-1} f_0(k_{n_{j-1}}, n_{j-1}) f_{n_j - n_{j-1}}(k_{n_j} - k_{n_{j-1}}) \left[1 - F(k_{n_{j-1}})\right] dk_{n_1} \dots dk_{n_{j-1}}.$$

Substituting the general exponential form (A.10) into (A.14), and applying properties of exponential family probability distributions, gives

$$\begin{aligned} 0 = & \int h_{n_L - n_1} e^{(\theta k_{n_1})} \int g(k_{n_1}, n_1) F(k_{n_1}) h_{n_1}(k_{n_1}) \exp(\theta k_{n_1}) dk_{n_1} \\ & + \sum_{j=2}^{L-2} \int h_{n_L - n_j} e^{(\theta k_{n_j})} \int g(k_{n_j}, n_j) \widetilde{H}(k_{n_j}) \exp(\theta k_{n_j} - n_j) F(k_{n_j}) dk_{n_j} \\ & + \int g(k_{n_L}, n_L) \widetilde{H}(k_{n_L}) \exp(\theta k_{n_L}) F(k_{n_L}) dk_{n_L}, \end{aligned} \tag{A.15}$$

where

$$\widetilde{H}(k_{n_j}) = \left[\underbrace{\int \dots \int}_{j-1} \prod_{i=1}^{j-1} h_{n_1}(k_{n_1}) h_{n_{i+1} - n_i}(k_{n_{i+1}} - k_{n_i}) \left[1 - F(k_{n_i})\right] dk_{n_1} \dots dk_{n_{j-1}} \right].$$

Upon noting that the right hand side is a convolution and making use of properties of linearity and uniqueness of the Laplace transform it can be shown that:

$$g(k_{n_L}, n_L) \widetilde{H}(k_{n_L}) = -\sum_{j=1}^{L-1} \int g(z_j, n_j) \widetilde{H}(z_j) F(z_j) dz_j,$$

$$g(k_{n_L}, n_L) = \frac{\sum_{j=1}^{L-1} \int g(z_j, n_j) \widetilde{H}(z_j) F(z_j) dz_j}{\widetilde{H}(k_{n_L})}.$$

Note that the Laplace transform is unique in both the unilateral as well as the bilateral case. In the unilateral case, this property is straightforward. In the bilateral case, the additional requirement needs to be added that this uniqueness holds over the region of absolute convergence. As mentioned earlier, this region of convergence is not restricted as a stochastic stopping rule is applied. Assigning, for example, arbitrary constants to $g(n_1, k_{n_1}), \ldots, g(n_{L-1}, k_{n_{L-1}})$, a value can be found for $g(n_L, k_{n_L}) \neq 0$, contradicting the requirement for (K_3, N) to be complete, hence establishing incompleteness. From applying the Lehmann-Scheffé theorem, no best mean-unbiased estimator is guaranteed to exist. The practical consequence of this is that even estimators as simple as a sample average need careful consideration and comparison with alternatives. To do this, we embed the sample average in a broader class of linear estimator, and also study it from a likelihood perspective.

Consider the special case of $L = 2$, $n_1 = n$, and $n_2 = 2n$, a normally distributed endpoint with mean μ and variance 1, and probit probability of stopping after the first look equal to $\Phi(\alpha + \beta k/n)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Then, following Molenberghs *et al.* (2014), incompleteness is established by constructing a non-trivial function $g(K_3, N)$ (where K_3 is the sample sum and N is the realized sample size, i.e., N can take values n and $2n$), satisfying for all μ:

$$g(k, 2n) \cdot p_0(2n, k) = -\int \phi_n(k - z) \cdot g(z, n) \cdot \phi_n(z) \cdot \Phi\left(\alpha + \frac{\beta}{n} z\right) dz, \quad \text{(A.16)}$$

where $\phi(\cdot)$ is the standard normal density. Molenberghs *et al.* (2014) gave two examples of such a function, one of which being:

$$g(k, n) = \frac{\lambda}{\Phi\left(\alpha + \frac{\beta}{n} k\right)}, \quad \text{(A.17)}$$

$$g(k, 2n) = -\frac{\lambda}{1 - \Phi\left(\frac{\alpha + \frac{\beta k}{2n}}{\sqrt{\frac{2n + \beta^2}{2n}}},\right)}, \quad \text{(A.18)}$$

with $\lambda \neq 0$ an arbitrary constant.

Example 5 (Known parameter)**.** *Consider the bivariate case studied in Example 2 but now such that θ_2 is known.*

The requirement for an expectation-zero function is:

$$\int dk_2 h_2(k_2) e^{-A(\theta_2)} e^{\theta_2 k_2} \int dk_1 g(k_1, k_2) h_1(k_1|k_2) e^{-A(\theta_1, k_2)} e^{\theta_1 k_1} = 0. \tag{A.19}$$

Choose $g(k_1, k_2) = g(k_2)$. Condition (A.19) becomes:

$$\begin{aligned} & e^{-A(\theta_2)} \int dk_2 h_2(k_2) e^{\theta_2 k_2} g(k_2) \int dk_1 h_1(k_1|k_2) e^{-A(\theta_1, k_2)} e^{\theta_1 k_1} \\ = \quad & e^{-A(\theta_2)} \int dk_2 h_2(k_2) e^{\theta_2 k_2} g(k_2) = 0. \end{aligned} \tag{A.20}$$

Hence, we merely need to satisfy:

$$\int g(k_2) h_2(k_2) e^{\theta_2 k_2} dk_2 = 0. \tag{A.21}$$

Importantly, because θ_2 is known, the left hand side of (A.21) is not a Laplace transform. Interpreting (A.21) as an inner product, we need only find a function $g(k_2) \perp h_2(k_2) e^{\theta_2 k_2}$, which is straightforward.

Example 6 (Univariate normal sample with identical mean and variance)**.** *Let $Y_i \sim N(\mu, \mu)$, $i = 1, \ldots, n$. Then $K_2 = \sum_{i=1}^n Y_i^2$ is a complete sufficient statistic for μ.*

This example is surprisingly different from Example 3, because now the kernel of the log-likelihood is:

$$\ell = -\frac{n}{2} \ln \mu - \frac{1}{2\mu} \sum_{i=1}^n y_i^2 - \frac{n\mu}{2},$$

so K_1 disappears. We clearly have a scalar sufficient statistic, and completeness is trivial. Note that the score equation takes the simple form

$$\mu^2 + \mu = \frac{K_2}{n},$$

leading to the maximum likelihood estimator:

$$\widehat{\mu} = \frac{\sqrt{4K_2/n + 1} - 1}{2}.$$

In Example 4, the conditional likelihood accommodating both K_3 and N has a non-linear correction term relative to the ordinary least squares solution to the likelihood equations in the standard case of a fixed sample size (Molenberghs *et al.*, 2014; Milanzi *et al.*, 2016, 2015).

Example 7 (Univariate normal sample with general coupling of mean and variance)**.** *Let* $Y_i \sim N(\mu, \mu^{2\lambda})$*,* $i = 1, \ldots, n$*. Then there is a complete sufficient statistic for* μ *only for* $\lambda = 0$ *or* $\lambda = 1/2$*.*

When $\lambda = 0$, Example 1 is recovered. Example 6 follows for $\lambda = 1/2$. In all other cases, the sufficient statistic is bivariate, which follows from the kernel of the log-likelihood:

$$\ell(\mu) \propto -n\lambda \ln \mu - \frac{K_2}{2\mu^{2\lambda}} + \frac{K_1}{\mu^{2\lambda-1}} - \frac{n}{2\mu^{2\lambda-2}}.$$

Given that $K_1 \sim N(n\mu, n\mu^{2\lambda})$ and $K_2 \sim \chi^2_{n\mu^{2\lambda}}$, it follows that $E(K_1) = n\mu$, $E(K_2) = 2n\mu^{2\lambda}$, and $E(K_1^2) = n^2\mu^2 + n\mu^{2\lambda}$. Consider a function

$$g(k_1, k_2) = \alpha k_1 + \beta k_1^2 + \gamma k_2. \tag{A.22}$$

The expectation is

$$E\left\{g(K_1, K_2)\right\} = \alpha n\mu + \beta n^2\mu^2 + (\beta n + 2\gamma n)\mu^{2\lambda}.$$

When $\lambda = 1$ every choice $\gamma = -\beta(n+1)/2$ produces a non-zero function with zero expectation. For $\lambda \neq 1$, in addition to being different from 0 and $1/2$ as well of course, there is no non-trivial solution. However, from Proposition 1, we know that for all $\lambda \neq 0, 1/2$, the sufficient statistic is incomplete. So it is seen that it is not because the posited function (A.22) fails to provide a counterexample that there exists none. We now know there are such functions, but the proposition obviates the need to explicitly construct one.

Next, we provide an additional example, using a contingency table.

Example 8 (Bivariate contingency table)**.** *Consider an* $M_1 \times M_2$ *contingency table with conditional row probabilities* $\varphi(k_1|k_2)$ *and marginal column probabilities* $\pi(k_2)$*.*

First, assume that all probabilities are unknown and to be estimated. Assume that there is a function $g(k_1, k_2)$ with zero expectation. Then

$$\sum_{k_1=1}^{M_1} \sum_{k_2=1}^{M_2} g(k_1, k_2)\varphi(k_1|k_2)\pi(k_2) = 0, \tag{A.23}$$

with sum constraints on the parameters: $\sum_{k_2=1}^{M_2} \pi(k_2) = 1$ and $\sum_{k_1=1}^{M_1} \varphi(k_1|k_2) = 1$, for every value of k_2. Because (A.23) should hold for all values of the parameters, $g(k_1, k_2) = 0$ follows immediately from algebraic results on polynomials.

Second, assume that $\pi(k_2)$ is given and choose $g(k_1, k_2) = g(k_2)$. Then, (A.23) simplifies to

$$\sum_{k_1=1}^{M_1} \sum_{k_2=1}^{M_2} g(k_2)\varphi(k_1|k_2)\pi(k_2) = \sum_{k_2=1}^{M_2} g(k_2)\pi(k_2) = 0.$$

Because the vector $\boldsymbol{\pi}$ is given, we merely need a set of constants $\boldsymbol{g}$ such that $\boldsymbol{g} \perp \boldsymbol{\pi}$.

Example 9 (Univariate outcomes with random sample size)**.** *Consider* $Y_i \sim N(\mu, 1)$*, with sample size* N*, with* $1 \leq N \leq n$ *and the probability of realizing sample size* N *equal to* π_N*. The sufficient statistic for* μ *is incomplete.*

The sufficient statistic is (K_3, N) with $K_3 = \sum_{i=1}^{N} Y_i$ and N the usual sample size. Assume that all $\pi_N > 0$, for $N = 1, \ldots, n$; this simplifies the calculations without loss of generality. Choose a function $g(k, N) = a_N$. It then follows that

$$\begin{aligned} E\{g(K_3, N)\} &= \int \sum_{N=1}^{n} g(k, N)\pi_N \phi(k; N\mu, N)dk \\ &= \sum_{N=1}^{n} a_N \pi_N \int \phi(k; N\mu, N)dk \\ &= \sum_{N=1}^{n} a_N \pi_N. \end{aligned}$$

This expectation equals zero if a vector $\boldsymbol{a} \perp \boldsymbol{\pi}$ is chosen. Choosing $(a_1, \ldots, a_{n-1})$ freely, then

$$a_n = -\frac{1}{\pi_n} \sum_{N=1}^{n-1} a_N \pi_N$$

satisfies the requirement. In the next example, we consider clustering between the outcomes.

Example 10 (Correlated outcomes with compound-symmetry structure and random sample size)**.** *The setting is similar to that of Example 9, except that the vector* $\boldsymbol{Y} \sim N(\mu 1_N, \sigma^2 I_N + \tau^2 J_N)$*, with* 1_N *a vector of ones of length* N*,* I_N *the* N*-dimensional identity matrix, and* J_N *an* $N \times N$ *matrix of ones. The sufficient statistic for* (μ, σ^2, τ^2) *is incomplete.*

The sufficient statistic is $(K_3 = \sum_{i=1}^{N} Y_i, K_4 = \boldsymbol{Y}'\boldsymbol{Y}, K_5 = \boldsymbol{Y}' J_N \boldsymbol{Y}, N)$, as will be clear from Example 13. By choosing a function $g(k_1, k_2, k_3, N) = a_N$ the same solution $\boldsymbol{a} \perp \boldsymbol{\pi}$ follows. This result does not depend in any way on this particular normality assumption, as can be formalized in the next example.

Example 11 (Vector-valued data and parameter, with completely random sample size)**.** *Assume an exponential family structure* $f(\boldsymbol{k}, N) = f_N(\boldsymbol{k})\pi(N|\boldsymbol{k}) \overset{\text{notation}}{=} f_N(\boldsymbol{k})\pi_N(\boldsymbol{k})$*. The sufficient statistic is incomplete.*

Choose $g(\boldsymbol{k}, N) = g_N(\boldsymbol{k}) = a_N/\pi_N(\boldsymbol{k})$ for $\pi_N(\boldsymbol{k}) \neq 0$ and 0 otherwise. Then

$$E\{g(K_3, N)\} = \sum_{N=1}^{n} \int f_N(\boldsymbol{k})\pi_N(\boldsymbol{k})g_N(\boldsymbol{k})d\boldsymbol{k} = \sum_{N=1}^{n} a_N \int f_N(\boldsymbol{k})d\boldsymbol{k} = \sum_{N=1}^{n} a_N = 0$$

for any zero-sum sequence.

Of course, by using the term clustered data we do imply that N clusters of sizes N_i $(i = 1, \ldots, m)$ are sampled. We have not considered this level of generality yet. Example 11 will be generalized next.

Example 12 (N clusters of completely random size)**.** *Consider N clusters of sizes N_i $(i = 1, \ldots, N)$, with sufficient statistics $[\boldsymbol{K} = \boldsymbol{K}\{(\boldsymbol{Y}_i)\}; \boldsymbol{N} = \boldsymbol{N}\{(N_i)\}]$. The sufficient statistic is incomplete.*

This result has the same form as in the in the previous example, with $g_{\boldsymbol{N}}(\boldsymbol{k}) = a_{\boldsymbol{N}}/\pi_{\boldsymbol{N}}(\boldsymbol{k})$ this time, and

$$\sum_{\boldsymbol{N}} a_{\boldsymbol{N}} = 0.$$

Example 13 (Compound-symmetry clusters of random size)**.** *Consider clustered data $\boldsymbol{Y}_i \sim N(\mu 1_{N_i}, \sigma^2 I_{N_i} + \tau^2 J_{N_i})$, for $i = 1, \ldots, N$. The sufficient statistic for (μ,σ^2,τ^2) is incomplete.*

The terms in the log-likelihood that are data-dependent, and hence produce the sufficient statistic, follow from

$$\begin{aligned}
&\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} -\frac{\mathbf{1}}{\mathbf{2}}(\boldsymbol{Y_i} - \boldsymbol{\mu 1_{N_i}})'(\boldsymbol{\sigma_2 I_{N_i}} + \boldsymbol{\tau^2 J_{N_i}})^{-1}(\boldsymbol{Y_i} - \boldsymbol{\mu 1_{N_i}}) \\
&= \sum_{i=1}^{N} -\frac{1}{2}(\boldsymbol{Y}_i - \mu 1_{N_i})' \left(I_{N_i} - \frac{\tau^2}{\sigma^2 + N_i\tau^2} J_{N_i} \right) (\boldsymbol{Y}_i - \mu 1_{N_i}) \\
&= \sum_{i=1}^{N} \frac{\mu}{\sigma^2 + N_i\tau^2} \left(\sum_{j=1}^{N_i} Y_{ij} \right) - \frac{1}{2\sigma^2} \left(\sum_{i=1}^{N} \sum_{j=1}^{N_i} Y_{ij}^2 \right) \\
&\quad + \sum_{i=1}^{N} \frac{\tau^2}{2\sigma^2(\sigma^2 + N_i\tau^2)} \left(\sum_{j=1}^{N_i} Y_{ij} \right)^2 . \qquad \text{(A.24)}
\end{aligned}$$

The three terms in (A.24) 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. To properly formalize this, assume that there are L different cluster sizes, and that there are c_ℓ clusters among the data of size n_ℓ. Evidently, $m = \sum_{\ell=1}^{L} c_\ell$. Based on (A.24) and the multiplicity

of the cluster sizes, the sufficient statistics are:

$$S_{1\ell} = \sum_{i=1}^{c_\ell} \sum_{j-1}^{n_\ell} Y_{ij}^{(\ell)}, \tag{A.25}$$

$$S_2 = \sum_{\ell=1}^{L} \sum_{i=1}^{c_\ell} \sum_{j-1}^{n_\ell} \left(Y_{ij}^{(\ell)}\right)^2, \tag{A.26}$$

$$S_{3\ell} = \sum_{i=1}^{c_\ell} \left(\sum_{j-1}^{n_\ell} Y_{ij}^{(\ell)} \right)^2, \tag{A.27}$$

$$S_{4\ell} = c_\ell, \tag{A.28}$$

where the superscript (ℓ) is used to indicate that the summation is restricted to data from clusters of size n_ℓ. The conditional and marginal expectations of (A.25)–(A.28) are:

$$\begin{aligned}
E(S_{1\ell}|c_\ell) &= c_\ell n_\ell \mu, \\
E(S_{1\ell}) &= m\mu\pi_\ell n_\ell, \\
E(S_2|c_\ell) &= \sum_{\ell=1}^{L} c_\ell n_\ell (\sigma^2 + \tau^2 + \mu^2), \\
E(S_2) &= N(\sigma^2 + \tau^2 + \mu^2) \sum_{\ell=1}^{L} \pi_\ell n_\ell, \\
E(S_{3\ell}|c_\ell) &= c_\ell \left\{ n_\ell(\sigma^2 + \tau^2 + \mu^2) + n_\ell(n_\ell - 1)(\tau^2 + \mu^2) \right\}, \\
E(S_{3\ell}) &= m\pi_\ell n_\ell \left\{ (\sigma^2 + \tau^2 + \mu^2) + (n_\ell - 1)(\tau^2 + \mu^2) \right\}, \\
E(S_{4\ell}) &= m\pi_\ell.
\end{aligned}$$

Group all sufficient statistics into S and define a function

$$g(s) = \sum_{\ell=1}^{L} \lambda_\ell \frac{s_{1\ell}}{s_{4\ell}}. \tag{A.29}$$

Then,

$$E\left\{g(S|S_4)\right\} = \sum_{\ell=1}^{L} \lambda_\ell \frac{E(S_{1\ell}|S_{4\ell})}{S_{4\ell}} = \mu \sum_{\ell=1}^{L} \lambda_\ell n_\ell,$$

and hence

$$E\left\{g(S)\right\} = \mu \sum_{\ell=1}^{L} \lambda_\ell n_\ell.$$

Once again, every solution $\boldsymbol{\lambda} \perp \boldsymbol{n}$, where $\boldsymbol{n} = (n_1, \ldots, n_L)'$, provides a counterexample, establishing incompleteness.

Example 14 (General clustered-data setting with random cluster sizes)**.** *Consider clustered data $\boldsymbol{Y}_i$ of size N_i, for $i = 1, \ldots, N$, following an exponential family with data- and cluster-size components $f(\boldsymbol{y}_i|\boldsymbol{\theta}, N_i)$ and $f(N_i|\psi)$. Whenever N_i can take more than one value, the sufficient statistic for $\boldsymbol{\theta}$ is generally incomplete.*

Example 15 (Fully observed 2×1 contingency table)**.** *Consider a binomial experiment based on a binary variable Y_i taking values 1 and 2, with n trials and parameter p ($i = 1, \ldots, n$). Denote the number of 1s and 2s by Z_{21} and Z_{22}, respectively, such that $Z_{21} + Z_{22} = n$. The sufficient statistic for p is complete.*

(The first of the double index is redundant in this example, but is needed in the following one.) Because of the sum constraint, the sufficient statistic is Z_{21} (or Z_{22}), and the MLE is $\widehat{p} = Z_{21}/n$. The result is obvious. Now turn to the same setting where not all observations are made.

Example 16 (Partially missing 2×1 contingency table)**.** *Consider a binomial experiment based on a binary variable Y_i taking values 1 and 2, with n trials and parameter p ($i = 1, \ldots, n$). Denote the number of 1s and 2s by Z_{21} and Z_{22}, respectively, and let the number of trials with unobserved outcome be Z_1. Then, $Z_{21} + Z_{22} + Z_1 = n$. Assume that the missing data are missing at random. The sufficient statistic is incomplete if ignorable likelihood is used.*

In the above, missing at random means that the missing data mechanism does not depend on unobserved information, given observed information. Under missingness at random, mild regularity conditions, and drawing likelihood inferences, it is well-known that the missing-data mechanism can be ignored. For details, see Little and Rubin (2002).

Let $R_i = 1$ if Y_i is observed and $R_i = 0$ otherwise. Further, let $q = P(Y_i = 1)$. Full likelihood means that p and q are both estimated from the data. It is easy to show that $\widehat{p} = Z_{21}/(Z_{21} + Z_{22})$ and $\widehat{q} = (Z_{21} + Z_{22})/n$. When both parameters are estimated, the sufficient statistic (because of the sum constraint) and the parameter vector are both two-dimensional, establishing completeness. However, under missingness at random the likelihood factors into a factor containing p only and a factor with only q. It is then common practice to ignore the factor containing q and to restrict efforts to estimation of p. This leads to the same estimator for p. The sufficient statistic remains two-dimensional: both Z_{21} and Z_{22}, because their sum is random as well, unlike in the non-missing-data case. It is then easy to construct a function $g(z_{21}, z_{22})$, such that $E[g(Z_{21}, Z_{22})] = 0$ for every value of p:

$$g(z_{21}, z_{22}) = \frac{Z_{21} + Z_{22}}{q} - \frac{Z_1}{1 - q}. \tag{A.30}$$

Example 17 (Partially missing 2×2 contingency table)**.** *Consider a contingency table with supplemental margin, cross-classifying two binary outcomes* (Y_{i1}, Y_{i2})*,* $(i = 1, \ldots, n)$*, and with counts* Z_{2jk} *and* Z_{1j} $(j, k = 1, 2)$*. Unless the supplemental margin is empty, the sufficient statistic for the response profile probabilities* p_{jk} *under ignorable likelihood is incomplete.*

Under ignorability, only the probabilities p_{jk} are estimated (subject to their sum being one), and not the missingness probabilities q_j, where q_j is the probability of observing the second outcome Y_{i2} for a subject with $Y_{i1} = j$. Because $E(Z_{2jk}) = np_{jk}q_j$ and $E(Z_{1j}) = np_{j+}(1 - q_j)$, where the $+$ sign instead of k indicates summation over k, it follows that the functions

$$E[g_j(Z_{2j1}, Z_{2j2})] = (1 - q_j)(Z_{2j1} + Z_{2j2}) - q_j Z_{1j},$$

$(j = 1, 2)$, have zero expectation.

Example 18 (Standard exponential distribution for continuous times)**.** *Consider* Y_i *(*$i = 1, \ldots, n$*) i.i.d. with exponential density* $f(y_i) = \lambda e^{-\lambda y_i}$*. The parameter is* λ*, the sufficient statistic* K_1 *is complete.*

The first derivative of the log-likelihood based on the above model is

$$\frac{\partial \ell}{\partial \lambda} = \frac{n}{\lambda} - K_1,$$

from which it clearly follows that the dimension of both parameter and minimal sufficient statistic are equal to one.

Example 19 (Standard Poisson distribution for count data)**.** *Consider* Y_i *(*$i = 1, \ldots, n$*) i.i.d. with Poisson probability* $P(y_i) = \frac{1}{y_i!}\lambda^{y_i} e^{-\lambda}$*. The parameter is* λ*, the sufficient statistic* K_1 *is complete.*

The first derivative of the log-likelihood is

$$\frac{\partial \ell}{\partial \lambda} = \frac{1}{\lambda} K_1 - n,$$

from which it follows also here that the dimension of both parameter and minimal sufficient statistic are equal to one.

Example 20 (Integrated exponential probabilities for counts)**.** *Consider* Y_i *(*$i = 1, \ldots, n$*) i.i.d. with probabilities following from integrating the exponential density between two subsequent integer values:* $P(y_i) = e^{-\lambda y_i}(1 - e^{-\lambda})$*. The parameter is* λ*, the sufficient statistic* K_1 *is complete.*

The first derivative of the log-likelihood is

$$\frac{\partial \ell}{\partial \lambda} = -K_1 + n\frac{e^{-\lambda}}{1-e^{-\lambda}},$$

from which it follows once more that the dimension of both the parameter as well as the minimal sufficient statistics are equal to one.

Note that, while in Examples 18–19 the estimators for λ are equal to the sample average $\widehat{\lambda} = \overline{Y} = K_1/n$, for Example 20 the estimator is

$$\widehat{\lambda} = -\ln\left(\frac{\overline{Y}}{1+\overline{Y}}\right).$$

Of course, this difference is inconsequential for the completeness result.

Example 21 (Integrated Weibull probabilities for counts)**.** *Consider Y_i $(i = 1, \ldots, n)$ i.i.d. with probabilities following from integrating the Weibull density between two subsequent integer values:*

$$P(y_i) = e^{-\lambda y_i^\rho} - e^{\lambda(y_i+1)^\rho}.$$

The parameter is (λ, ρ), representing location and shape, but no reduction in statistics is possible, i.e., it consists of all individual values $(Y_i)_i$.

The log-likelihood derivatives are:

$$\begin{aligned}
\frac{\partial \ell}{\partial \lambda} &= \sum_{i=1}^{n} \frac{-e^{-\lambda y_i^\rho} y_i^\rho + e^{-\lambda(y_i+1)^\rho}(y_i+1)^\rho}{e^{-\lambda y_i^\rho} - e^{\lambda(y_i+1)^\rho}}, \\
\frac{\partial \ell}{\partial \rho} &= \sum_{i=1}^{n} \frac{-e^{-\lambda y_i^\rho} y_i^\rho \ln(y_i) + e^{-\lambda(y_i+1)^\rho}(y_i+1)^\rho \ln(y_i+1)}{e^{-\lambda y_i^\rho} - e^{\lambda(y_i+1)^\rho}}.
\end{aligned}$$

Clearly, no dimension reduction of the data is possible: the parameter is two-dimensional, but the sufficient statistic is of length n. Upon noting that

$$E(y_i) = \sum_{n=0}^{+\infty} e^{-\lambda n^\rho} - 1 = \alpha,$$

(α is used for notational purposes) it follows that a function

$$g(y_1, \ldots, y_n) = \sum_{i=1}^{n} \beta_i y_i,$$

has expectation

$$E[g(y_1, \ldots, y_n)] = \alpha \sum_{i=1}^{n} \beta_i,$$

which is equal to zero for any zero-sum (contrast) vector $(\beta_1, \ldots, \beta_n)'$.

Appendix B

Appendix for Chapter 5

Section B.1 explains the incompleteness in the compound-symmetry model using the definition. The resulting lack of closed-form solutions for MLE are outlined in Section B.2 and further calculations in Section B.3. More on the derivation of weights for the compound-symmetry case are given in Section B.4. Section B.5 and B.6 give more details about respectively a first and second simulation study. Section B.7 describes the use of R for the analysis of the case study.

B.1 Incompleteness in the Compound-symmetry Model

Based on (5.1) The conditional and marginal expectations of (A.25)–(A.28) are:

$$\begin{aligned}
E(W_{1k}|c_k) &= c_k n_k \mu, \\
E(W_{1k}) &= N\mu\pi_k n_k, \\
E(W_2|c_k) &= \sum_{k=1}^{L} c_k n_k(\sigma^2 + d + \mu^2), \\
E(W_2) &= N(\sigma^2 + d + \mu^2)\sum_{k=1}^{L} \pi_k n_k, \\
E(W_{3k}|c_k) &= c_k \left\{ n_k(\sigma^2 + d + \mu^2) + n_k(n_k - 1)(d + \mu^2) \right\}, \\
E(W_{3k}) &= N\pi_k n_k \left\{ (\sigma^2 + d + \mu^2) + (n_k - 1)(d + \mu^2) \right\}, \\
E(W_{4k}) &= N\pi_k.
\end{aligned}$$

Group all sufficient statistics in $\boldsymbol{W}$ and define a function

$$g(\boldsymbol{w}) = \sum_{k=1}^{L} \lambda_k \frac{w_{1k}}{w_{4k}}. \tag{B.1}$$

Then,

$$E\left\{g(\boldsymbol{W}|\boldsymbol{W}_4)\right\} = \sum_{k=1}^{L} \lambda_k \frac{E(W_{1k}|W_{4k})}{W_{4k}} = \mu \sum_{k=1}^{L} \lambda_k n_k,$$

and hence

$$E\left\{g(\boldsymbol{W})\right\} = \mu \sum_{k=1}^{K} \lambda_k n_k.$$

Thus, every solution $\boldsymbol{\lambda} \perp \boldsymbol{n}$, where $\boldsymbol{n} = (n_1, \ldots, n_K)'$, provides a counterexample, establishing incompleteness.

Such a vector $\boldsymbol{\lambda}$ exists if and only if $K \geq 2$, for which it is assumed that at least two $c_k > 0$ (i.e., at least two different cluster sizes occur).

B.2 Likelihood-based Estimation of the CS Model

B.2.1 Score Functions

The score function has components:

$$\frac{\partial \ell}{\partial \mu_k} = \frac{1}{\sigma_k^2 + n_k d_k} \left(\sum_{i=1}^{c_k} \sum_{j=1}^{n_1} y_{ij}^{(k)} - c_k n_k \mu_k \right), \tag{B.2}$$

$$\frac{\partial \ell}{\partial \sigma_k^2} = \frac{-c_k n_k}{2\sigma_k^2} \cdot \frac{\sigma_k^2 + (n_k - 1)d_k}{\sigma_k^2 + n_k d_k} + \frac{c_k n_k S_k}{2\sigma_k^4} - \frac{d_k(2\sigma_k^2 + n_k d_k) c_k n_k T_k}{2\sigma_k^4 (\sigma_k^2 + n_k d_k)^2}, \tag{B.3}$$

$$\frac{\partial \ell}{\partial d_k} = \frac{-c_k n_k}{2(\sigma_k^2 + n_k d_k)} + \frac{c_k n_k T_k}{2(\sigma_k^2 + n_k d_k)^2}, \tag{B.4}$$

with

$$S_k = \frac{1}{c_k n_k} Q_k = \frac{1}{c_k n_k} \sum_{i=1}^{c_k} \boldsymbol{Z}_i^{(k)\prime} \boldsymbol{Z}_i^{(k)}, \tag{B.5}$$

$$T_k = \frac{1}{c_k n_k} R_k = \frac{1}{c_k n_k} \sum_{i=1}^{c_k} \boldsymbol{Z}_i^{(k)\prime} \boldsymbol{J}_{\boldsymbol{n_k}} \boldsymbol{Z}_i^{(k)}. \tag{B.6}$$

B.2.2 Lack of Closed-form Solution when $K \geq 2$

The lack of a closed form when $K \geq 2$ is well known, but we highlight a few relevant features here. More detail is given in Appendix B.3. Function (5.8) can be turned into the log-likelihood kernel for the conventional situation where there is a common mean

parameter and common variance components across all cluster sizes, i.e., $\ell(\mu, \sigma^2, d)$. The score functions follow from summing the terms in (B.2)–(B.4) across cluster sizes:

$$\frac{\partial \ell}{\partial \mu} = \sum_{k=1}^{K} \frac{\partial \ell}{\partial \mu_k}\bigg|_{\mu_k=\mu}, \qquad \frac{\partial \ell}{\partial \sigma^2} = \sum_{k=1}^{K} \frac{\partial \ell}{\partial \sigma_k^2}\bigg|_{\sigma_k^2=\sigma^2}, \qquad \frac{\partial \ell}{\partial d} = \sum_{k=1}^{K} \frac{\partial \ell}{\partial d_k}\bigg|_{d_k=d}. \tag{B.7}$$

Solving the score equation in (B.7) for the mean, using that

$$\Sigma_{n_k}^{-1} = \frac{1}{\sigma^2} I_{n_k} - \frac{d}{\sigma^2(\sigma^2 + n_k d)} J_{n_k},$$

leads to the identity:

$$\widehat{\mu} = \frac{\sum_{k=1}^{K} \frac{n_k c_k}{\sigma^2 + n_k d} \overline{Y}^{(k)}}{\sum_{k=1}^{K} \frac{n_k c_k}{\sigma^2 + n_k d}} = \frac{\sum_{k=1}^{K} \frac{n_k c_k}{\sigma^2 + n_k d} \widehat{\mu_k}}{\sum_{k=1}^{K} \frac{n_k c_k}{\sigma^2 + n_k d}}, \tag{B.8}$$

where $\widehat{\mu_k}$ as in (5.9). For the variance components, only implicit identities follow; they are functions of (B.5)–(B.6). These take the form of high-degree polynomials, for which no general explicit solution exists. While (B.8) is explicit, it is a weighted average of the cluster-size specific averages $\overline{Y}^{(k)}$, with weights depending on the variance components. This, combined with the result for the variance components, implies that there is no explicit solution, unless the variance components are known or the cluster size is constant.

B.3 Full Likelihood

Referring to the conventional situations, i.e. $\ell(\mu, \sigma^2, d)$ in (5.6) and the score equation in (B.7), also second derivatives can be calculated:

$$\frac{\partial^2 \ell}{\partial \mu^2} = \sum_{k=1}^{K} \frac{-c_k n_k}{\sigma^2 + n_k d} \tag{B.9}$$

$$\frac{\partial^2 \ell}{\partial \sigma^2 \partial \mu} = \sum_{k=1}^{K} \frac{-1}{(\sigma^2 + n_k d)^2} \left(\sum_{i=1}^{c_k} \sum_{j=1}^{n_k} y_{ij}^{(k)} - c_k n_k \mu_k \right) \tag{B.10}$$

$$\frac{\partial^2 \ell}{\partial d \partial \mu_k} = \sum_{k=1}^{K} \frac{-n_k}{(\sigma^2 + n_k d)^2} \left(\sum_{i=1}^{c_k} \sum_{j=1}^{n_k} y_{ij}^{(k)} - c_k n_k \mu_k \right) \tag{B.11}$$

$$\frac{\partial^2 \ell}{\partial \mu \partial \sigma^2} = \sum_{k=1}^{K} \left(\frac{-1}{\sigma^4} + \frac{d(2\sigma^2 + n_k d) n_k}{\sigma^4 (\sigma^2 + n_k d)^2} \right) \sum_{i=1}^{c_k} \sum_{j=1}^{n_k} Z_{ij}^{(k)} \tag{B.12}$$

$$\frac{\partial^2 \ell}{(\partial \sigma^2)^2} = \sum_{k=1}^{K} \left(\frac{c_k n_k}{2\sigma^4} \cdot \frac{\sigma^2 + (n_k - 1)d}{\sigma^2 + n_k d} - \frac{c_k n_k}{2\sigma^2} \cdot \frac{d}{(\sigma^2 + n_k d)^2} - \frac{c_k n_k S_k}{\sigma^6} \right.$$
$$\left. - d c_k n_k T_k \frac{\sigma^2(\sigma^2 + n_k d) - (2\sigma^2 + n_k d)^2}{\sigma^6 (\sigma^2 + n_k d)^3} \right) \tag{B.13}$$

$$\frac{\partial^2 \ell}{\partial d \partial \sigma^2} = \sum_{i=1}^{K} \left(\frac{c_k n_k}{2(\sigma^2 + n_k d)} - \frac{c_k n_k T_k}{\sigma^4} \cdot \frac{(\sigma^2 + n_k d)^2 - n_k d(2\sigma^2 + n_k d)}{(\sigma^2 + n_k d)^3} \right) \tag{B.14}$$

$$\frac{\partial^2 \ell}{\partial \mu \partial d} = \sum_{k=1}^{K} \frac{-n_k}{(\sigma^2 + n_k d)^2} \sum_{i=1}^{c_k} \sum_{j=1}^{n_k} Z_{ij}^{(k)} \tag{B.15}$$

$$\frac{\partial^2 \ell}{\partial \sigma^2 \partial d} = \sum_{k=1}^{K} \left(\frac{c_k n_k}{2(\sigma^2 + n_k d)^2} - \frac{c_k n_k T_k}{(\sigma^2 + n_k d)^3} \right) \tag{B.16}$$

$$\frac{\partial^2 \ell}{\partial d^2} = \sum_{k=1}^{K} \left(\frac{c_k n_k^2}{2(\sigma^2 + n_k d)^2} - \frac{c_k n_k^2 T_k}{(\sigma^2 + n_k d)^3} \right). \tag{B.17}$$

Should we use conditional likelihood, then the log-likelihood's kernel equals:

$$\begin{aligned} L \quad \propto \quad & \prod_{i=1}^{k} \frac{1}{(2\pi)^{n_1/2} |\Sigma_{n_1}|^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{y}_{i1} - \boldsymbol{\mu}_{n_1})' \Sigma_{n_1}^{-1} (\boldsymbol{y}_{i1} - \boldsymbol{\mu}_{n_1}) \right\} \\ & \times \left[\frac{\Phi(\alpha + \boldsymbol{y}_{i1}' \boldsymbol{\beta})}{\Phi\left(\frac{\alpha + \boldsymbol{\mu}_{n_1}' \boldsymbol{\beta}}{\sqrt{1 + \boldsymbol{\beta}' \Sigma_{n_1} \boldsymbol{\beta}/n_1}} \right)} \right] \\ & \times \prod_{i=k+1}^{N} \frac{1}{(2\pi)^{n_2/2} |\Sigma_{n_2}|^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{y}_i - \boldsymbol{\mu}_{n_2})' \Sigma_{n_2}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_{n_2}) \right\} \\ & \times \left[\frac{1 - \Phi(\alpha + \boldsymbol{y}_{i1}' \boldsymbol{\beta})}{1 - \Phi\left(\frac{\alpha + \boldsymbol{\mu}_{n_1}' \boldsymbol{\beta}}{\sqrt{1 + \boldsymbol{\beta}' \Sigma_{n_1} \boldsymbol{\beta}/n_1}} \right)} \right] \end{aligned} \tag{B.18}$$

$$\begin{aligned} \ell \quad \propto \quad & -\frac{1}{2} \sum_{i=1}^{k} \left\{ \ln |\Sigma_{n_1}| + (\boldsymbol{y}_i - \boldsymbol{\mu}_{n_1})' \Sigma_{n_1}^{-1} (\boldsymbol{y}_{i1} - \boldsymbol{\mu}_{n_1}) \right\} \\ & - k \ln \Phi(\tilde{\alpha} + \boldsymbol{\mu}_{n_1}' \tilde{\boldsymbol{\beta}}) \\ & - \frac{1}{2} \sum_{i=k+1}^{N} \left\{ \ln |\Sigma_{n_2}| + (\boldsymbol{y}_i - \boldsymbol{\mu}_{n_2})' \Sigma_{n_2}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_{n_2}) \right\} \\ & - (N - k) \ln \left\{ 1 - \Phi(\tilde{\alpha} + \boldsymbol{\mu}_{n_1}' \tilde{\boldsymbol{\beta}}) \right\}, , \end{aligned} \tag{B.19}$$

with $\tilde{\alpha} = \frac{\alpha}{\sqrt{1 + \boldsymbol{\beta}' \Sigma_{n_1} \boldsymbol{\beta}/n_1}}$ and $\tilde{\boldsymbol{\beta}} = \frac{\boldsymbol{\beta}}{\sqrt{1 + \boldsymbol{\beta}' \Sigma_{n_1} \boldsymbol{\beta}/n_1}}$

The corresponding score equations are:

$$\begin{aligned}\frac{\partial \ell}{\partial \mu} &= \frac{1}{\sigma^2 + n_1 d}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_1} \boldsymbol{y}_{ij} - k n_1 \mu\right) - k n_1 \boldsymbol{j}'_{\boldsymbol{n_1}} \tilde{\boldsymbol{\beta}} \frac{\phi(\tilde{\alpha} + \boldsymbol{\mu}'_{n_1}\tilde{\boldsymbol{\beta}})}{\Phi(\tilde{\alpha} + \boldsymbol{\mu}'_{n_1}\tilde{\boldsymbol{\beta}})} \\ &+ \frac{1}{\sigma^2 + n_2 d}\left(\sum_{i=k+1}^{N}\sum_{j=1}^{n_2} \boldsymbol{y}_{ij} - (N-k) n_2 \mu\right) \\ &- (N-k) n_2 \boldsymbol{j}'_{\boldsymbol{n_1}} \tilde{\boldsymbol{\beta}} \frac{\phi(\tilde{\alpha} + \boldsymbol{\mu}'_{n_1}\tilde{\boldsymbol{\beta}})}{\Phi(\tilde{\alpha} + \boldsymbol{\mu}'_{n_1}\tilde{\boldsymbol{\beta}})}, \end{aligned} \tag{B.20}$$

with $\frac{\partial \ell}{\partial \sigma^2}$ and $\frac{\partial \ell}{\partial d}$ identical to (B.3) and (B.4). The components of the Hessian are:

$$\begin{aligned}\frac{\partial^2 \ell}{\partial \mu^2} &= \frac{-k n_1}{\sigma^2 + n_1 d} - \frac{(N-k) n_2}{\sigma^2 + n_2 d} \\ &- k n_1 \boldsymbol{j}'_{\boldsymbol{n_1}} \tilde{\boldsymbol{\beta}} \left[\frac{-\Phi(\tilde{\alpha} + \boldsymbol{\mu}'_{n_1}\tilde{\boldsymbol{\beta}}) \cdot \phi(\tilde{\alpha} + \boldsymbol{\mu}'_{n_1}\tilde{\boldsymbol{\beta}}) \cdot (\tilde{\alpha} + \boldsymbol{\mu}'_{n_1}\tilde{\boldsymbol{\beta}}) \cdot \boldsymbol{j}'_{\boldsymbol{n_1}}\tilde{\boldsymbol{\beta}}}{\Phi^2(\tilde{\alpha} + \boldsymbol{\mu}'_{n_1}\tilde{\boldsymbol{\beta}})}\right. \\ &\left. - \frac{\phi^2(\tilde{\alpha} + \boldsymbol{\mu}'_{n_1}\tilde{\boldsymbol{\beta}}) \cdot \boldsymbol{j}'_{\boldsymbol{n_1}}\tilde{\boldsymbol{\beta}}}{\Phi^2(\tilde{\alpha} + \boldsymbol{\mu}'_{n_1}\tilde{\boldsymbol{\beta}})}\right] \\ &+ (N-k) n_2 \boldsymbol{j}'_{\boldsymbol{n_1}} \tilde{\boldsymbol{\beta}} \\ &\times \left[\frac{-(1 - \Phi(\tilde{\alpha} + \boldsymbol{\mu}'_{n_1}\tilde{\boldsymbol{\beta}})) \cdot \phi(\tilde{\alpha} + \boldsymbol{\mu}'_{n_1}\tilde{\boldsymbol{\beta}}) \cdot (\tilde{\alpha} + \boldsymbol{\mu}'_{n_1}\tilde{\boldsymbol{\beta}}) \cdot \boldsymbol{j}'_{\boldsymbol{n_1}}\tilde{\boldsymbol{\beta}}}{(1 - \Phi(\tilde{\alpha} + \boldsymbol{\mu}'_{n_1}\tilde{\boldsymbol{\beta}}))^2}\right. \\ &\left. + \frac{\phi^2(\tilde{\alpha} + \boldsymbol{\mu}'_{n_1}\tilde{\boldsymbol{\beta}}) \cdot \boldsymbol{j}'_{\boldsymbol{n_1}}\tilde{\boldsymbol{\beta}}}{(1 - \Phi(\tilde{\alpha} + \boldsymbol{\mu}'_{n_1}\tilde{\boldsymbol{\beta}}))^2}\right] \end{aligned} \tag{B.21}$$

$$\begin{aligned}\frac{\partial^2 \ell}{\partial \sigma^2 \partial \mu} &= \frac{-1}{(\sigma^2 + n_1 d)^2}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_1} \boldsymbol{y}_{ij} - k n_1 \mu\right) \\ &- \frac{1}{(\sigma^2 + n_2 d)^2}\left(\sum_{i=k+1}^{N}\sum_{j=1}^{n_2} \boldsymbol{y}_{ij} - (N-k) n_2 \mu\right) \end{aligned} \tag{B.22}$$

$$\begin{aligned}\frac{\partial^2 \ell}{\partial d \partial \mu} &= \frac{-n_1}{(\sigma^2 + n_1 d)^2}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_1} \boldsymbol{y}_{ij} - k n_1 \mu\right) \\ &- \frac{n_2}{(\sigma^2 + n_2 d)^2}\left(\sum_{i=k+1}^{N}\sum_{j=1}^{n_2} \boldsymbol{y}_{ij} - (N-k) n_2 \mu\right) \end{aligned} \tag{B.23}$$

with $\frac{\partial^2 \ell}{\partial \mu \partial \sigma^2}$, $\frac{\partial^2 \ell}{(\partial \sigma^2)^2}$, $\frac{\partial^2 \ell}{\partial d \partial \sigma^2}$, $\frac{\partial^2 \ell}{\partial \mu \partial d}$, $\frac{\partial^2 \ell}{\partial \sigma^2 \partial d}$, and $\frac{\partial^2 \ell}{\partial (d)^2}$ indentical to (B.12), (B.13), (B.14), (B.15), (B.16), and (B.17).

B.4 Derivation of Optimal Scalar Weights for Compound-symmetry Case

To find the optimal scalar weight with minimum variance for μ we use the method of Lagrange with the constraint that the weights a_k need to sum to 1:

$$Q = \sum_{k=1}^{K} a_k^2 \frac{\sigma^2 + n_k d}{c_k n_k} - \lambda \left(\sum_{k=1}^{K} a_k - 1 \right). \tag{B.24}$$

Solving the first partial derivative we become an expression for a_k involving λ. Summing this one produces an expression for λ, leading to the complete formula for a_k. Precisely, the system of equations is:

$$\begin{aligned} \frac{\partial Q}{\partial a_k} = 2a_k \frac{\sigma^2 + n_k d}{c_k n_k} - \lambda &= 0, \\ \frac{\partial Q}{\partial \lambda} = \sum_{k=1}^{K} a_k - 1 &= 0, \end{aligned}$$

or, alternatively:

$$\begin{aligned} a_k &= \frac{\lambda}{2} \frac{c_k n_k}{\sigma^2 + n_k d}, \\ \lambda &= \left(\frac{1}{2} \sum_{k=1}^{K} \frac{c_k n_k}{\sigma^2 + n_k d} \right)^{-1}, \end{aligned}$$

and hence

$$a_k = \frac{\frac{c_k n_k}{\sigma^2 + n_k d}}{\sum_{m=1}^{K} \frac{c_m n_m}{\sigma^2 + n_m d}}.$$

In the same manner, expressions for b_k and g_k can be found, as we will show next. For σ^2:

$$Q = 2\sigma^4 \sum_{k=1}^{K} b_k^2 \frac{1}{c_k (n_k - 1)} - \lambda \left(\sum_{k=1}^{K} b_k - 1 \right),$$

producing the system:

$$\begin{aligned} \frac{\partial Q}{\partial b_k} = \frac{4\sigma^4 b_k}{c_k (n_k - 1)} - \lambda &= 0, \\ \frac{\partial Q}{\partial \lambda} = \sum_{k=1}^{K} b_k - 1 &= 0, \end{aligned}$$

which can be rewritten as:

$$\begin{aligned} 4\sigma^4 b_k &= \lambda c_k (n_k - 1), \\ \lambda &= \frac{4\sigma^4}{\sum_{k=1}^{K} c_k (n_k - 1)}, \end{aligned}$$

and finally:

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

For d, the objective function is:

$$Q = \sum_{k=1}^{K} g_k^2 v_k - \lambda(\sum_{k=1}^{K} g_k - 1),$$

with

$$v_k = \frac{2}{c_k n_k}\left(\frac{\sigma^4}{n_k - 1} + 2d\sigma^2 + n_k d^2\right).$$

The system of equations now is:

$$\begin{aligned} \frac{\partial Q}{\partial g_k} = 2g_k v_k - \lambda &= 0, \\ \frac{\partial Q}{\partial \lambda} = \sum_{k=1}^{K} g_k - 1 &= 0, \end{aligned}$$

leading to:

$$\begin{aligned} g_k &= \lambda \frac{1}{2v_k}, \\ \lambda &= \frac{2}{\sum_{k=1}^{K} \frac{1}{v_k}}, \end{aligned}$$

giving the solution:

$$g_k = \frac{\frac{c_k n_k}{\frac{\sigma^4}{n_k - 1} + 2d\sigma^2 + n_k d^2}}{\sum_{m=1}^{K} \frac{c_m n_k}{n_m - 1} + 2d\sigma^2 + n_m d^2}.$$

B.4.1 Cluster-by-cluster Analysis

We study the case of the most extreme partitioning, i.e., where each of the clusters is analyzed separately. This can be relevant in cases with perhaps a limited number of very to extremely large clusters. This means that $c_k \equiv 1$ throughout. Clearly, the n_k will then no longer be unique. We will examine this case in detail, and contrast a first weighted estimator with an *ad hoc* one.

B.4.1.1 The Weighted Estimator for the Cluster-by-cluster Case

The estimator follows from setting $c_k \equiv 1$ and hence $K \equiv N$ throughout. For example, this special case can easily be considered for all expressions in Sections 5.4.1.1–5.4.1.3. Because c_k enters the inverse of the variance-covariance matrix multiplicatively, as is seen from (5.12)–(5.13), the optimal estimator that is obtained when each cluster is considered

to be its own stratum, is identical to the one obtained when strata are defined in terms of all clusters of a given size. The same is true for the scalar weights.

It is insightful to consider in more detail the special case where further the cluster sizes are all identical to n. One then easily obtains:

$$\widehat{\mu} = \frac{1}{Nn}\sum_{i=1}^{N}\sum_{j=1}^{n} Y_{ij}, \tag{B.25}$$

$$\begin{aligned}\widehat{\sigma}^2 &= \frac{1}{Nn(n-1)}\left(n\sum_{i=1}^{N}\boldsymbol{Z}_i'\boldsymbol{Z}_i - \sum_{i=1}^{N}\boldsymbol{Z}_i' J_n \boldsymbol{Z}_i\right)\\ &= \frac{1}{Nn(n-1)}(nQ-R),\end{aligned} \tag{B.26}$$

$$\begin{aligned}\widehat{d} &= \frac{1}{Nn(n-1)}\left(\sum_{i=1}^{N}\boldsymbol{Z}_i' J_n \boldsymbol{Z}_i - \sum_{i=1}^{N}\boldsymbol{Z}_i'\boldsymbol{Z}_i\right)\\ &= \frac{1}{Nn(n-1)}(Q-R),\end{aligned} \tag{B.27}$$

with obvious notation for Q and R, inspired by (B.5)–(B.6). The corresponding variance-covariance elements, similar in spirit to (5.12)–(5.13), are:

$$\mathrm{var}(\widehat{\mu}) = \frac{\sigma^2 + nd}{Nn}, \tag{B.28}$$

$$\mathrm{var}\begin{pmatrix}\widehat{\sigma^2}\\ \widehat{d}\end{pmatrix} = \begin{pmatrix}\frac{2\sigma^4}{N(n-1)} & -\frac{2\sigma^4}{Nn(n-1)}\\ -\frac{2\sigma^4}{Nn(n-1)} & \frac{2}{Nn}\left[\frac{\sigma^4}{n-1} + 2\sigma^2 d + nd^2\right]\end{pmatrix}. \tag{B.29}$$

This estimator coincides with the MLE, as is known from Molenberghs, Verbeke, and Iddi (2011).

B.4.1.2 A Two-stage Estimator for Compound Symmetry

In linear mixed models, there is a method of estimation, sometimes called the two-stage approach (Laird and Ware, 1982; Verbeke and Molenberghs, 2000), in which each cluster is analyzed separately to begin with, using linear regression, after which the cluster-specific parameters are summarized into fixed effects. Although the above cluster-by-cluster analysis is superficially similar to this, it is not equivalent. In particular, there is no bias (as can be seen in the two stage method), and the maximum likelihood estimator is recovered.

This approach is most useful when cluster sizes are not constant, and in models that are more complex than compound symmetry. However, to gain some insight, we develop the details of the method for the CS model with constant cluster size.

For the mean, (B.25) is retained, as the average of the cluster-specific averages $\overline{Y}_i$. Further, define:

$$s^2 = \frac{1}{Nn}\sum_{i=1}^{N}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_i)^2, \tag{B.30}$$

$$t^2 = \frac{1}{N}\sum_{i1}^{N}(\overline{Y}_i-\widehat{\mu})^2. \tag{B.31}$$

Straightforward algebra shows:

$$E(s^2) = \frac{n-1}{n}\sigma^2, \tag{B.32}$$

$$E(t^2) = \frac{N-1}{N}\left(d+\frac{1}{n}\sigma^2\right). \tag{B.33}$$

Should n and N approach infinity, then it follows that s^2 and t^2 are asymptotically unbiased estimators for σ^2 and d, respectively. However, this is not always reasonable. In applications such as the NTP data (Section 3.1),**??**), it is fair to say that the cluster size has a biological upper limit. In other situations, however, such as meta-analyses, it is sensible to assume that both n and N approach infinity.

In the next section, we will study the consequences of removing the bias. For now, a small, obvious modification is:

$$s_*^2 = \frac{n}{n-1}s^2, \tag{B.34}$$

$$t_*^2 = \frac{N}{N-1}t^2. \tag{B.35}$$

Now, s_*^2 is unbiased, while $E(t_*^2)=d+\sigma^2/n$, the bias σ^2/n can be made to disappear asymptotically provided it is sensible to let n grow large.

It is of interest to consider the variance-covariance structure of the estimators s^2, t^2, s_*^2, and t_*^2, as well as to make relative efficiency considerations. This will be done next.

B.4.1.3 Connections Between Estimators

Comparing algebraic expressions (B.26)–(B.27) with (B.30)–(B.31), leads to the linear relationships:

$$s^2 = \frac{n-1}{n}\widehat{\sigma^2}+0\cdot\widehat{d}, \tag{B.36}$$

$$t^2 = \frac{N-1}{Nn}\widehat{\sigma^2}+\frac{N-1}{N}\widehat{d}. \tag{B.37}$$

Relationships (B.36)–(B.37) can be combined with (B.29) to produce:

$$\text{var}\begin{pmatrix} s^2 \\ t^2 \end{pmatrix} = \begin{pmatrix} \frac{2(n-1)\sigma^4}{Nn^2} & 0 \\ 0 & \frac{2(N-1)^2}{N^2n}\left[\frac{\sigma^4}{n}+2\sigma^2 d+nd^2\right] \end{pmatrix} \tag{B.38}$$

and, similarly,

$$\text{var}\begin{pmatrix} s_*^2 \\ t_*^2 \end{pmatrix} = \begin{pmatrix} \frac{2\sigma^4}{N(n-1)} & 0 \\ 0 & \frac{2}{Nn}\left[\frac{\sigma^4}{n} + 2\sigma^2 d + nd^2\right] \end{pmatrix}. \tag{B.39}$$

From its definition it follows that $s_*^2 \equiv \widehat{\sigma^2}$. The same is not true for t_*^2. One reason to consider it nevertheless is its independence from s_*^2. Indeed, $(\widehat{\mu}, s_*^2 \equiv \widehat{\sigma^2}, t_*^2)'$ is an estimator with mutually independent components. While the same is true when s^2 and t^2 are used instead, the biases are larger.

For this case then, the choice between $\widehat{d}$ and t_*^2 is in terms of a trade-off between efficiency and independence.

To gauge the efficiency loss when using t_*^2, the mean squared error is:

$$MSE(t_*^2) = \frac{2}{Nn}\left(\frac{\sigma^4}{n} + 2\sigma^2 d + nd^2\right) + \frac{1}{n^2}\sigma^4,$$

and hence the relative MSE:

$$RMSE(t^2; \widehat{d}) = \frac{2\left(\frac{\sigma^4}{n} + 2\sigma^2 d + nd^2\right) + \frac{N}{n}\sigma^4}{2\left(\frac{\sigma^4}{n-1} + 2\sigma^2 d + nd^2\right)}, \tag{B.40}$$

which approaches infinity when N does, while n would remain constant. In other words, this estimator is inconsistent unless it is being applied in situations where n can also be considered to be large.

There are three distinct situations. First, when $N/n = \lambda n + o(n)$, for some λ, i.e., when N is of the order of n^2, then, based on (B.40), the ARE is $2(d^2 + \lambda\sigma^4)/[2d^2]$. The magnitude of the efficiency loss depends on sizes of the parameters involved. Second, when $\mathcal{O}(N) < \mathcal{O}(n^2)$, the ARE equals 1. This includes the cases where N is constant, $N = n^{1/2}$, $N = n$, and $N = n^{3/2}$, for example. A constant or slowly increasing N is plausible in a meta-analytic context. Third, if N/n increases too quickly, i.e., $\mathcal{O}(N/n) > \mathcal{O}(n)$, then the estimator t_*^2 is inconsistent. This is the case, in particular, for bounded n.

The estimators s^2 and t^2 can be combined linearly to produce unbiased estimators. In other words, based on (B.32)–(B.33), the following corrections can be applied to (B.30)–(B.31):

$$s_{\text{corr}}^2 = \frac{n}{n-1}s^2, \tag{B.41}$$

$$t_{\text{corr}}^2 = \frac{N}{N-1}t^2 - \frac{N}{(n-1)(N-1)}s^2. \tag{B.42}$$

Interestingly, this requirement reproduces (B.36)–(B.37): the requirement of an unbiased estimator reproduces $\widehat{\sigma^2}$ and $\widehat{d}$, presented in (B.26)–(B.27) and hence also with their variance.

B.5 Details About the First Simulation Study

The simulation study, summarized in Section 5.5, is described in detail here.

B.5.1 Simulation Method

The design of the simulation study is as follows.

- Each generated set of data consists of c_k clusters of size n_k, for $k = 1, \ldots, K$. We choose $K = 4$ throughout.
- For a generated set of data, the splitting is done by placing all clusters of a given size in one sub-sample.
- The CS model parameters are $\mu = 0$, $d = 1$, and $\sigma^2 = 2$.
- After estimating the three model parameters within each sub-sample, they are combined using the following weighting methods: (a) equal, (b) proportional, where the weights are
$$w_k = \frac{c_k}{\sum_{\ell=1}^{4} c_\ell},$$
and (c) size-proportional, where the weights for μ and d are:
$$w_k \frac{c_k n_k}{\sum_{\ell=k}^{4} c_\ell n_\ell},$$
while for σ^2 we take:
$$w_k = \frac{c_k(n_k - 1)}{\sum_{\ell=1}^{4} c_\ell(n_\ell - 1)}.$$
- Per setting, 100 replications are considered.

These settings are applied to various combinations of the n_k and c_k, now described in turn.

B.5.2 Setting 1: Equal $c_k \cdot n_k$, Different c_k and n_k.

Consider 150 samples in each split, as follows: $(c_1, n_1) = (3, 50)$, $(c_2, n_2) = (5, 30)$, $(c_3, n_3) = (10, 15)$, and $(c_4, n_4) = (15, 10)$. The results are presented in Table B.1. Graphical depictions can be found in Figures B.1 and B.2. Figure B.1 shows that there is a different amount of information in the various sub-samples. This is not a problem, rather a consequence of the way the splits are created and the different amounts of information carried in each. It reminds us that we need to be judicious how the information from the splits will be weighted. It is not a surprise that equal weights are a poor choice. The

other methods perform similarly, and all do very well. To varying degrees, the same will be seen in Settings 2 and 3.

Table B.1: *First simulation study. Setting 1. Average of split-specific and combined (weighted) parameters and their precision estimates.*

	μ	var(μ)	d	var(d)	σ^2	var(σ^2)
split1	-0.00396	0.05779	0.68143	0.41395	1.98676	0.00296
split2	0.05697	0.03071	0.80997	0.19712	1.98578	0.00304
split3	-0.02111	0.01174	0.95161	0.07869	1.97690	0.00319
split4	0.01123	0.00626	0.98870	0.04677	1.98056	0.00347
Equal	0.01078	0.03769	0.85793	0.09988	1.98250	0.01406
Prop	0.00698	0.03230	0.92245	0.08568	1.98081	0.01907
Size prop	0.01078	0.03769	0.85793	0.09988	1.98260	0.01405
Full	0.00780	0.03513	0.98016	0.08614	1.98257	0.01392

B.5.3 Setting 2: Different $c_k \cdot n_k$, Equal c_k, Different n_k

To see the effect of split size, the following choices are made: $(c_1, n_1) = (4, 25)$, $(c_2, n_2) = (4, 50)$, $(c_3, n_3) = (4, 125)$, and $(c_4, n_4) = (4, 250)$. As a consequence, the size of the splits will be 100, 200, 500, and 1000, respectively. Table B.2 summarized the results, with graphical displays presented in Figures B.3 and B.4.

B.5.4 Setting 3: Different $c_k \cdot n_k$, Different c_k, Equal n_k

We now choose: $(c_1, n_1) = (10, 20)$, $(c_2, n_2) = (20, 20)$, $(c_3, n_3) = (50, 20)$, and $(c_4, n_4) = (100, 20)$. Table B.3 summarizes the results. Graphs can be found in Figures B.5 and B.6.

B.5.5 Optimal, Approximate Optimal, and Iterated Optimal Weights

Optimal weights were discussed in Section 5.4.1.1. When we plug the MLE's into the optimal weights, the result of using these weights is the MLE's itself. Of course, this is a circular reasoning, which is why one needs to resort to, for example, the approximate or iterated optimal weights derived in Section 5.4.1.2. For both of these, using Settings 1–3, we conducted simulations. They are reported in Figures B.7–B.9.

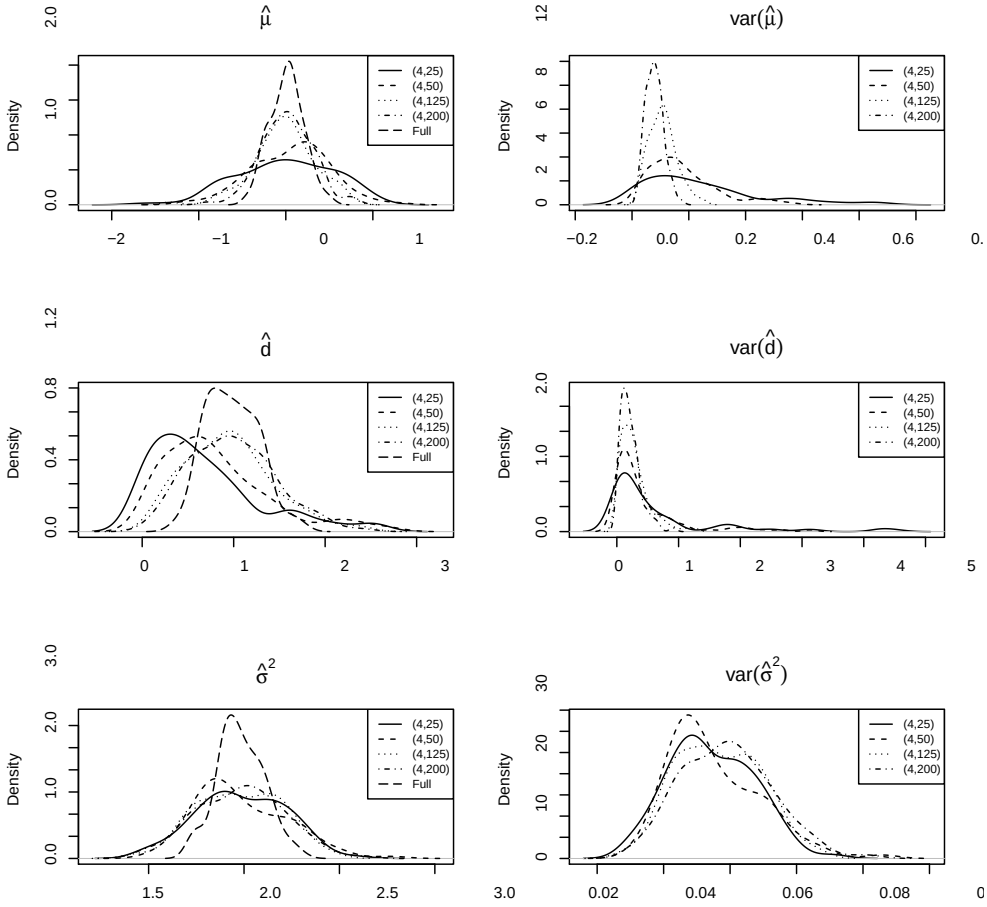

Figure B.1: *First simulation study. Setting 1. Split-specific results.*

It is noteworthy that the behavior of the iterated optimal weights depends on c_k and n_k. First, they often but not always converge in a single iteration; the maximum number of iterations observed in our simulations being 6. Second, the iterated optimal weights converge to size optimal weights for σ^2 and to proportional weights for d.

Taken together, it follows that both approximately optimal and iterated optimal weights provide excellent results. The specific attraction of the approximate optimal weights is that they obviate the need for iteration, which is a factor of stability and speed.

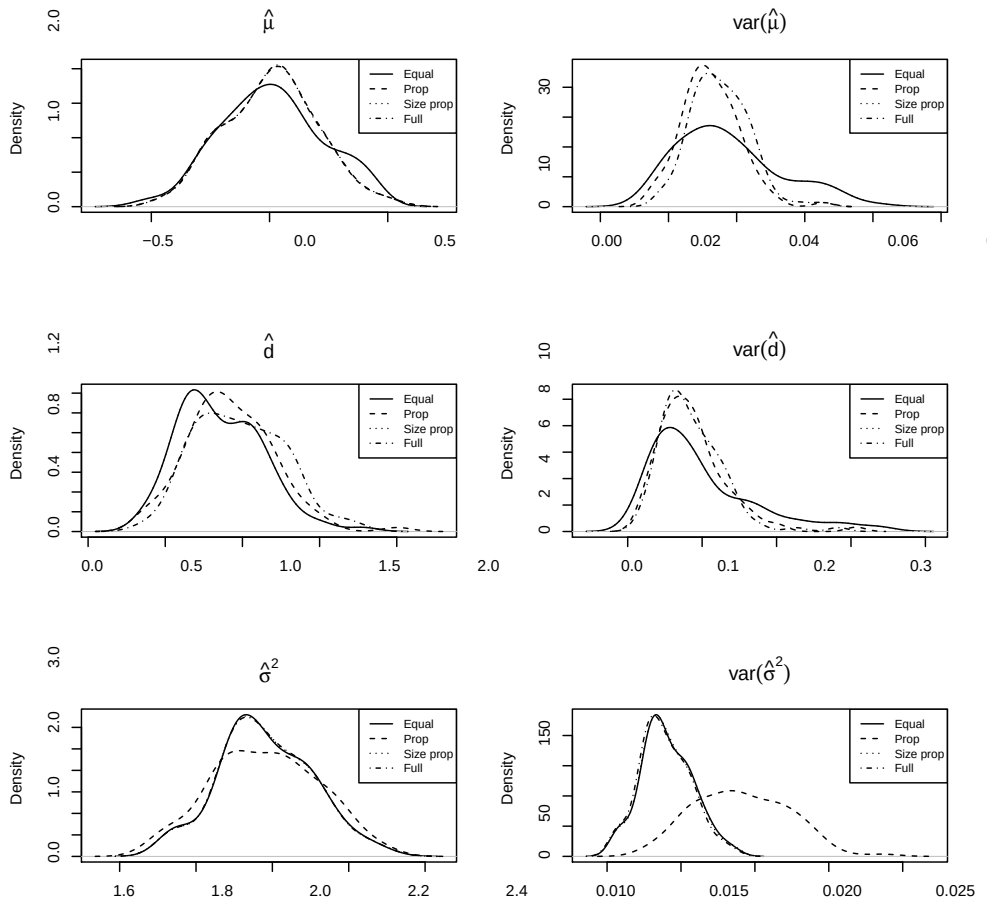

Figure B.2: *First simulation study. Setting 1. Combining the results from the four splits, using equal, proportional, and size proportional weights. This is compared with full maximum likelihood.*

B.6 Details About the Second Simulation Study

The aim of this study is to compare the proposed method to two alternatives:

1. full maximum likelihood;
2. the proposed sample-splitting method, allowing for closed forms;
3. using multiple imputation (MI) first, to render the clusters of equal sizes, and then apply closed-form solutions to the augmented balanced data, together with the

Table B.2: *First simulation study. Setting 2. Average of split-specific and combined (weighted) parameters and their precision estimates.*

	μ	var(μ)	d	var(d)	σ^2	var(σ^2)
split1	-0.02515	0.05227	0.83440	0.52423	2.00347	0.00730
split2	0.01287	0.05157	0.86891	0.57904	1.97285	0.00160
split3	0.06812	0.03586	0.74147	0.23681	2.00165	0.00026
split4	-0.03676	0.02979	0.68241	0.14117	1.99216	0.00006
Equal	0.00477	0.05111	0.78180	0.14770	1.99253	0.00935
Prop	0.00477	0.05111	0.78180	0.14770	1.99253	0.00935
Size prop	-0.00147	0.07139	0.72798	0.16585	1.99328	0.00447
Full	0.00530	0.06339	0.89599	0.14604	1.99333	0.00446

Table B.3: *First simulation study. Setting 3. Average of split-specific and combined (weighted) parameters and their precision estimates.*

	μ	var(μ)	d	var(d)	σ^2	var(σ^2)
split1	0.00343	0.00900	0.84739	0.05169	2.02445	0.00190
split2	0.02553	0.00304	1.00224	0.01754	2.01962	0.00047
split3	-0.00010	0.00045	0.95794	0.00212	1.99765	0.00007
split4	0.01151	0.00012	1.01226	0.00064	1.98944	0.00002
Equal	0.01009	0.01139	0.95496	0.02694	2.00779	0.00486
Prop	0.00939	0.00604	0.98690	0.01369	1.99702	0.00234
Size prop	0.00939	0.00604	0.98690	0.01369	1.99702	0.00234
Full	0.00939	0.00614	1.00487	0.01372	1.99702	0.00233

combination rules.

B.6.1 Simulation Plan

In order to study the effect of cluster sizes (n_k) and number of clusters of each size (c_k), 5 different configurations are considered:

Config. 1. $c_k = (15, 25, 30, 20, 10)$, $n_k = (8, 5, 3, 9, 15)$;

Config. 2. $c_k = (150, 250, 300, 200, 100)$, $n_k = (8, 5, 3, 9, 15)$;

Config. 3. $c_k = (1500, 2500, 3000, 2000, 1000)$, $n_k = (8, 5, 3, 9, 15)$;

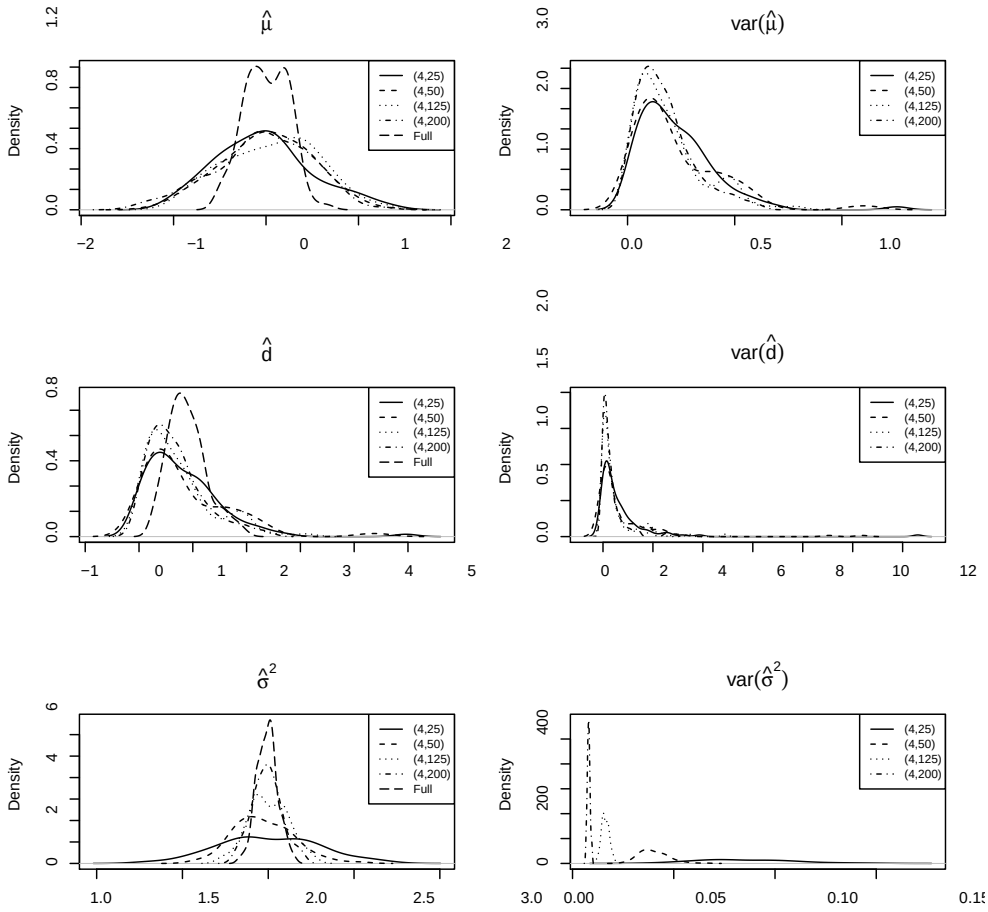

Figure B.3: *First simulation study. Setting 2. Split-specific results.*

Config. 4. $c_k = (15, 25, 30, 20, 10)$, $n_k = (80, 50, 30, 90, 150)$;

Config. 5. $c_k = (15, 25, 30, 20, 10)$, $n_k = (800, 500, 300, 900, 1500)$.

Each configuration is repeated 100 times.

Each cluster is generated from a CS model with $\mu_0 = 0$, $d_0 = 1$, and $\sigma_0^2 = 4$.

For estimating the parameters using the full unbalanced data, PROC MIXED in SAS (Version 9.4) is used with the covariance structure in the REPEATED statement set to `type=cs`.

The closed form solutions and their variances are implemented in `R` in three different ways. First, the formulas are implemented directly using 'for' loops. Following the ideas in

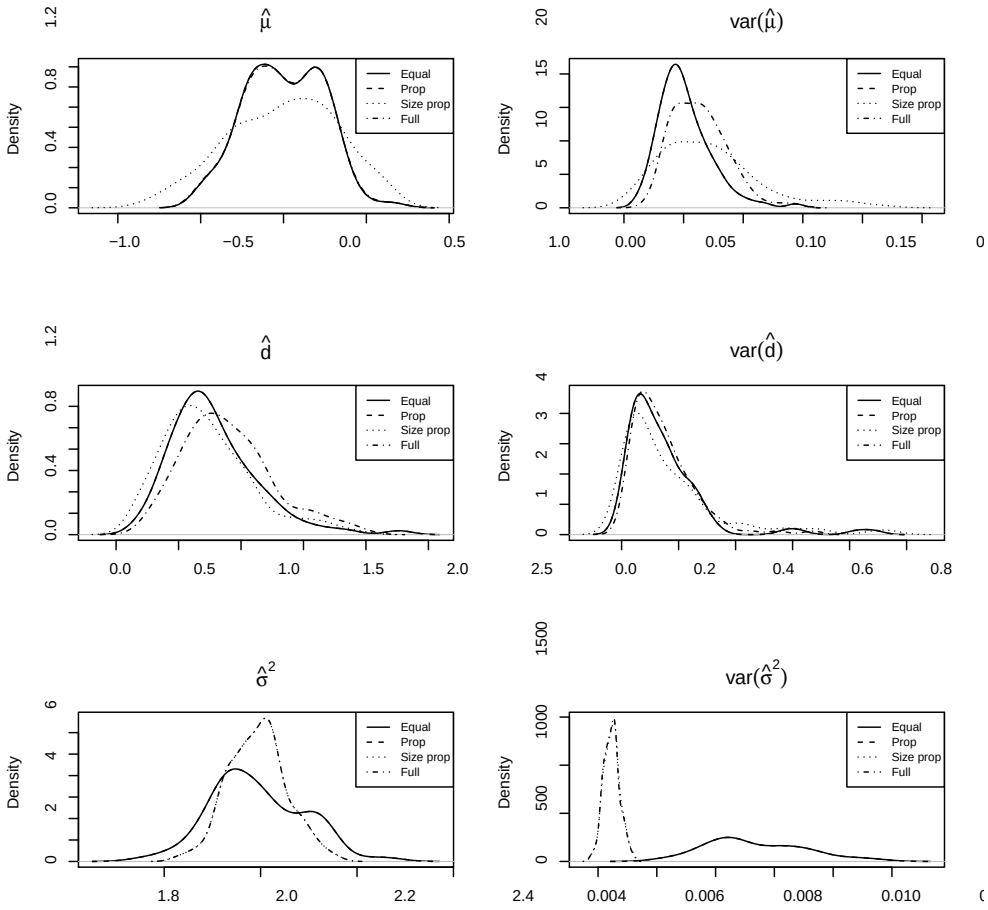

Figure B.4: *First simulation study. Setting 2. Combining the results from the four splits, using equal, proportional, and size proportional weights. This is compared with full maximum likelihood.*

Sikorska *et al.* (2013), it might be faster to replace 'for' loops with vectorized computation. For $\widehat{\mu}_k$ it is straightforward, since one just needs to compute an arithmetic average. If Z is a n_k times c_k matrix with its ith column defined as $Z_i^{(k)} = \left(Y_i^{(k)} - \mu_k \mathbf{1}_{n_k}\right)$, then computing $\sum_{i=1}^{c_k} Z_i^{(k)'} Z_i^{(k)}$ is equivalent to replacing each element in matrix Z by its square, and then sum over the sum of its columns. Furthermore, $J_{n_k} Z_i^{(k)}$ would simply compute the sum of columns in matrix Z. Therefore, $\sum_{i=1}^{c_k} Z_i^{(k)'} J_{n_k} Z_I^{(k)}$ is equivalent to post-multiplying Z by the sum of its columns and then sum over this vector. In this

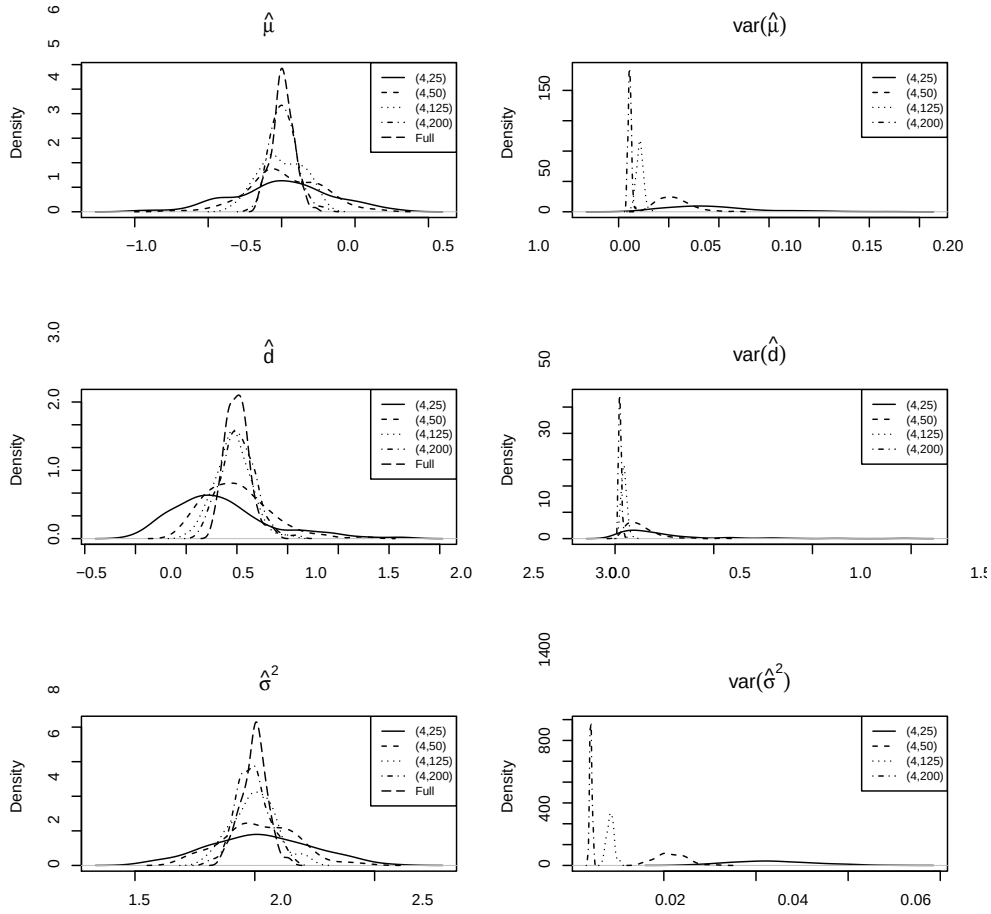

Figure B.5: *First simulation study. Setting 3. Split-specific results.*

way, within each split, the parameters can be estimated avoiding 'for' loops.

Second, it is also possible to find the estimates for all of the splits at once instead of computing them separately.

A third way consists of calculating all the estimates together and not split by split in a 'for' loop. This approach is possible via imposing balance through adding missing values in the matrix but, when multiplying and summing, ignoring the missing values. This is very easy in R.

We will compare computation time between these three approaches, for the five configurations.

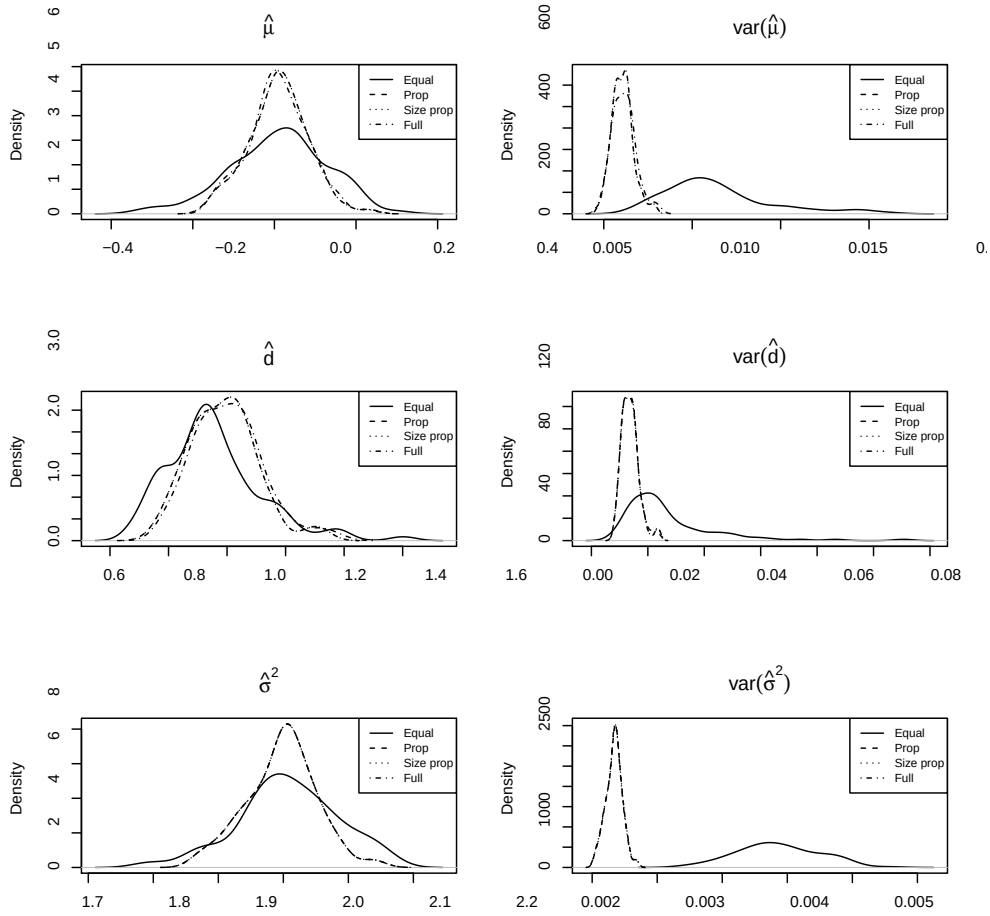

Figure B.6: *First simulation study. Setting 3. Combining the results from the four splits, using equal, proportional, and size proportional weights. This is compared with full maximum likelihood.*

Additionally, to combine the results from sample splitting, the same weights as used in the case study are considered here as well: equal, proportional, approximate scalar, scalar, and approximate optimal. In the case of the approximate optimal weights both simple and proper variances are calculated.

For the multiple imputation based approach, $M = 20$ imputations are considered and the conventional combination rules applied.

Note that the MI approach cannot be used with configurations 1, 4, and 5, because the number of available subjects in the observed dataset is less than the number of

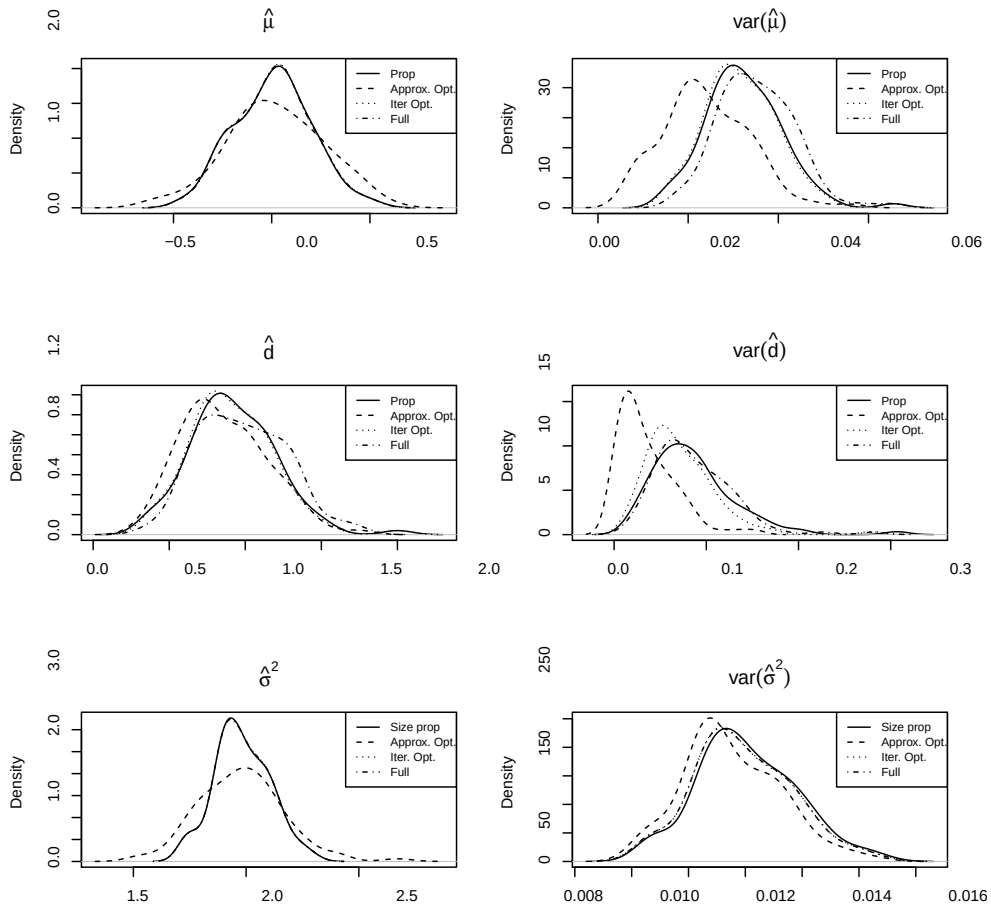

Figure B.7: *First simulation study. Setting 1. (Size) proportional, approximate, and iterated optimal weights, as well as full maximum likelihood.*

repeated measurements, leading to a singular covariance matrix. From the remaining configurations 2 and 3, we have chosen #2, which implies smaller numbers and hence is more challenging.

For each configuration we report three results: the estimated parameters, their standard errors, and the mean square error (MSE). Furthermore, we report computation time.

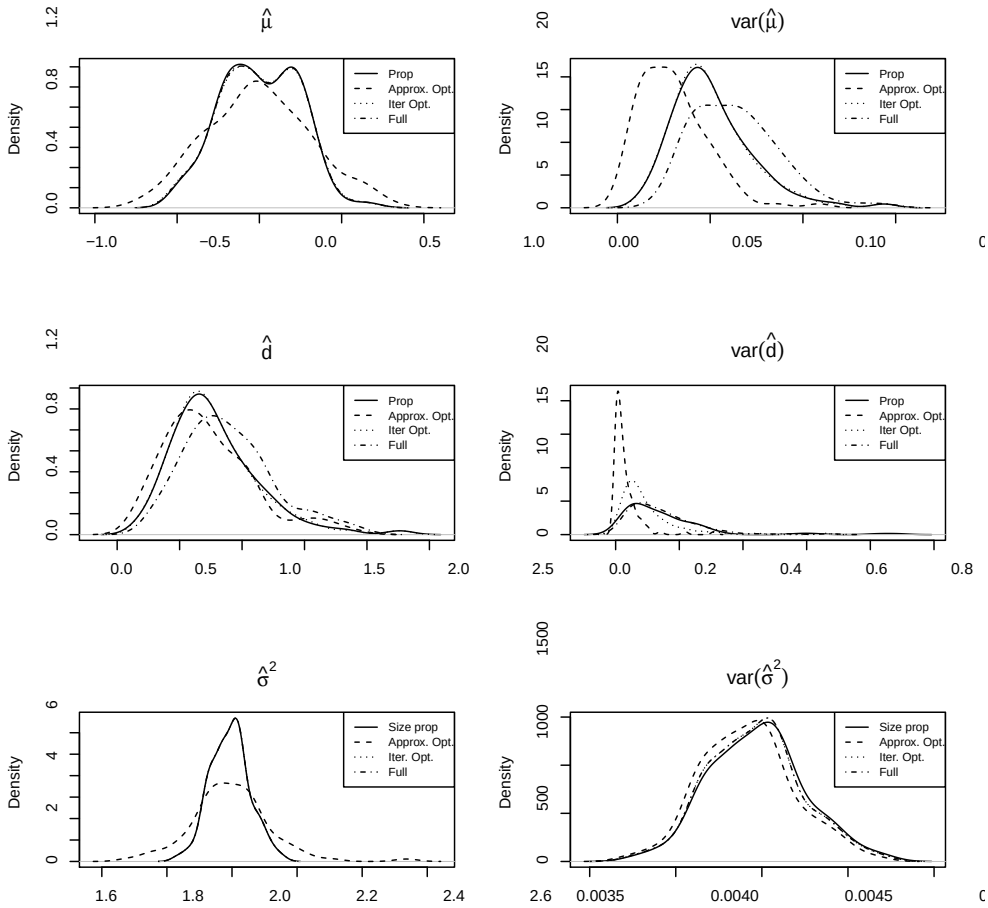

Figure B.8: *First simulation study. Setting 2. (Size) proportional, approximate, and iterated optimal weights, as well as full maximum likelihood.*

B.6.2 Simulation results

Based on the simulation results, it appears that using equal weights is not recommended, while using proportional weights produces results comparable with ML. Of course, in case of σ^2 the approximate scalar weights work better comparing with ML. An interesting outcome of the simulation is that by keeping the number of clusters of different sizes constant, but allowing the cluster sizes to increase, improves estimation of σ^2, while increasing the number of clusters and keeping their sizes constant improves the estimation of d. This is not surprising, because d is the between-cluster variability, which is easier to

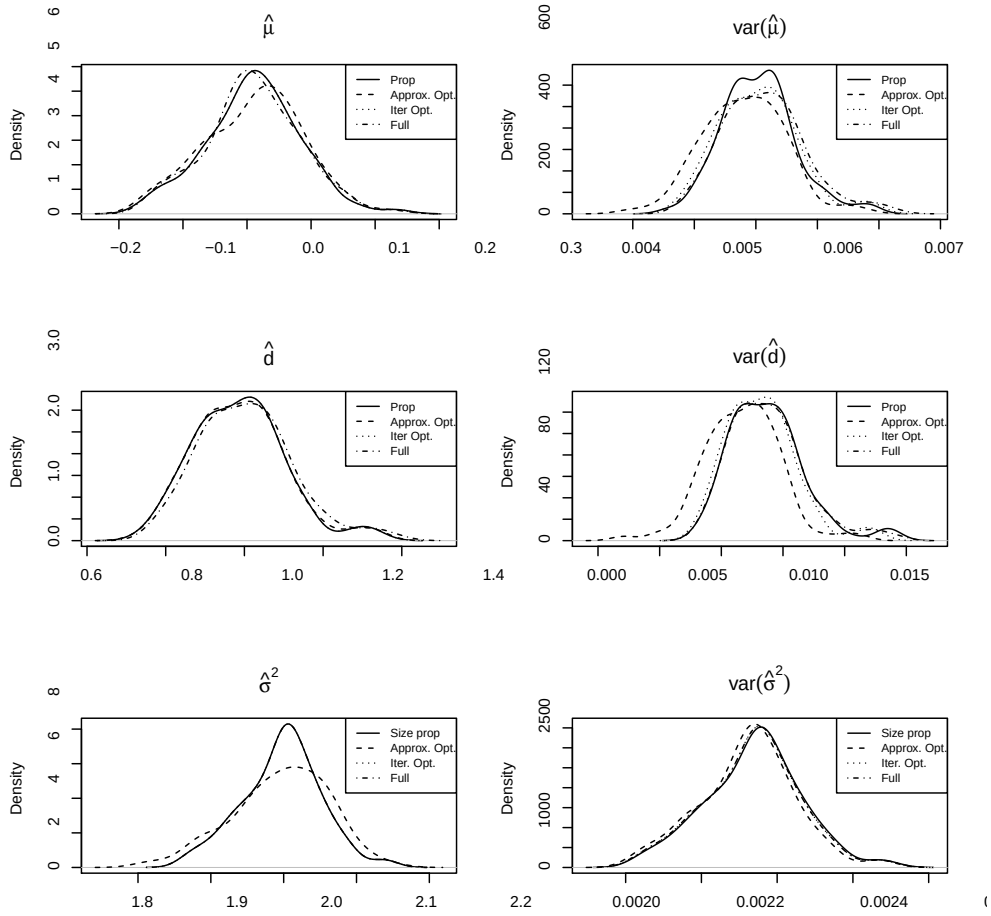

Figure B.9: *First simulation study. Setting 3. (Size) proportional, approximate, and iterated optimal weights, as well as full maximum likelihood.*

estimate from a larger number of clusters. This should be seen against the background of relatively small differences anyway.

The results based on MI are not comparable with sample-splitting results. In particular, the variance component d is underestimated using MI, while σ^2 is overestimated. The larger standard errors in this case suggest that the sample-splitting methods use information more efficiently.

Comparing computation times, the closed-form approaches are the clear winners, further enhanced by smaller standard errors. Furthermore, it follows that computing the estimates in a semi-parallel fashion, thus avoiding 'for' loops within the splits but using

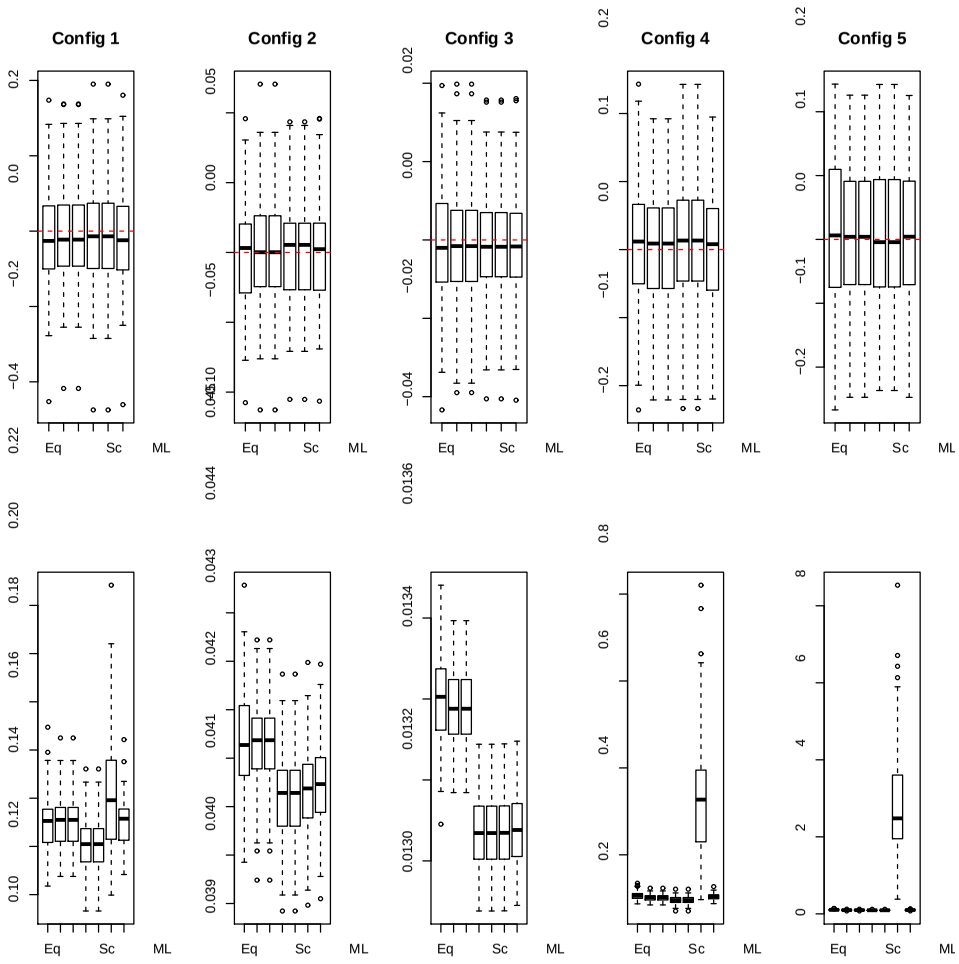

Figure B.10: *Second simulation study. Estimates for μ (first row) and its standard error (second row).*

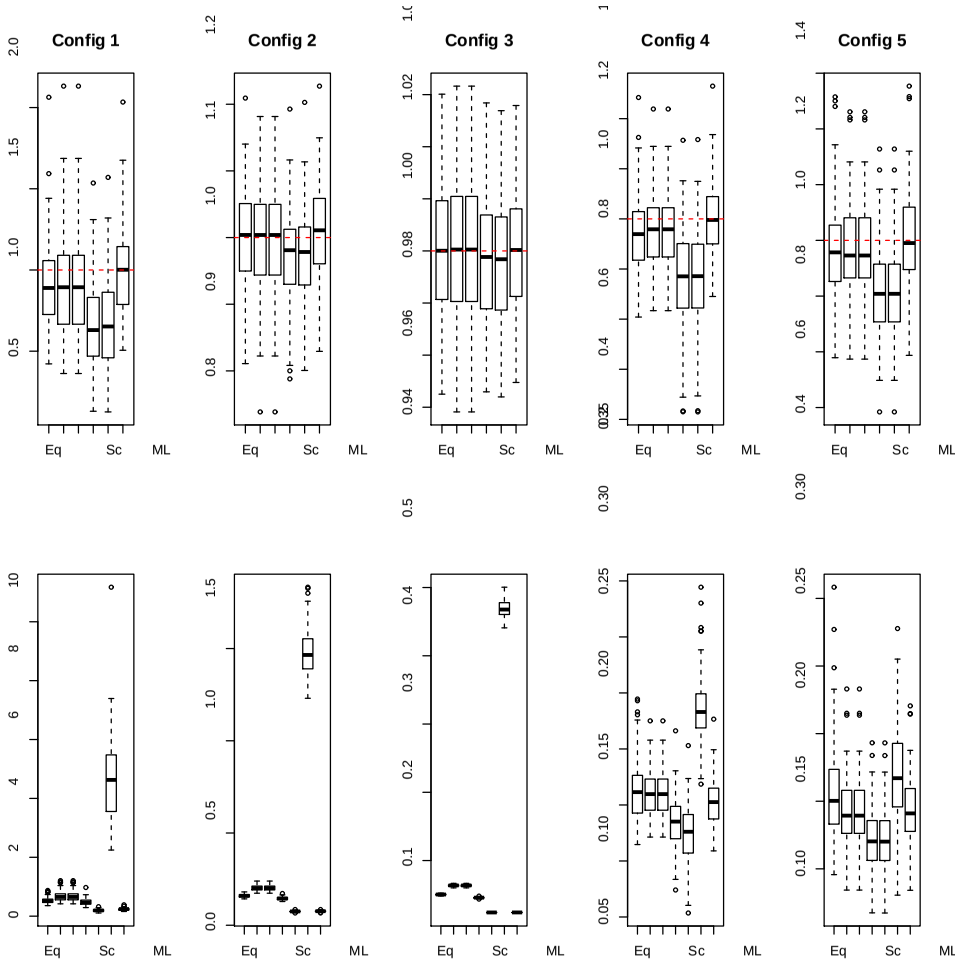

Figure B.11: *Second simulation study. Estimates for d (first row) and standard errors (second row).*

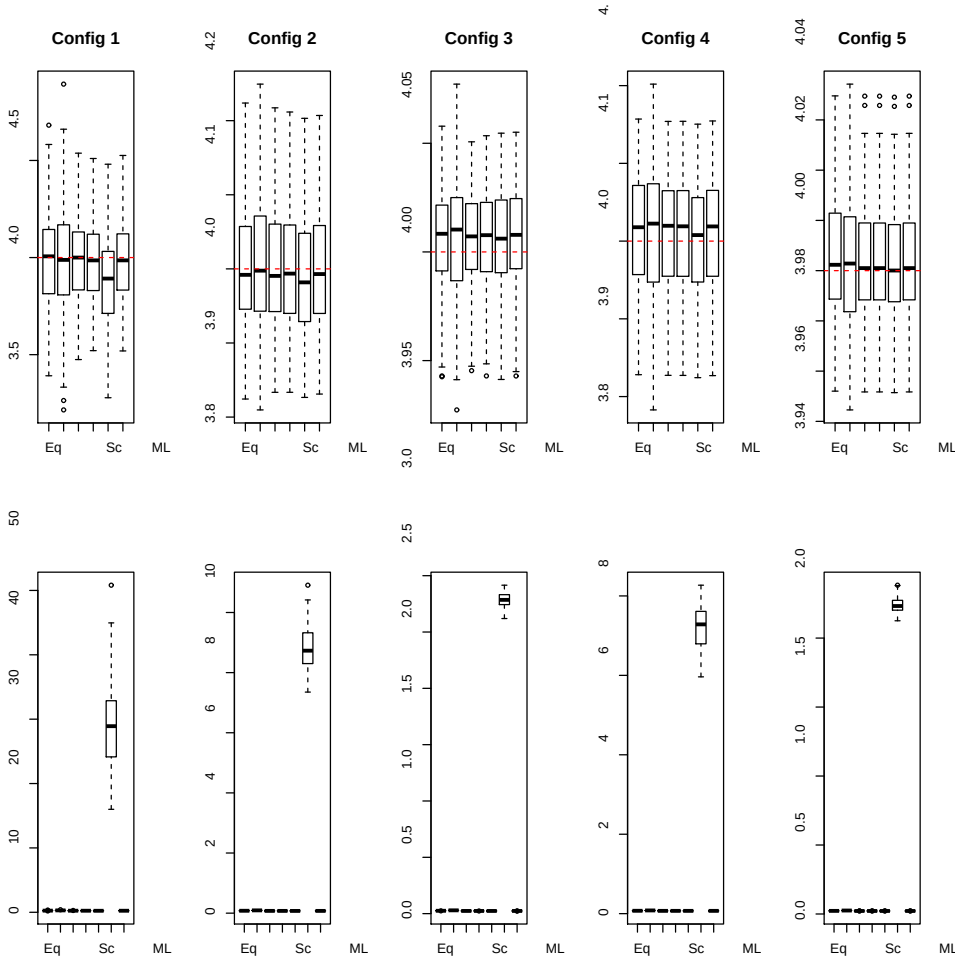

Figure B.12: *Second simulation study. Estimates for σ^2 (first row) and standard errors (second row).*

Table B.4: *Second simulation study. Mean, standard deviation (S.D.) and MSE for μ among 100 replications for each configuration using different combination weights comparing with full sample MLE.*

Config.		Equal	Prop	Approx. sc.	Scalar	Approx. opt.	ML
	Mean	-1.72277E-02	-1.51605E-02	-1.51605E-02	-1.21296E-02	-1.21296E-02	-1.56028E-02
1	S.D.	(1.32751E-01)	(1.28989E-01)	(1.28989E-01)	(1.32435E-01)	(1.32435E-01)	(1.29150E-01)
	MSE	1.77434E-02	1.67015E-02	1.67015E-02	1.75108E-02	1.75108E-02	1.67564E-02
	Mean	-9.39926E-04	6.93419E-04	6.93419E-04	1.27613E-03	1.27613E-03	7.59533E-04
2	S.D.	(3.93526E-02)	(3.95534E-02)	(3.95534E-02)	(3.84321E-02)	(3.84321E-02)	(3.83996E-02)
	MSE	1.53403E-03	1.54930E-03	1.54930E-03	1.46389E-03	1.46389E-03	1.46036E-03
	Mean	-8.30934E-04	-1.25609E-03	-1.25609E-03	-1.26810E-03	-1.26810E-03	-1.31356E-03
3	S.D.	(1.44839E-02)	(1.47545E-02)	(1.47545E-02)	(1.41310E-02)	(1.41310E-02)	(1.41704E-02)
	MSE	2.08376E-04	2.17095E-04	2.17095E-04	1.99298E-04	1.99298E-04	2.00517E-04
	Mean	9.30928E-03	2.53713E-03	2.53713E-03	8.29367E-03	8.29367E-03	2.82086E-03
4	S.D.	(9.26881E-02)	(8.32009E-02)	(8.32009E-02)	(9.76672E-02)	(9.76672E-02)	(8.28999E-02)
	MSE	8.59183E-03	6.85960E-03	6.85960E-03	9.51227E-03	9.51227E-03	6.81163E-03
	Mean	9.77532E-03	1.02173E-02	1.02173E-02	8.72381E-03	8.72381E-03	1.02422E-02
5	S.D.	(1.09769E-01)	(1.04876E-01)	(1.04876E-01)	(1.04982E-01)	(1.04982E-01)	(1.04847E-01)
	MSE	1.20243E-02	1.09934E-02	1.09934E-02	1.09872E-02	1.09872E-02	1.09880E-02

then between splits, is most efficient. Unless the n_k's are very large, computing all estimates at once is the most efficient. Of course, if the estimates for different splits can be done in parallel (without 'for' loops), this is more efficient than estimating them all at once.

B.7 Analysis of the NTP Data Using R

First, this Appendix briefly describes the use of some building blocks in R for the analysis of the case study discussed in Section 3.1. The text file containing these functions is available for download at `www.ibiostat.be`.

- `splitmeth(idvector, yvector)`: computes the size of each cluster, giving the identification vector and the responses. Afterwards it combines the clusters with the same size together in a block.
- `ckblock(idvector, yvector)`: Identification vectors and outcomes isolated per cluster size in the previous function are used in this function seperately. For each block the cluster size, the size of the block, the mean, variance, correlation and the variance-covariance matrix are computed.
- `estimators(idvector, yvector)`: computes all the estimators mentioned in this chapter, given the identification vector and the outcomes. It uses the two functions

Table B.5: *Second simulation study. Mean and standard deviation (S.D.) for standard errors of μ estimates in 100 replications for each configuration using different combination weights comparing with full sample MLE.*

Config.		Equal	Prop	Approx. sc.	Scalar	Simple opt.	Proper opt.	ML
1	Mean	1.29844E-01	1.29733E-01	1.29733E-01	1.20593E-01	1.20593E-01	1.41085E-01	1.29648E-01
	S.D.	(1.18175E-02)	(1.07496E-02)	(1.07496E-02)	(1.09122E-02)	(1.09122E-02)	(2.25155E-02)	(9.95614E-03)
2	Mean	4.23655E-02	4.23056E-02	4.23056E-02	4.11657E-02	4.11657E-02	4.12635E-02	4.14298E-02
	S.D.	(1.08386E-03)	(8.88747E-04)	(8.88747E-04)	(8.71465E-04)	(8.71465E-04)	(8.90259E-04)	(8.66432E-04)
3	Mean	1.34008E-02	1.33822E-02	1.33822E-02	1.30725E-02	1.30725E-02	1.30728E-02	1.30799E-02
	S.D.	(1.21880E-04)	(9.92324E-05)	(9.92324E-05)	(9.68871E-05)	(9.68871E-05)	(9.68974E-05)	(9.62652E-05)
4	Mean	1.06373E-01	1.01232E-01	1.01232E-01	9.60807E-02	9.60807E-02	3.30358E-01	1.03382E-01
	S.D.	(9.80278E-03)	(7.26031E-03)	(7.26031E-03)	(8.36242E-03)	(8.36242E-03)	(1.39042E-01)	(7.17708E-03)
4	Mean	1.05176E-01	9.84615E-02	9.84615E-02	9.45066E-02	9.45066E-02	2.81427E+00	1.00533E-01
	S.D.	(1.10711E-02)	(8.18259E-03)	(8.18259E-03)	(8.14229E-03)	(8.14229E-03)	(1.41211E+00)	(8.28537E-03)

above, `splitmeth` to split the data according to cluster size and `ckblock`to compute the estimators in each block. The function gives the estimators of μ, σ^2 and d together with their precision.

Next, the R functions to compute CS sample splitting estimates and variances are described. Section B.7.1 will describe the input of the functions, the output of each function will de described in Section B.7.2. Each function is followed by an example as well. The functions themselves are presented in Section B.7.3.

There are three functions provided to estimate CS parameters (μ, σ^2, d). The function `est.CS` estimates the CS parameters and their variances using vectorized (semi-parallel) calculations, i.e. the for loops are avoided for calculation within each split. The function `est.CS.for` implements the formulas in 5.9 directly using for loops. The function `est.CS.all` estimates the parameters for all of the splits simultaneously.

The function `param.free.CS` can be used to combine results from different sub-samples using parameter-free weights: Prop., Equal, and Appr.sc. The function `scalar.weights.CS` gives the same results but using scalar weights (approximated by sub-sample specific estimates). The function `approx.optimal.CS` combines the results of different sub-samples using approximated optimal weights, the proper variances for these estimates are also provided. Finally, the function `clusterBYcluster.CS` computes the cluster specific estimate of (μ, σ^2, d) using cluster-by-cluster approaches. Combining them by the desired rule using the three previous functions is straightforward.

B.7.1 Input

Table B.13 describes the input for the various functions.

Table B.6: *Second simulation study. Mean, standard deviation (S.D.) and MSE for d estimates in 100 replications for each configuration using different combination weights comparing with full sample MLE.*

Config.		Equal	Prop	Approx. sc.	Scalar	Approx. opt	ML
	Mean	9.09580E-01	9.00788E-01	9.00788E-01	6.63293E-01	6.71111E-01	9.86549E-01
1	S.D.	(2.66885E-01)	(2.99736E-01)	(2.99736E-01)	(2.70556E-01)	(2.79060E-01)	(2.55874E-01)
	MSE	7.86910E-02	9.87862E-02	9.87862E-02	1.85840E-01	1.85264E-01	6.49975E-02
	Mean	9.98984E-01	9.97534E-01	9.97534E-01	9.72269E-01	9.73771E-01	1.00712E+00
2	S.D.	(7.44958E-02)	(8.28143E-02)	(8.28143E-02)	(7.28186E-02)	(7.22762E-02)	(7.08463E-02)
	MSE	5.49515E-03	6.79570E-03	6.79570E-03	6.01854E-03	5.85958E-03	5.01969E-03
	Mean	1.00004E+00	9.99697E-01	9.99697E-01	9.97218E-01	9.97153E-01	1.00019E+00
3	S.D.	(2.61046E-02)	(2.82480E-02)	(2.82480E-02)	(2.48989E-02)	(2.48068E-02)	(2.44715E-02)
	MSE	6.74637E-04	7.90063E-04	7.90063E-04	6.21496E-04	6.17332E-04	5.92900E-04
	Mean	9.40257E-01	9.50756E-01	9.50756E-01	7.65219E-01	7.65362E-01	9.96005E-01
4	S.D.	(1.53414E-01)	(1.48216E-01)	(1.48216E-01)	(1.91865E-01)	(1.91614E-01)	(1.49451E-01)
	MSE	2.68697E-02	2.41734E-02	2.41734E-02	9.15661E-02	9.14038E-02	2.21282E-02
	Mean	9.63459E-01	9.68182E-01	9.68182E-01	8.21266E-01	8.21270E-01	1.00958E+00
5	S.D.	(1.73767E-01)	(1.63471E-01)	(1.63471E-01)	(1.72162E-01)	(1.72163E-01)	(1.69235E-01)
	MSE	3.12283E-02	2.74677E-02	2.74677E-02	6.12894E-02	6.12881E-02	2.84459E-02

B.7.2 Output

The output of each function is a list (except for `scalar.weights.CS`) containing the calculated quantities. The functions `est.CS` and `est.CS.for` compute the parameters estimates with their variances within each split. Table B.14 presents the output for these two functions. An example of using them is given in Example 22. For estimating the parameters for all of the splits simultaneously, function `est.CS.all` can be used. The output description of this function can be found in Table B.15. An example of using it is presented in Example 23. One may denote that the input dataset for using this function should make all clusters of equal size using missing values `NaN`. The function `param.free.CS` computes the three parameter free combining rules: Prop., Equal, and Appr.sc., Table B.16 presents the output of this function, an example of using it is given in Example 24. For computing the scalar weights one may use the function `scalar.weights.CS`. The output of this function are described in Table B.17 and Example 25 shows how to use it in practice. Function `approx.optimal.CS` computes the approximate optimal weights together with their proper variances. Table B.18 gives the output of this function and Example 26 will show the use of this function. Finally, for cluster-by-cluster analyses, the three methods discussed (weighted, two stage, unbiased two stage) are implemented in the function `clusterBYcluster.CS`. One may find descriptions of the output of this function together with an example of using it in Table B.19 and Example 27, respectively.

Table B.7: *Second simulation study. Mean and standard deviation (S.D.) for standard errors of d estimates in 100 replications for each configuration using different combination weights comparing with full sample MLE.*

Config.		Equal	Prop.	Approx. sc.	Scalar	Simple opt.	Proper opt.	ML
1	Mean	5.30888E-01	6.76159E-01	6.76159E-01	4.83519E-01	1.95328E-01	4.66561E+00	2.36408E-01
	S.D.	(1.00495E-01)	(1.66365E-01)	(1.66365E-01)	(1.10589E-01)	(4.01066E-02)	(1.40173E+00)	(3.91752E-02)
2	Mean	1.60365E-01	2.03126E-01	2.03126E-01	1.45284E-01	7.48056E-02	1.47955E+00	7.60798E-02
	S.D.	(8.29444E-03)	(1.42616E-02)	(1.42616E-02)	(8.70034E-03)	(3.35411E-03)	(1.31587E-01)	(3.26483E-03)
3	Mean	5.02596E-02	6.34861E-02	6.34861E-02	4.56061E-02	2.39463E-02	4.67989E-01	2.39748E-02
	S.D.	(8.43976E-04)	(1.44201E-03)	(1.44201E-03)	(8.34510E-04)	(3.68251E-04)	(1.24414E-02)	(3.66865E-04)
4	Mean	1.62410E-01	1.59656E-01	1.59656E-01	1.34256E-01	1.24552E-01	2.35351E-01	1.51740E-01
	S.D.	(2.83009E-02)	(2.05482E-02)	(2.05482E-02)	(2.31559E-02)	(2.51091E-02)	(3.00109E-02)	(2.11541E-02)
5	Mean	1.54122E-01	1.43313E-01	1.43313E-01	1.22117E-01	1.22024E-01	1.70868E-01	1.43764E-01
	S.D.	(3.48725E-02)	(2.50371E-02)	(2.50371E-02)	(2.30972E-02)	(2.31064E-02)	(3.70951E-02)	(2.38992E-02)

Example 22.

```
> est.CS(n,C,Y)
$mu.hat
[1] -0.02008066

$d.hat
[1] 1.05208

$sigma2.hat
[1] 3.893208

$var.mu.hat
[1] 0.01025821

$cov.varcomp
             [,1]         [,2]
[1,]  0.028870611 -0.003608826
[2,] -0.003608826  0.032020343
```

Example 23.

```
> est.CS.all(data2,ck,nk)
$mu.hat
             [,1]
[1,]  0.017154093
```

Table B.8: *Second simulation study. Mean, standard deviation (S.D.) and MSE for σ^2 estimates in 100 replications for each configuration using different combination weights comparing with full sample MLE.*

Config.		Equal	Prop.	Approx. sc.	Scalar	Approx. opt.	ML
	Mean	3.98608E+00	3.99739E+00	3.98571E+00	3.98364E+00	3.87487E+00	3.98075E+00
1	S.D.	(2.50882E-01)	(2.95821E-01)	(2.35138E-01)	(2.33650E-01)	(2.38167E-01)	(2.30477E-01)
	MSE	6.25062E-02	8.66420E-02	5.49414E-02	5.43140E-02	7.18141E-02	5.29588E-02
	Mean	4.00184E+00	4.00681E+00	4.00177E+00	4.00087E+00	3.99027E+00	4.00064E+00
2	S.D.	(7.85190E-02)	(9.05849E-02)	(7.57443E-02)	(7.56486E-02)	(7.50477E-02)	(7.51739E-02)
	MSE	6.10698E-03	8.16998E-03	5.68296E-03	5.66624E-03	5.67055E-03	5.59502E-03
	Mean	4.00509E+00	4.00491E+00	4.00575E+00	4.00590E+00	4.00472E+00	4.00587E+00
3	S.D.	(2.45760E-02)	(2.82395E-02)	(2.36027E-02)	(2.36983E-02)	(2.37694E-02)	(2.36697E-02)
	MSE	6.23828E-04	8.13646E-04	5.84605E-04	5.90806E-04	5.81652E-04	5.89081E-04
	Mean	4.01402E+00	4.01190E+00	4.01302E+00	4.01304E+00	4.00363E+00	4.01306E+00
4	S.D.	(7.69655E-02)	(8.78962E-02)	(7.32088E-02)	(7.31202E-02)	(7.20589E-02)	(7.31207E-02)
	MSE	6.06101E-03	7.79021E-03	5.47558E-03	5.46319E-03	5.15372E-03	5.46370E-03
	Mean	4.00346E+00	4.00338E+00	4.00292E+00	4.00292E+00	4.00192E+00	4.00292E+00
5	S.D.	(2.56561E-02)	(2.79741E-02)	(2.46670E-02)	(2.46664E-02)	(2.46901E-02)	(2.46669E-02)
	MSE	6.63599E-04	7.86128E-04	6.10884E-04	6.10854E-04	6.07187E-04	6.10909E-04

```
[2,]  0.043479563
[3,] -0.156958273
[4,] -0.031153966
[5,] -0.007420771

$sigma2.hat
[1] 3.893208 4.085951 3.901042 4.015487 3.900566

$d.hat
[1] 1.0506935 1.0162879 0.7930084 0.8279435 0.9344471

$var.mu.hat
[1] 0.010248964 0.007333913 0.006977852 0.006370544 0.011944848

$var.d.hat
[1] 0.03196348 0.02822874 0.03485060 0.01648236 0.02863248

$var.sigma2.hat
[1] 0.02887061 0.03339000 0.05072710 0.02015517 0.02173487
```

Table B.9: *Second simulation study. Mean and standard deviation (S.D.) for standard errors of σ^2 estimates in 100 replications for each configuration using different combination weights comparing with full sample MLE.*

Config.		Equal	Prop.	Approx. sc.	Scalar	Simple opt.	Proper opt.	ML
1	Mean	2.53758E-01	2.95524E-01	2.40482E-01	2.38699E-01	2.30855E-01	2.88730E+01	2.35804E-01
	S.D.	(1.98892E-02)	(3.33124E-02)	(1.45355E-02)	(1.39849E-02)	(1.38813E-02)	(6.93951E+00)	(1.34905E-02)
2	Mean	7.98210E-02	9.26570E-02	7.57927E-02	7.53242E-02	7.48395E-02	8.82494E+00	7.49872E-02
	S.D.	(1.84469E-03)	(3.08841E-03)	(1.45443E-03)	(1.42354E-03)	(1.40575E-03)	(6.92140E-01)	(1.39989E-03)
3	Mean	2.52202E-02	2.92034E-02	2.39687E-02	2.38353E-02	2.37400E-02	2.78559E+00	2.37444E-02
	S.D.	(1.85701E-04)	(3.12856E-04)	(1.42587E-04)	(1.40888E-04)	(1.40071E-04)	(6.70930E-02)	(1.39784E-04)
4	Mean	7.22378E-02	8.07793E-02	7.01684E-02	7.01663E-02	6.99970E-02	7.23107E+00	7.01238E-02
	S.D.	(1.54384E-03)	(2.35043E-03)	(1.28712E-03)	(1.28385E-03)	(1.26263E-03)	(5.28710E-01)	(1.27765E-03)
5	Mean	2.25782E-02	2.52054E-02	2.19702E-02	2.19702E-02	2.19647E-02	2.23651E+00	2.19689E-02
	S.D.	(1.55999E-04)	(2.29649E-04)	(1.35364E-04)	(1.35358E-04)	(1.35462E-04)	(5.50362E-02)	(1.35379E-04)

```
$cov.d.sigma2.hat
[1] -0.003608826 -0.006678000 -0.016909033 -0.002239464
-0.001448992
```

Example 24.

```
> param.free.CS (nk,ck,mu.split.est,sigma2.split.est,
                 d.split.est,mu.split.var,sigma2.split.var,
                 d.split.var)
$mu
                Est          Var
Equal     0.8486860 0.0001863644
Prop      0.8350032 0.0001755510
Appr.sc. 0.8350032 0.0001755510

$sigma2
                  Est          Var
Equal     0.008473228 2.368028e-07
Prop      0.008471484 1.973433e-07
Appr.sc. 0.008327288 1.725952e-07

$d
                 Est          Var
Equal     0.01443741 5.351632e-06
Prop      0.01538224 5.861071e-06
```

Table B.10: *Second simulation study. Computation time (in seconds) using closed-form solutions with different implementation forms, compared to PROC MIXED.*

		Split by split			PROC
Config.		without loops	using loops	Together	MIXED
1	Mean	**0.00520**	0.00980	0.00640	0.34155
	S.D.	(0.00882)	(0.00943)	(0.00823)	(0.17711)
2	Mean	**0.02150**	0.07340	0.05660	0.35575
	S.D.	(0.01617)	(0.02006)	(0.09662)	(0.03073)
3	Mean	**0.17980**	0.73480	0.43400	0.81783
	S.D.	(0.02292)	(0.05835)	(0.11543)	(0.02582)
4	Mean	**0.00220**	0.00610	0.00360	2.58808
	S.D.	(0.00579)	(0.01497)	(0.00689)	(0.40720)
5	Mean	0.04030	0.27490	**0.02130**	629.58333
	S.D.	(0.01521)	(0.01941)	(0.00872)	(116.09435)

```
Appr.sc. 0.01538224 5.861071e-06
```

Example 25.

```
scalar.weights.CS (nk,ck,mu.split.est,sigma2.split.est,
                   d.split.est,mu.split.var,sigma2.split.var,
                   d.split.var)
              Est.          Var
mu     0.841588794 1.700247e-04
sigma2 0.008313578 1.716135e-07
d      0.013715408 4.986101e-06
```

Example 26.

```
approx.optimal.CS (nk,ck,sigma2.split.est,d.split.est,
                   sigma2.split.var,d.split.var)
$mu.est
[1] 0.8415888

$varcomp.est
```

Table B.11: *Second simulation study. Mean, standard deviation (S.D.) and MSE for CS parameter estimates in 100 replications for configuration 2 using different combination weights comparing with full sample MLE and MI-MLE.*

		Equal	Prop.	Approx. sc.	Scalar	Approx. opt.	MI	ML
	Mean	-5.09643E-03	-2.62061E-03	-2.62061E-03	-2.88933E-03	-2.88933E-03	-8.04195E-03	-3.28224E-03
μ	S.D.	(4.85424E-02)	(4.91772E-02)	(4.91772E-02)	(4.68032E-02)	(4.68032E-02)	(6.00662E-02)	(4.71723E-02)
	MSE	2.35877E-03	2.40108E-03	2.40108E-03	2.17698E-03	2.17698E-03	3.63654E-03	2.21375E-03
	Mean	9.98392E-01	9.95216E-01	9.95216E-01	9.70589E-01	9.71627E-01	3.51123E-02	9.92960E-01
d	S.D.	(7.33193E-02	(7.46946E-02)	(7.46946E-02)	(7.59516E-02)	(7.53362E-02)	(1.83983E-02)	(7.50610E-02)
	MSE	5.32455E-03	5.54638E-03	5.54638E-03	6.57598E-03	6.42380E-03	9.31343E-01	5.62737E-03
	Mean	4.00782E+00	4.00791E+00	4.00581E+00	4.00544E+00	3.99400E+00	5.32627E+00	4.00347E+00
σ^2	S.D.	(7.66500E-02)	(8.98881E-02)	(7.19708E-02)	(7.13441E-02)	(7.19080E-02)	(1.55770E-01)	(7.56968E-02)
	MSE	5.87763E-03	8.06169E-03	5.16181E-03	5.06871E-03	5.15505E-03	1.78301E+00	5.68472E-03

Table B.12: *Second simulation study. Mean and standard deviation (S.D.) for the standard error of CS parameter estimates in 100 replications for configuration 2 using different combination weights comparing with full sample MLE and MI-MLE.*

		Equal	Prop.	Approx. sc.	Scalar	Simple opt.	Proper opt.	MI	ML
μ	Mean	4.24162E-02	4.22766E-02	4.22766E-02	4.11356E-02	4.11356E-02	4.12439E-02	4.95431E-02	4.14004E-02
	S.D.	(1.21548E-03)	(8.59430E-04)	(8.59430E-04)	(9.31222E-04)	(9.31222E-04)	(9.42598E-04)	(7.65032E-03)	(9.11785E-04)
d	Mean	1.60545E-01	2.02922E-01	2.02922E-01	1.45623E-01	7.47531E-02	1.48865E+00	9.10211E-02	7.54391E-02
	S.D.	(8.88524E-03)	(1.47416E-02)	(1.47416E-02)	(9.84170E-03)	(3.76197E-03)	(1.29914E-01)	(5.18621E-03)	(3.38044E-03)
σ^2	Mean	7.99442E-02	9.26315E-02	7.58626E-02	7.54139E-02	7.49171E-02	8.86450E+00	1.64310E-01	7.50377E-02
	S.D.	(1.87358E-03)	(3.14842E-03)	(1.39531E-03)	(1.34244E-03)	(1.33807E-03)	(6.56714E-01)	(2.27104E-02)	(1.42688E-03)

```
[1] 0.006059707 0.014069710

$mu.var
[1] 0.0001771469

$varcomp.var
              [,1]          [,2]
[1,]  7.503554e-08 -1.089792e-08
[2,] -1.089792e-08  5.144086e-06

$proper.var.mu
[1] 0.0001771471

$proper.var.varcomp
              [,1]          [,2]
[1,]  8.045528e-06 -9.606275e-06
```

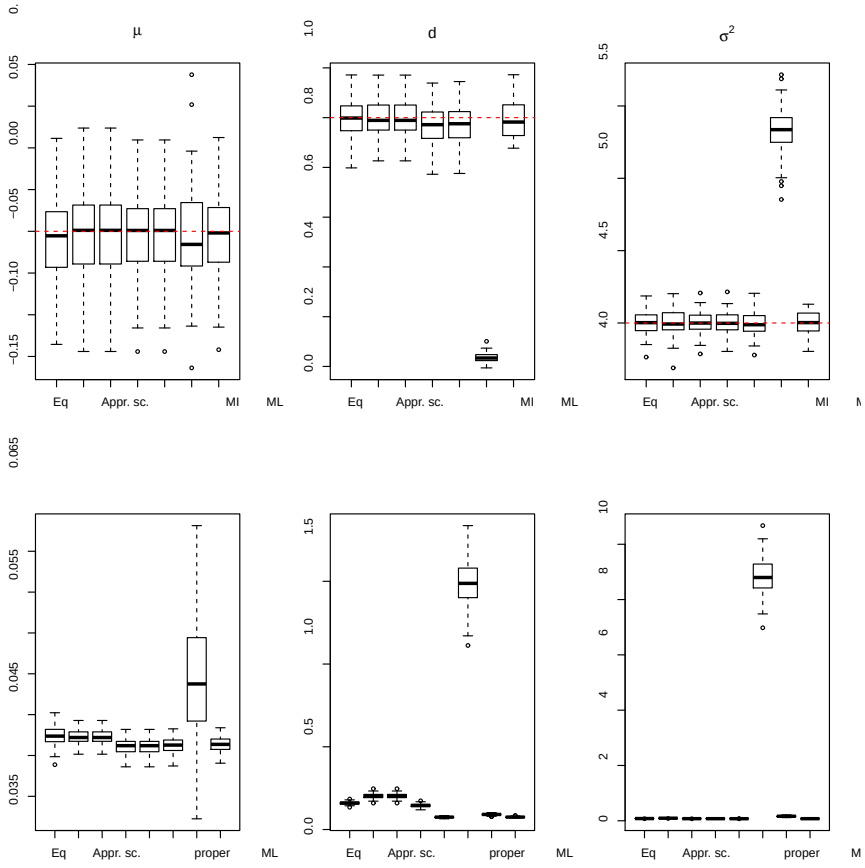

Figure B.13: *Second simulation study. Estimated CS parameters (first row) and their standard error (second row) using sample splitting, MI-MLE, and MLE.*

```
[2,] -9.606275e-06  2.111345e-05
```

One may note that once the cluster-specific estimates have been obtained using any of the cluster-by-cluster approaches, one can easily combine them with the desired rule using the available functions. The output are organized in accordance with the input of other functions, so they can be directly plugged in.

Example 27.

```
> clusterBYcluster.CS(nk,ck,Data)
$mu.split.est
```

Table B.13: *Function input*.

Input	Description
`n`	an integer indicating the common cluster size within each split
`C`	an integer indicating the number of clusters within each split
`Y`	a vector containing the response values corresponding to each split
`data`	a 2-column matrix with response variable as its first column and splits indicator as its second column. All clusters should be of equal size by placing `NaN` to make smaller clusters the same size as the largest one.
`nk`	a vector containing the cluster sizes
`ck`	a vector containing the number of clusters of size n_k
`mu.split.est`	a vector containing the $\widehat{\mu}$'s from all sub-samples.
`sigma2.split.est`	a vector containing the $\widehat{\sigma}^2$'s from all sub-samples.
`d.split.est`	a vector containing the $\widehat{d}$'s from all sub-samples.
`mu.split.var`	a vector containing the $\mathrm{Var}(\widehat{\mu})$'s from all sub-samples.
`sigma2.split.var`	a vector containing the $\mathrm{Var}(\widehat{\sigma}^2)$'s from all sub-samples.
`d.split.var`	a vector containing the $\mathrm{Var}(\widehat{d})$'s from all sub-samples.
`Data`	(only applicable in function `clusterBYcluster.CS`) a 3-column matrix with first column the subject, second column the response variable, and third column the split indexes, which show which observation belongs to which sub-sample.

```
[1] 0.8058733 0.8500635 0.8350303 0.8645111 0.9579167 0.7938929
0.8111373  0.8538187

$mu.split.var
        Weighted    TwoStage TwoStageUnbiased
[1,] 0.001307539 0.001376916      0.001478020
[2,] 0.001650106 0.001810324      0.002043291
[3,] 0.001334065 0.001390168      0.001517057
[4,] 0.001092201 0.001189274      0.001332732
[5,] 0.003312091 0.003506955      0.005190838
[6,] 0.003285979 0.003373701      0.003849875
```

Table B.14: *Function output:* `est.CS` *and* `est.CS.for`

Output	Description
`mu.hat`	a scalar presenting $\widehat{\mu}$.
`d.hat`	a scalar presenting $\widehat{d}$.
`sigma2.hat`	a scalar presenting $\widehat{\sigma^2}$.
`var.mu.hat`	a scalar presenting variance of $\widehat{\mu}$.
`cov.varcomp`	a matrix presenting covariance matrix of $(\widehat{\sigma^2}, \widehat{d})$

Table B.15: *Function output:* `est.CS.all`

Output	Description
`mu.hat`	a vector presenting $\widehat{\mu}$ for each split.
`d.hat`	a vector presenting $\widehat{d}$ for each split.
`sigma2.hat`	a vector presenting $\widehat{\sigma^2}$ for each split.
`var.mu.hat`	a vector presenting variance of $\widehat{\mu}$ for each split.
`var.d.hat`	a vector presenting variance of $\widehat{d}$ for each split.
`var.sigma2.hat`	a vector presenting variance of $\widehat{\sigma^2}$ for each split.
`cov.d.sigma2.hat`	a vector presenting covariance of $(\widehat{\sigma^2}, \widehat{d})$ for each split.

```
[7,] 0.001098199 0.001127220      0.001198495
[8,] 0.001034920 0.001055199      0.001136498

$sigma2.split.est
        Weighted    TwoStage TwoStageUnbiased
[1,] 0.011562832 0.010406549      0.011562832
[2,] 0.011776069 0.010093773      0.011776069
[3,] 0.008146211 0.007405646      0.008146211
[4,] 0.014040871 0.013104813      0.014040871
[5,] 0.005344821 0.004676719      0.005344821
[6,] 0.010580691 0.009824927      0.010580691
[7,] 0.006458361 0.005920164      0.006458361
[8,] 0.003998299 0.003690738      0.003998299

$sigma2.split.var
         Weighted     TwoStage TwoStageUnbiased
```

Table B.16: *Function output:* `param.free.CS`

Output	Description
`mu`	a matrix with columns $\tilde{\mu}$ and $\mathrm{Var}(\tilde{\mu})$ for Equal, Proper, and Appr.sc. weights, rows 1–3, respectively.
`sigma2`	a matrix with columns $\tilde{\sigma}^2$ and $\mathrm{Var}(\tilde{\sigma}^2)$ for Equal, Proper, and Appr.sc. weights, rows 1–3, respectively.
`d`	a matrix with columns $\tilde{d}$ and $\mathrm{Var}(\tilde{d})$ for Equal, Proper, and Appr.sc. weights, rows 1–3, respectively.

Table B.17: *Function output:* `scalar.weights.CS`

Output	Description
-	a matrix with columns $\tilde{\mu}$ and $\mathrm{Var}(\tilde{\mu})$ for scalar weights.

```
[1,] 1.980727e-06 1.299555e-06     1.980727e-06
[2,] 5.136141e-06 2.772361e-06     5.136141e-06
[3,] 1.106012e-06 7.554214e-07     1.106012e-06
[4,] 3.129302e-06 2.374623e-06     3.129302e-06
[5,] 2.720678e-06 1.594811e-06     2.720678e-06
[6,] 2.152904e-06 1.600612e-06     2.152904e-06
[7,] 4.461008e-07 3.149770e-07     4.461008e-07
[8,] 1.903143e-07 1.381729e-07     1.903143e-07

$d.split.est
        Weighted    TwoStage TwoStageUnbiased
[1,] 0.018456799 0.019613082       0.02101402
[2,] 0.013168657 0.014850953       0.01670732
[3,] 0.015268213 0.016008777       0.01746412
[4,] 0.008893749 0.009829807       0.01105853
[5,] 0.009268172 0.009936274       0.01490441
[6,] 0.025532065 0.026287829       0.03004323
[7,] 0.018131189 0.018669386       0.01983622
[8,] 0.014181322 0.014488883       0.01560341

$d.split.var
```

Table B.18: *Function output:* `approx.optimal.CS`

Output	Description
`mu.est`	$\tilde{\mu}$ obtained by approximate optimal weights
`varcomp.est`	$(\tilde{\sigma}^2, \tilde{d})$ obtained by approximated optimal weights
`mu.var`	$\mathrm{Var}(\tilde{\mu})$ obtained by approximate optimal weights
`varcomp.var`	covariance matrix of $(\tilde{\sigma}^2, \tilde{d})$ obtained by approximated optimal weights
`proper.var.mu`	$\mathrm{Var}_{\mathrm{Proper}}(\tilde{\mu})$ obtained by approximate optimal weights
`proper.var.varcomp`	proper covariance matrix of $(\tilde{\sigma}^2, \tilde{d})$ obtained by approximate optimal weights

```
         Weighted    TwoStage TwoStageUnbiased
[1,] 5.130954e-05 0.007715211     6.553629e-05
[2,] 4.911611e-05 0.008261736     7.515067e-05
[3,] 4.272263e-05 0.005140078     5.523509e-05
[4,] 2.148616e-05 0.003950645     3.197116e-05
[5,] 6.586221e-05 0.002231365     1.616688e-04
[6,] 1.727735e-04 0.006083359     2.371446e-04
[7,] 4.100850e-05 0.005122242     4.883729e-05
[8,] 2.999080e-05 0.003263327     3.616558e-05

$Var.varcomp
$Var.varcomp[[1]]
              [,1]          [,2]
[1,]  1.980727e-06 -1.980727e-07
[2,] -1.980727e-07  5.130954e-05

$Var.varcomp[[2]]
              [,1]          [,2]
[1,]  5.136141e-06 -7.337344e-07
[2,] -7.337344e-07  4.911611e-05

$Var.varcomp[[3]]
              [,1]          [,2]
[1,]  1.106012e-06 -1.005466e-07
```

Table B.19: *Function output:* `clusterBYcluster.CS`

Output	Description
`mu.split.est`	a vector containing cluster-by-cluster estimated μ for each cluster
`mu.split.var`	a matrix with rows the variances of estimated μ for each cluster and with columns indicating the three methods: weighted, two-stage, and unbiased two-stage
`sigma2.split.est`	a matrix with rows the estimated σ^2 for each cluster and columns indicating the three methods: weighted ($\widehat{\sigma}^2$), two-stage (s^2) and unbiased two-stage (s_*^2)
`sigma2.split.var`	a matrix with rows the variances of estimated σ^2 for each cluster and its columns indicate the three methods: weighted ($\widehat{\sigma}^2$), two stage (s^2) and unbiased two stage (s_*^2)
`d.split.est`	a matrix with rows the estimated d^2 for each cluster and columns indicating the three methods: weighted ($\widehat{\sigma}^2$), two-stage (t^2), and unbiased two-stage (t_*^2)
`d.split.var`	a matrix with rows the variances of estimated d^2 for each cluster and columns indicating the three methods: weighted ($\widehat{\sigma}^2$), two-stage (t^2) and unbiased two-stage (t_*^2)
`Var.varcomp`	a list containing matrices as its elements. The matrices are the full covariance matrices for $(\widehat{\sigma}^2, \widehat{d})$ from a weighted cluster-by-cluster analysis

```
[2,] -1.005466e-07  4.272263e-05

$Var.varcomp[[4]]
              [,1]          [,2]
[1,]  3.129302e-06 -2.086202e-07
[2,] -2.086202e-07  2.148616e-05

$Var.varcomp[[5]]
              [,1]          [,2]
[1,]  2.720678e-06 -3.400847e-07
[2,] -3.400847e-07  6.586221e-05
```

```
$Var.varcomp[[6]]
              [,1]          [,2]
[1,]  2.152904e-06 -1.537789e-07
[2,] -1.537789e-07  1.727735e-04

$Var.varcomp[[7]]
              [,1]          [,2]
[1,]  4.461008e-07 -3.717507e-08
[2,] -3.717507e-08  4.100850e-05

$Var.varcomp[[8]]
              [,1]          [,2]
[1,]  1.903143e-07 -1.463956e-08
[2,] -1.463956e-08  2.999080e-05
```

B.7.3 R Functions

Here are the R functions.

```
est.CS <- function(n,C,Y){
y.matrix=matrix(Y,n,C)
mu.hat=mean(Y)
Z=Y-mu.hat
Z.matrix=matrix(Z,n,C)
J=matrix(1,n,n)
tmp1=sum(apply(Z.matrix^2,2,sum))
tmp2=apply(Z.matrix%*%matrix(apply(Z.matrix,2,sum),C,1),2,sum)
tmp3=1/((C*n)*(n-1))
sigma2.hat=tmp3*((n*sum(tmp1))-sum(tmp2))
d.hat=tmp3*(sum(tmp2)-sum(tmp1))
var.mu.hat=(sigma2.hat+(n*d.hat))/(C*n)
cov.varcomp=(2*(sigma2.hat^2)/((C*n)*(n-1)))*matrix(
c(n,-1,-1,(((sigma2.hat^2) + ((2*(n-1))*(d.hat*sigma2.hat)) +
((n*(n-1))*(d.hat^2)))/(sigma2.hat^2))),2,2)
return(list(mu.hat=mu.hat,d.hat=d.hat,sigma2.hat=sigma2.hat
,var.mu.hat=var.mu.hat,cov.varcomp=cov.varcomp))
```

```
}

est.CS.for <- function(n,C,Y){
y.matrix=matrix(Y,n,C)
mu.hat=mean(Y)
Z=Y-mu.hat
Z.matrix=matrix(Z,n,C)
tmp2=rep(0,C)
tmp1=rep(0,C)
J=matrix(1,n,n)
for (i in 1:C){
tmp1[i]=t(Z.matrix[,i])%*% Z.matrix[,i]
tmp2[i]= (t(Z.matrix[,i])%*%J)%*%Z.matrix[,i]
}
tmp3=1/((C*n)*(n-1))
sigma2.hat=tmp3*((n*sum(tmp1))-sum(tmp2))
d.hat=tmp3*(sum(tmp2)-sum(tmp1))
var.mu.hat=(sigma2.hat+(n*d.hat))/(C*n)
cov.varcomp=(2*(sigma2.hat^2)/((C*n)*(n-1)))*matrix(
c(n,-1,-1,(((sigma2.hat^2) + ((2*(n-1))*(d.hat*sigma2.hat)) +
((n*(n-1))*(d.hat^2)))/(sigma2.hat^2))),2,2)
return(list(mu.hat=mu.hat,d.hat=d.hat,sigma2.hat=sigma2.hat,
var.mu.hat=var.mu.hat,cov.varcomp=cov.varcomp))
}

est.CS.all <- function(data,ck,nk){

Y.mat=matrix(data[,1],max(nk),dim(data)[1]/max(nk))
split.idx.sub=matrix(data[,2],max(nk),dim(data)[1]/max(nk))[1,]
split.matrix=matrix(0,length(ck),sum(ck))
for (i in 1:length(ck)){
split.matrix[i,split.idx.sub==i]=1
}
subj.mean=apply(Y.mat,2,sum,na.rm=T)
split.mu.hat=(split.matrix%*%subj.mean)/(ck*nk)
subj.mu.hat=t(split.matrix)%*%split.mu.hat
mean.mat=matrix(rep(c(subj.mu.hat),max(nk)),max(nk),
```

```
                length(subj.mean),byrow=T)
Z.mat=(Y.mat-mean.mat)
tmp1.1=matrix(rep(apply(Z.mat^2,2,sum,na.rm=T),length(ck)),
              length(ck),sum(ck),byrow=T)
tmp1=apply(split.matrix*tmp1.1,1,sum)
tmp2.1=matrix(rep(apply(Z.mat,2,sum,na.rm=T),length(ck)),
              length(ck),sum(ck),byrow=T)
tmp2.2=split.matrix*tmp2.1
Z.mat.zero=Z.mat
Z.mat.zero[is.na(Z.mat)==TRUE]=0
tmp2=apply(Z.mat.zero%*%t(tmp2.2),2,sum,na.rm=T)
tmp3=1/((ck*nk)*(nk-1))
split.sigma2.hat=tmp3* ((nk*tmp1)-tmp2)
split.d.hat=tmp3*(tmp2-tmp1)
split.var.mu.hat=(split.sigma2.hat + (nk*split.d.hat))/(ck*nk)
varcomp.factor=(2*(split.sigma2.hat^2)/((ck*nk)*(nk-1)))
split.var.d.hat=varcomp.factor*(((split.sigma2.hat^2) +
 ((2*(nk-1))*(split.d.hat*split.sigma2.hat)) +
((nk*(nk-1))*(split.d.hat^2)))/(split.sigma2.hat^2))
split.var.sigma2.hat=varcomp.factor*nk
split.cov.d.sigma2.hat=-1*varcomp.factor
return(list(mu.hat=split.mu.hat,sigma2.hat=split.sigma2.hat,
            d.hat=split.d.hat,var.mu.hat=split.var.mu.hat,
            var.d.hat=split.var.d.hat,
var.sigma2.hat=split.var.sigma2.hat,
cov.d.sigma2.hat=split.cov.d.sigma2.hat))
}

param.free.CS <- function(nk,ck,mu.split.est,sigma2.split.est,
                          d.split.est,mu.split.var,sigma2.split.var,
                          d.split.var){
num.split=length(ck)
# Calculating parameter free weights
Equal=rep(1/num.split,num.split)
Prop=ck/sum(ck)
Appr.sc=(ck*nk)/sum(ck*nk)
# mu
```

```
mu.eq=c(sum(Equal*mu.split.est),sum((Equal^2)*mu.split.var))
mu.prop=c(sum(Prop*mu.split.est),sum((Prop^2)*mu.split.var))
mu.appr.sc=mu.prop
mu=rbind(mu.eq,mu.prop,mu.appr.sc)
colnames(mu)=c("Est","Var")
rownames(mu)=c("Equal","Prop","Appr.sc.")
# Sigma2
sigma2.eq=c(sum(Equal*sigma2.split.est),
            sum((Equal^2)*sigma2.split.var))
sigma2.prop=c(sum(Prop*sigma2.split.est),
              sum((Prop^2)*sigma2.split.var))
sigma2.appr.sc=c(sum(Appr.sc*sigma2.split.est),
                 sum((Appr.sc^2)*sigma2.split.var))
sigma2=rbind(sigma2.eq,sigma2.prop,sigma2.appr.sc)
colnames(sigma2)=c("Est","Var")
rownames(sigma2)=c("Equal","Prop","Appr.sc.")
# d
d.eq=c(sum(Equal*d.split.est),sum((Equal^2)*d.split.var))
d.prop=c(sum(Prop*d.split.est),sum((Prop^2)*d.split.var))
d.appr.sc=d.prop
d=rbind(d.eq,d.prop,d.appr.sc)
colnames(d)=c("Est","Var")
rownames(d)=c("Equal","Prop","Appr.sc.")

return(list(mu=mu,sigma2=sigma2,d=d))
}

scalar.weights.CS
   <- function(nk,ck,mu.split.est,sigma2.split.est,d.split.est,
               mu.split.var,sigma2.split.var,d.split.var){

ak=(ck*nk)/(sigma2.split.est + (nk*d.split.est))
w.mu=ak/sum(ak)
bk=ck*(nk-1)
w.sigma2=bk/sum(bk)
gk=(ck*nk)/
(((sigma2.split.est^2)/(nk-1))
```

```
+((2*sigma2.split.est)*d.split.est)+(nk*(d.split.est^2)))
w.d=gk/sum(gk)

mu.scalar=c(sum(w.mu*mu.split.est), sum(mu.split.var* (w.mu^2)))
d.scalar=c(sum(w.d*d.split.est), sum(d.split.var* (w.d^2)))
sigma2.scalar=c(sum(w.sigma2*sigma2.split.est),
                sum(sigma2.split.var* (w.sigma2^2)))
param.scalar=rbind(mu.scalar,sigma2.scalar,d.scalar)
colnames(param.scalar)=c("Est.","Var")
rownames(param.scalar)=c("mu","sigma2","d")
return(param.scalar)
}

approx.optimal.CS
   <- function(nk,ck,mu.split.est,sigma2.split.est,d.split.est,
               sigma2.split.var,d.split.var){
library(magic)
num.split=length(ck)
#Calculating approximated optimal weights
W=NULL
for (i in 1:num.split){
V1=(2*(sigma2.split.est[i]^2))/(ck[i]*(nk[i]-1))
V2=(-1)*((2*(sigma2.split.est[i]^2))/(ck[i]*nk[i]*(nk[i]-1)))
V3=(2/(ck[i]*nk[i]))*((sigma2.split.est[i]^2
            /(nk[i]-1))+((2*d.split.est[i])
             *sigma2.split.est[i])+(nk[i]*(d.split.est[i]^2)))
W[[i]]=solve(matrix(c(V1,V2,V2,V3),2,2))
}

V.total=apply(simplify2array(W),c(1,2),sum)
W.inv=solve(V.total)
W.opt=NULL
sigma2.d=rbind(sigma2.split.est,d.split.est)
sigma2.d.est=matrix(0,dim(sigma2.d)[1],dim(sigma2.d)[2])
for (i in 1:num.split){
W.opt=W.inv %*% W[[i]]
sigma2.d.est[,i]=W.opt%*% sigma2.d[,i]
```

```
}
varcomp.est=apply(sigma2.d.est,1,sum)
varcomp.var=W.inv

# Calculating proper Variance

A1=((sigma2.split.est^2) + (nk*d.split.est))^3
A2=(sigma2.split.est*d.split.est)*((2*sigma2.split.est)
    *(nk*d.split.est))
A3=d.split.est*((sigma2.split.est+(nk*d.split.est))^2)
A=A1-A2-A3

dev.w.sigma2=NULL
dev.w.d=NULL
C3.tmp=NULL
C4.tmp=NULL
for (i in 1:num.split){
tmp1=(-1*ck[i]*nk[i])/A1[i]
tmp2=matrix(c(A[i] ,1 ,1, nk[i]),2,2)
dev.w.sigma2[[i]]=tmp1*tmp2
dev.w.d[[i]]=tmp1 * matrix(c(1,nk[i],nk[i],(nk[i]^2)),2,2)

II=as.matrix(diag(2)-(W.inv%*%W[[i]]))
Theta=c(sigma2.split.est[i],d.split.est[i])
Rho=II%*%Theta
C3.tmp[[i]]=adiag(dev.w.d[[i]],dev.w.sigma2[[i]])
C4.tmp[[i]]= kronecker (diag(2),Rho)
}
C1.tmp=t(kronecker(c(1,1),diag(2)))
C2.tmp=kronecker(diag(2),W.inv)
proper.var=NULL
for (i in 1:num.split){
C=((C1.tmp%*%C2.tmp)%*%C3.tmp[[i]])%*%C4.tmp[[i]]
AA=W.inv%*%(W[[i]])
B=AA+C
VV=diag(c(d.split.var[i],sigma2.split.var[i]))
proper.var[[i]]=(B%*%VV)%*%t(B)
```

```
}
Proper.var=apply(simplify2array(proper.var),c(1,2),sum)

# Approximate weights for mu
Am=(ck*nk)/(sigma2.split.est+(nk*d.split.est))

w.mu=Am/sum(Am)
mu.est=sum(w.mu*mu.split.est)
mu.var=1/sum(Am)
# Proper variance for mu
proper.var.mu1=mu.var
TMP2=rep(0,num.split)
TMP.1=rep(0,num.split)
for (k in 1:num.split){
TMP1=2*ck[k]*(nk[k]^2)
for (m in 1:num.split){
TMP2[m]=Am[m]*((mu.split.est[k]-mu.split.est[m])^2)
}
TMP.1[k]=TMP1*sum(TMP2)
}
TMP.2=sum(Am)^4
proper.var.mu=sum(TMP.1)/TMP.2
Proper.var.mu=(1/sum(Am))+ proper.var.mu

return(list(mu.est=mu.est,varcomp.est=varcomp.est,mu.var=mu.var,
 varcomp.var=varcomp.var,proper.var.mu=Proper.var.mu,
 proper.var.varcomp=Proper.var))
}

clusterBYcluster.CS <- function(nk,ck,Data){
# Data should be a 3-column matrix with first column the subject,
# second column the response and third column the split indexes
# which show which observation belongs to which sub-sample.
num.split=length(ck)
Var.varcomp=NULL
mu.split.est=rep(0,num.split)
mu.split.var=matrix(0,num.split,3)
```

```
d.split.est=matrix(0,num.split,3)
sigma2.split.est=matrix(0,num.split,3)
d.split.var=matrix(0,num.split,3)
sigma2.split.var=matrix(0,num.split,3)

for (k in 1:num.split){

# Making data for each cluster
split.data=Data[Data[,3]==k,]
n=nk[k]
N=ck[k]
# Computing t^2
data.matrix=matrix(split.data[,2],n,N)
mu.hat=sum(apply(data.matrix,2,sum))/(prod(dim(data.matrix)))
t2=sum((apply(data.matrix,2,mean)-mu.hat)^2)/(dim(data.matrix)[2])
#Computing S^2
mean.vec=apply(data.matrix,2,mean)
SS=rep(0,dim(data.matrix)[2])
for (i in 1:dim(data.matrix)[2]){
SS[i]=sum((data.matrix[,i]-mean.vec[i])^2)
}
s2=sum(SS)/prod(dim(data.matrix))
# Computing s^2* and t^2*
s2.star=(dim(data.matrix)[1]/(dim(data.matrix)[1]-1))*s2
t2.star=(dim(data.matrix)[2]/(dim(data.matrix)[2]-1))*t2
# Computing \widehat{\sigma}^2 and \widehat{d}
Z=data.matrix-mu.hat
J=matrix(1,dim(Z)[1],dim(Z)[1])
ZZ=rep(0,dim(Z)[2])
ZJZ=rep(0,dim(Z)[2])
for (i in 1:dim(Z)[2]){
ZZ[i]=t(Z[,i])%*% Z[,i]
ZJZ[i]=(t(Z[,i])%*%J)%*%Z[,i]
}
d.hat=(1/(N*n*(n-1))) * (sum(ZJZ)-sum(ZZ))
sigma2.hat=(1/(N*n*(n-1)))* ((n*sum(ZZ))-sum(ZJZ))
```

```
# computing the variance of \widehat{\mu} \widehat{\sigma}^2
  and \widehat{d}

var.mu=(sigma2.hat+ (n*d.hat))/(N*n)
var1=(2*(sigma2.hat^2))/(N*(n-1))
var12=(-1)*(2*(sigma2.hat^2))/(N*n*(n-1))
var2=(2/(n*N))*(((sigma2.hat^2)/(n-1)) + (2*sigma2.hat*d.hat)
              + (n*(d.hat^2)))
var.varcomp=matrix(c(var1,var12,var12,var2),2,2)

# computing the variance of \widehat{\mu} \widehat{\sigma}^2
  and \widehat{d}
#based on the two stage approach

var.mu.2stage=(s2+ (n*t2))/(N*n)
var.s2=(2*(n-1)*(s2^2))/(N*(n^2))
var.t2=(2*((N-1)^2))/((N^2)*n) * (((s2^2)/n) + (2*s2+t2 )
        + (n*(t2^2) ))

# computing the variance of \widehat{\mu} \widehat{\sigma}^2
  and \widehat{d}
#based on the unbiased two stage approach

var.mu.2stage.unbiased=(s2.star+ (n*t2.star))/(N*n)
var.s2.star=(2*(s2.star^2))/ (N*(n-1))
var.t2.star= (2/(N*n)) * (((s2.star^2)/n)+(2*s2.star*t2.star)
                      +(n*(t2.star^2)))

# Saving the results

mu.split.est[k]=mu.hat
Var.varcomp[[k]]=var.varcomp
mu.split.var[k,]=c(var.mu,var.mu.2stage,var.mu.2stage.unbiased)
d.split.est[k,]=c(d.hat,t2,t2.star)
sigma2.split.est[k,]=c(sigma2.hat,s2,s2.star)
d.split.var[k,]=c(var.varcomp[2,2],var.t2,var.t2.star)
sigma2.split.var[k,]=c(var.varcomp[1,1],var.s2,var.s2.star)
```

```
}
colnames(mu.split.var)=c("Weighted","TwoStage",
         "TwoStageUnbiased")
colnames(d.split.var)=c("Weighted","TwoStage",
         "TwoStageUnbiased")
colnames(sigma2.split.var)=c("Weighted","TwoStage",
         "TwoStageUnbiased")
colnames(d.split.est)=c("Weighted","TwoStage",
         "TwoStageUnbiased")
colnames(sigma2.split.est)=c("Weighted","TwoStage","TwoStageUnbiased")

return(list(mu.split.est=mu.split.est,mu.split.var=mu.split.var,
sigma2.split.est=sigma2.split.est, sigma2.split.var=sigma2.split.var,
d.split.est=d.split.est,d.split.var=d.split.var,
            Var.varcomp=Var.varcomp))}
```

Appendix C

Appendix for Chapter 6

Section C.1 gives the contrast between the AR(1) model and the balanced conditionally independent model. Section C.2 outlined detailed derivations expressions presented in Sections 6.2 and 6.3. Details on the simulation study and data analysis are presented respectively in Section C.3 and Section C.4.

C.1 The Balanced Conditionally Independent Model

In this case, one imposes the following structure on (6.1):

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

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

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

Here, $\boldsymbol{Y}$ is an $N \times n$ matrix stacking the outcomes of all clusters of size c, A, Z, and Q group the designs mentioned in Section 6.2, the vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ contain the fixed

effects, B contains N rows of length q, representing the q-dimensional random-effects vector, and ε shares its dimensions with $\boldsymbol{Y}$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

C.2 Algebraic Derivations in the AR(1) Case

Here, we present more detailed derivations of the key algebraic expressions presented in Sections 6.2 and 6.3.

C.2.1 Some Useful Expressions

Consider,

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

then,

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

It can be shown that:

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

The inverse of C can be calculated as follows:

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

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

Consider:

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

Then, by taking the derivative with respect to ρ:

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

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

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

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

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

C.2.2 The Likelihood Estimators in a Given Cluster

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Therefore, g and h can be computed as:

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

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

Now using (C.27):

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

which gives:

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

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

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

which gives:

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

For $n_k = 3$:

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

Let,

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

Using (C.27):

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

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

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

with:

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

Then using (C.26):

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

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

Proof. Consider:

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

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

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

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

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

C.2.3 Hessians, Covariance Matrices, and Optimal Weights

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

C.2.3.1 Second derivative with respect to fixed effects

As

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

we have:

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

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

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

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

C.2.3.2 Second derivative with respect to variance components

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

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

it follows that

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

The second-derivatives are:

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

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

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

This follows from the fact that:

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

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

For the second derivative with respect to ρ:

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

Likewise:

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

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

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

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

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

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

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

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

So,

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

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

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

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

Therefore,

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

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

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

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

C.2.4 Proof of Proposition 6.1

Proof. Consider an estimator of the form:

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

for a vector of weights $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)'$. Because the clusters are i.i.d. it is evident that the components of $\boldsymbol{\alpha}$ do not depend on the cluster index i. Clearly, the requirement that $E(\widetilde{\mu}_\alpha) = \mu$ implies the condition

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

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

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

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

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

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

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

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

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

establishing the MLE. This completes the proof.

C.2.5 Optimal weights in case of a general mean structure $X_i^{(k)}\boldsymbol{\beta}$

Cluster size specific expressions are:

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

and

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

The combination ruleis

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

with

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

and C_k is as described in Supplementary Materials C.2.1.

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

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

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

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

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

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

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

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

For the third term, define $\boldsymbol{X}_{ij}^{(k)-} = (x_{ki2j} \ \dots \ x_{kin_kj})^t$ and $\boldsymbol{X}_{ij}^{(k)+} = (x_{ki1j} \ \dots \ x_{kin_{k-1}j})^t$. As a consequence, (C.59) will equal

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

In summary:

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

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

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

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

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

C.2.6 Delta Method for the Mean Estimator

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

$$a_k = \frac{a_k'}{\sum_{m=1}^{K} a_m'}, \tag{C.66}$$

and (6.21) becomes

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

The first derivatives equal

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

and these can be combined using the delta method, resulting in (6.24):

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

C.2.7 Calculating $\widehat{\rho}$ and $\widehat{\sigma}^2$ in R

In this section, we consider the implementation of the calculations for the variance components in R. This can be done with a few simple lines of code. For fixed $C_k = C$ and $n_k = n$, and given the data y, the function `est.ar1` estimates the variance components and provides a plot for the third-degree polynomial in (C.31). This visually underscores that there is only one root in $[-1, 1]$. Figure C.1 shows (C.31) for 10 simulated data sets; clearly, there is a single root only in $[-1, 1]$. For convenience, the R code is given in Supplementary Materials C.5. Other functions to find variances and iterated optimal weights are also available.

C.3 Details on Additional Simulations

C.3.1 Simulations with Proportional and Size-proportional Weights

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

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

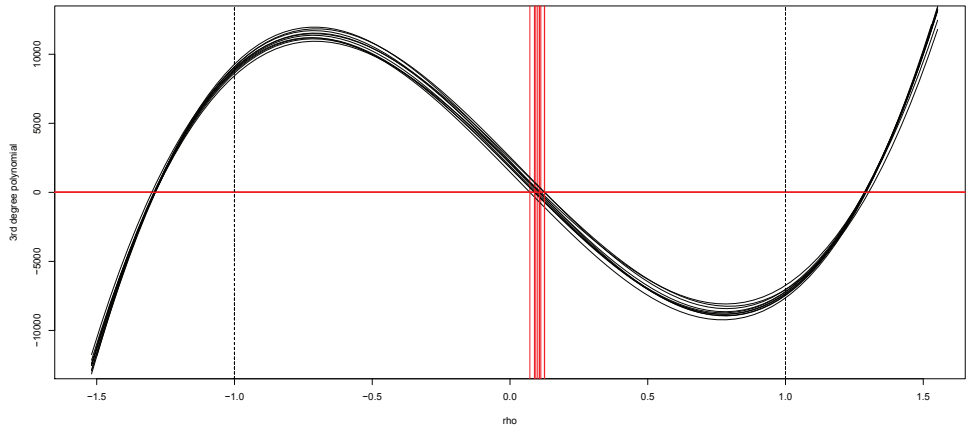

Figure C.1: *Calculations. The third degree polynomial in (C.31) for* 10 *different generated data. The red vertical line shows* $\widehat{\rho}$.

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

C.3.2 Simulations with Proportional and Size-proportional Weights: ρ near 0/1

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

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

In Figure C.5, for $\rho = 0.99$ and 0.95, the size-proportional weights perform worse than the proportional weights. This is expected, because in this case AR(1) approaches CS. This result is clearer in the left panel of Figure C.5, where the standard deviations are

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

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

shown. For ρ's near 1, the proportional weights are as efficient as full likelihood, while as ρ moves further from 1 this would happen for size-proportional weights.

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

C.3.3 Simulations With Optimal Weights

Given the covariance matrix of the parameter estimators, finding the optimal weights is straightforward, but in practice the unknown parameters therein need to be estimated. Here we compare the iterative weights with size-proportional weights and ML. See Figures C.7, C.8, C.9, and Table C.1. As expected, the optimal weights lead to estimates very close to the MLE; the difference between them being numerical.

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

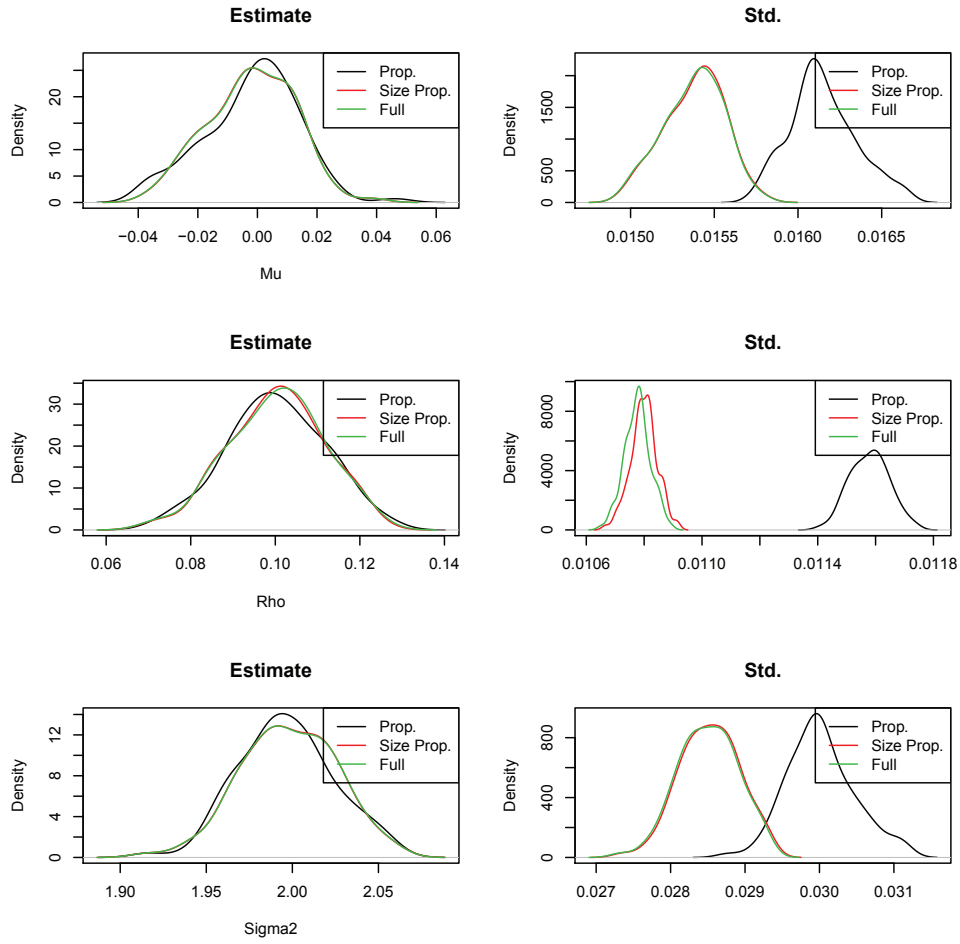

Figure C.2: *Simulation study. Comparing proportional and size-proportional weights with full likelihood for* 100 *replications with* $\mu = 0$, $\sigma^2 = 2$ *and* $\rho = 0.1$.

C.3.4 Simulations on Computation Time

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

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

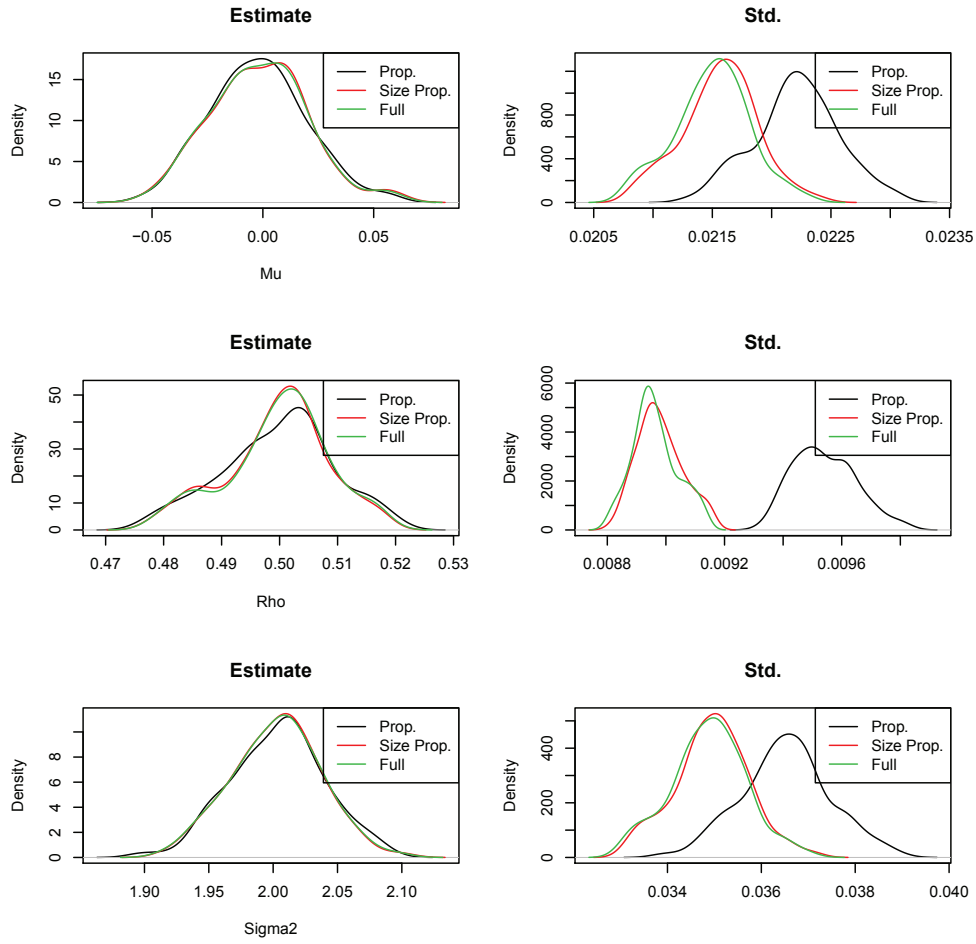

Figure C.3: *Simulation study. Comparing proportional and size-proportional weights with full likelihood for* 100 *replications with* $\mu = 0$, $\sigma^2 = 2$ *and* $\rho = 0.5$.

then the MSE is computed as follows:

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

Table C.2 summarizes the results for μ. The sample splitting estimates are computed using proportional and size-proportional (identical to equal weights in this case) weights. The results using the full sample are also given. The third column in Table C.2 presents the averaged (over 100 replications) estimated μ and its standard deviation. The fourth column presents the averaged estimated standard deviation for $\widehat{\mu}$ (over 100 replications)

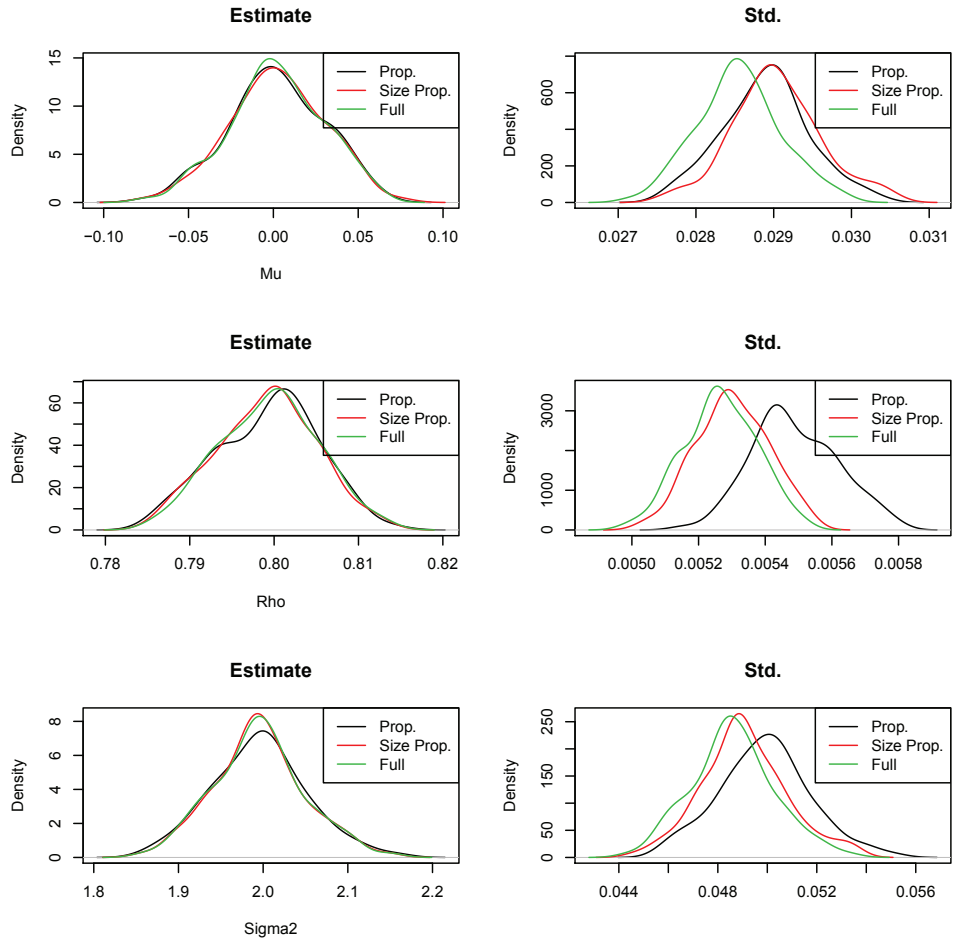

Figure C.4: *Simulation study. Comparing proportional and size-proportional weights with full likelihood for* 100 *replications with* $\mu = 0$, $\sigma^2 = 2$ *and* $\rho = 0.8$.

and its standard deviation. The last column shows the MSE computed using (C.70) for $\mu_0 = 1$. Tables C.3 and C.4 shows the same results for ρ and σ^2 ($\sigma_0^2 = 2$), respectively.

Table C.5 compares the computation time between closed-form and interative methods. The closed-form solutions are implemented in R and for the numerical methods the MIXED procedure in SAS is used, with error covariance structure set to AR(1).

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

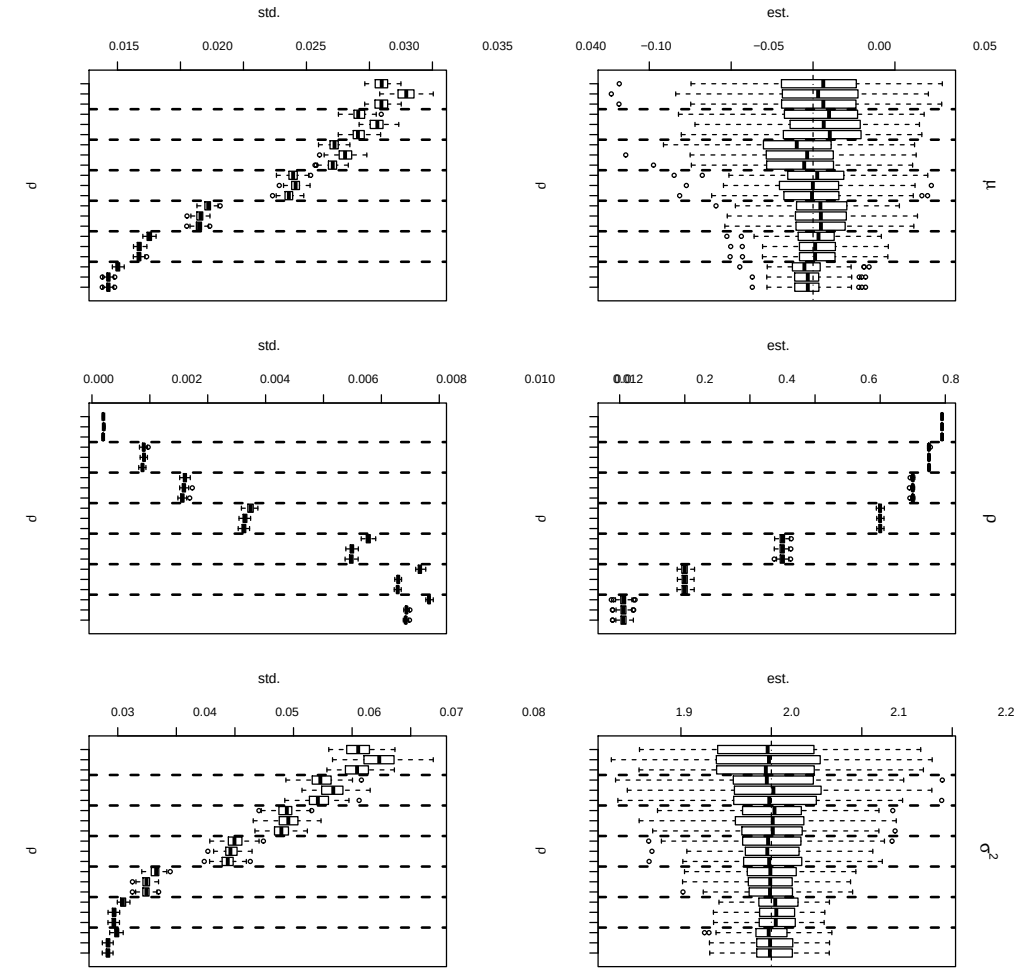

Figure C.5: *Simulation study. Boxplots comparing proportional and size-proportional weights with full likelihood for* 100 *replications with* $\mu = 0$, $\sigma^2 = 2$ *and* $\rho = 0.99, 0.95, 0.9, 0.8, 0.5, 0.2, 0.01$. *In every section of the boxplots (which are separated by dashed lines) the first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

As one may see in Table C.5 and Figure C.10, using closed form solutions significantly reduces the computation time. This means, as well as the computation time reduction due to splitting the data, using closed form solutions within each split the computation time

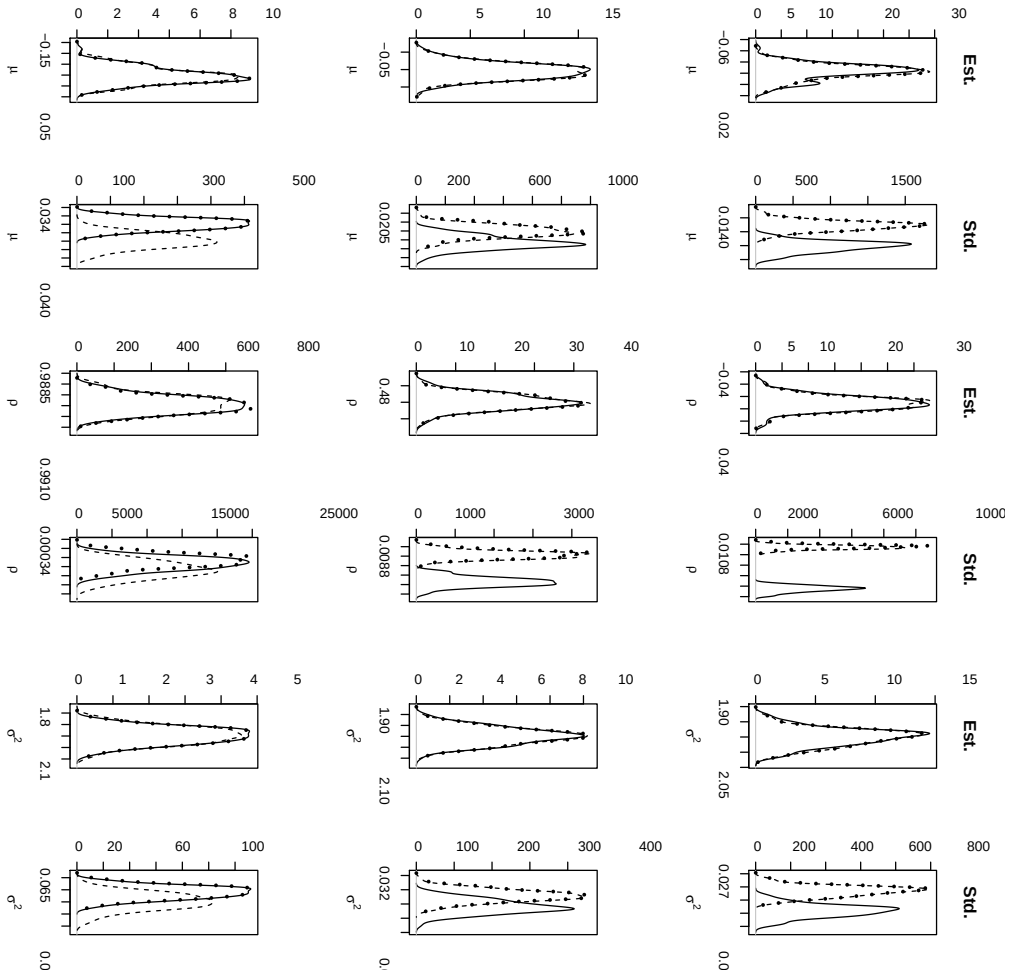

Figure C.6: *Simulation study. Comparing proportional, size-proportional and full likelihood results via their empirical density for the 100 replications. In all of the figures* $\mu = 0$ *and* $\sigma^2 = 2$*. The first row is for* $\rho = 0.01$*, the middle one is for* $\rho = 0.5$ *and last one corresponds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

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

To assess the effect of the overall size of the dataset, the model is fitted to two

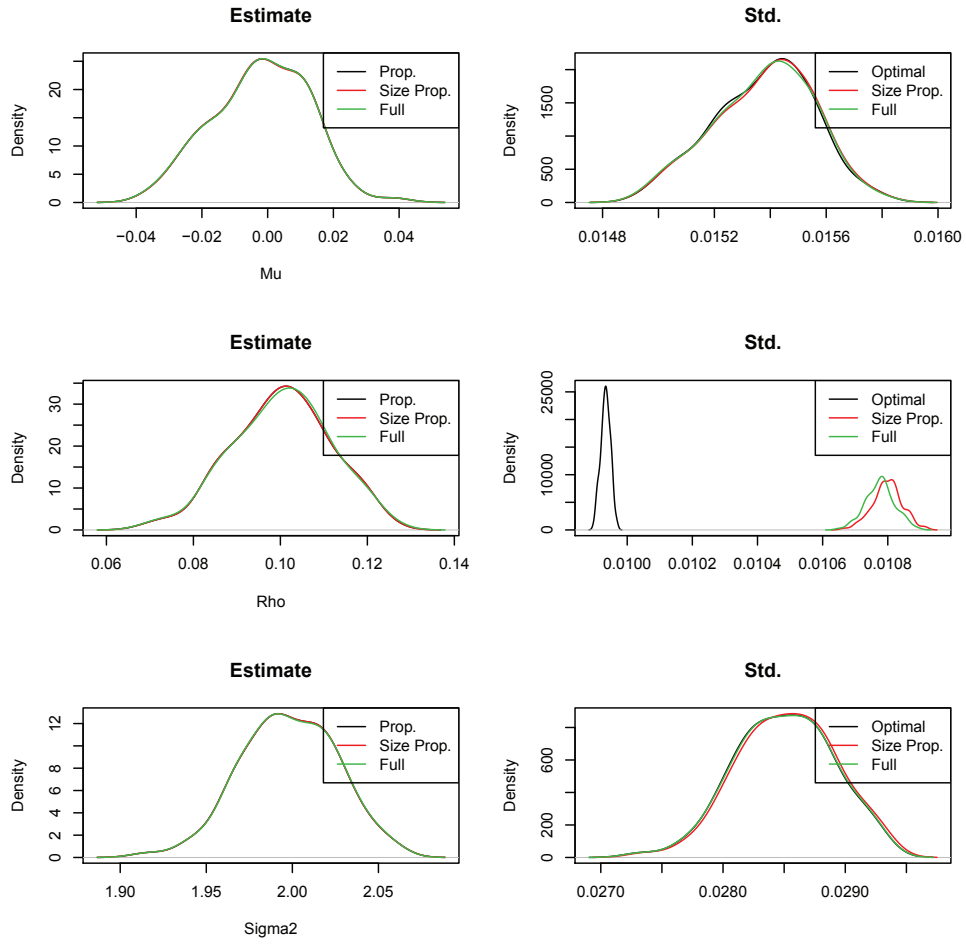

Figure C.7: *Simulation study. Comparing iterated optimal and size-proportioanl weights with full likelihood for* 100 *replications with* $\mu = 0$, $\sigma^2 = 2$ *and* $\rho = 0.1$.

concatenated copies of the same set. Computation time results are presented in Table C.5 and Figure C.10. The data are generated with $\mu = 0$, $\sigma^2 = 2$, and $\rho = 0.25$.

C.4 Details on PANSS Data Analysis

As one may see from Table C.6, by far the majority of the study subjects have complete data and hence belong to the first pattern.

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

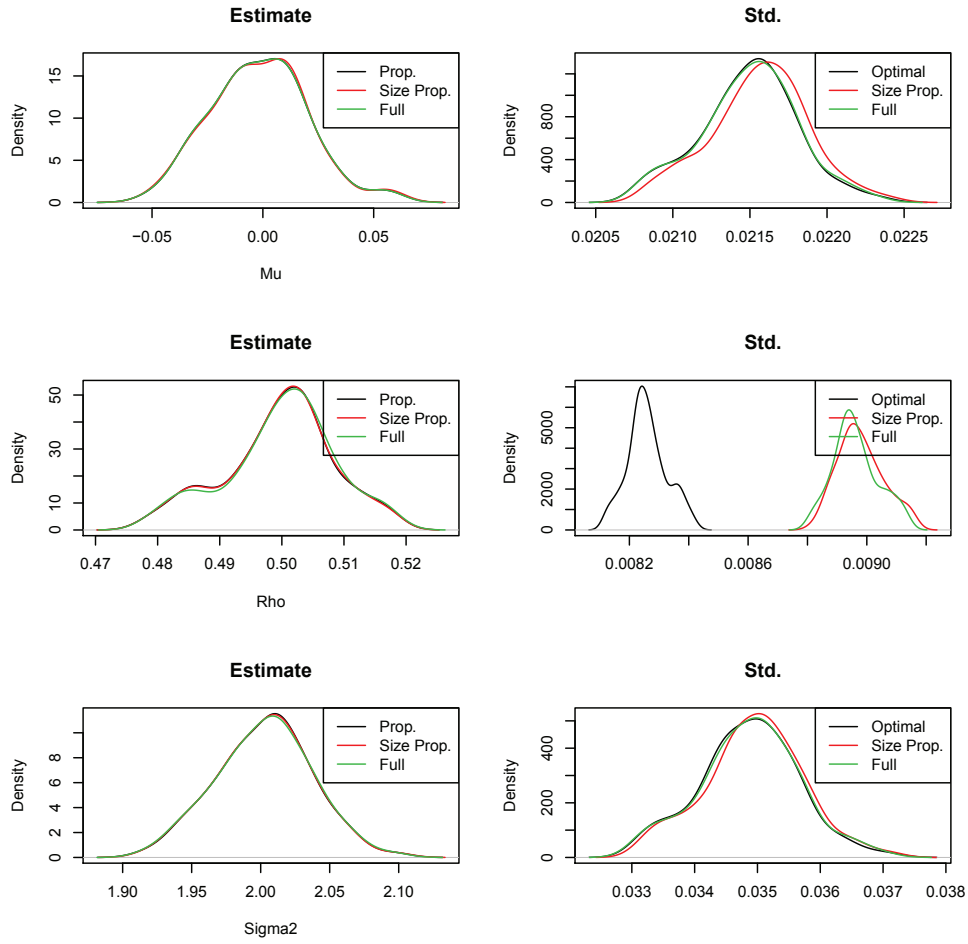

Figure C.8: *Simulation study. Comparing iterated optimal and size-proportioanl weights with full likelihood for* 100 *replications with* $\mu = 0$, $\sigma^2 = 2$ *and* $\rho = 0.5$.

To examine the choice of an AR(1) covariance structure, Table C.7 shows three model selection criteria to compare different error covariance structures. Changing from independence structure ($R = \sigma^2 I$) to compound-symmetry ($R = \sigma^2 I$) the criteria decrease with a large amount, and the same when changing to AR(1). The step to an unstructured covariance does not make a big difference (considering that the unstructured covariance would has 21 parameters to estimate compared to 2 parameters in the AR(1) model). Therefore, AR(1) seems to be a good choice.

The 95% confidence intervals, accompanying (6.26), are presented in Figure C.12. In

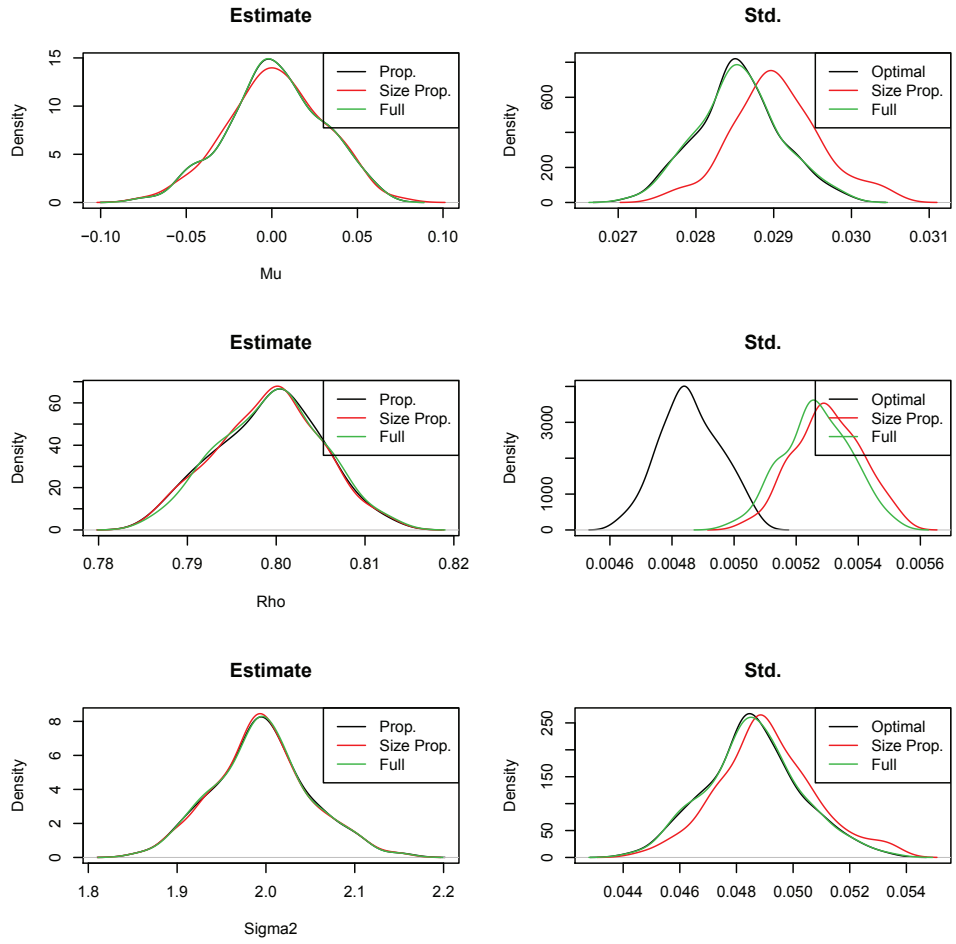

Figure C.9: *Simulation study. Comparing iterated optimal and size-proportioanl weights with full likelihood for* 100 *replications with* $\mu = 0$, $\sigma^2 = 2$ *and* $\rho = 0.8$.

order to give more insight in these results, Figure C.13 shows the 95% confidence interval in each split, comparing with the full sample splits (the horizontal dashed line in the figure).

The 95% confidence intervals, accompanying (6.27), are presented in Figure C.14. Figure C.15 shows the 95% confidence intervals for the parameter estimates in each split comparing with the full sample estimate (the horizontal dashed-like in the figure.)

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

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

C.5 R Code

Estimating variance components

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

# making a matrix out of the response vector
Resp=matrix(Y,n,C)

# Computing cross products
SS=crossprod(t(Resp))

# Computing S, \tilde{S} and R
```

Table C.3: *Simulation study. Estimating ρ and its standard deviation. The mean (standard deviation) of the 100 replications are given together with mean squared errors for $\rho = 0.01, 0.2, 0.5, 0.8, 0.9, 0.99$ using proportional and size-proportional weights comparing with the full likelihood results.*

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

```
S=sum(diag(SS))
S.tilde=sum(diag(SS)[2:(n-1)])
tmp.R=SS
diag(tmp.R)=NA
tmp.R2 = (matrix(tmp.R[which(!is.na(tmp.R))],nrow=n,ncol=n-1))
R=sum(tmp.R2[1,])

# Finding the coefficients of the 3rd degree polynomial and its roots
P1=(n-1)*S.tilde
P2=(n-2)*R
P3=((n*S.tilde)+S)
```

Table C.4: *Simulation study. Estimating σ^2 and its standard deviation. The mean (standard deviation) of the 100 replications are given together with mean squared errors for $\rho = 0.01, 0.2, 0.5, 0.8, 0.9, 0.99$ using proportional and size-proportional weights comparing with the full likelihood results.*

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

```
P4=n*R

PP=polynomial(c(P4,-P3,-P2,P1))
roots=polyroot(PP)
Roots=Re(roots)[abs(Im(roots)) < 1e-6]
rho.hat=Roots[abs(Roots)<1]

# Plotting the 3rd degree polynomial if requested

if (Plot==1){
plot(PP,xlim=c(-1.5,1.5),xlab="rho",ylab="3rd degree polynomial")
```

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

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

```
abline(h=0,col=2)
abline(v=Roots[abs(Roots)<1],lty=2,lwd=2,col=2)
abline(v=-1,lty=2)
abline(v=1,lty=2)
}

# Estimating \sigma2

tmp1=1/(C*n)
tmp2=1/(1-(rho.hat^2))
tmp3=S+((rho.hat^2)*S.tilde)
tmp4=2*rho.hat*R

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

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

Computing variance of parameter estimates

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

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

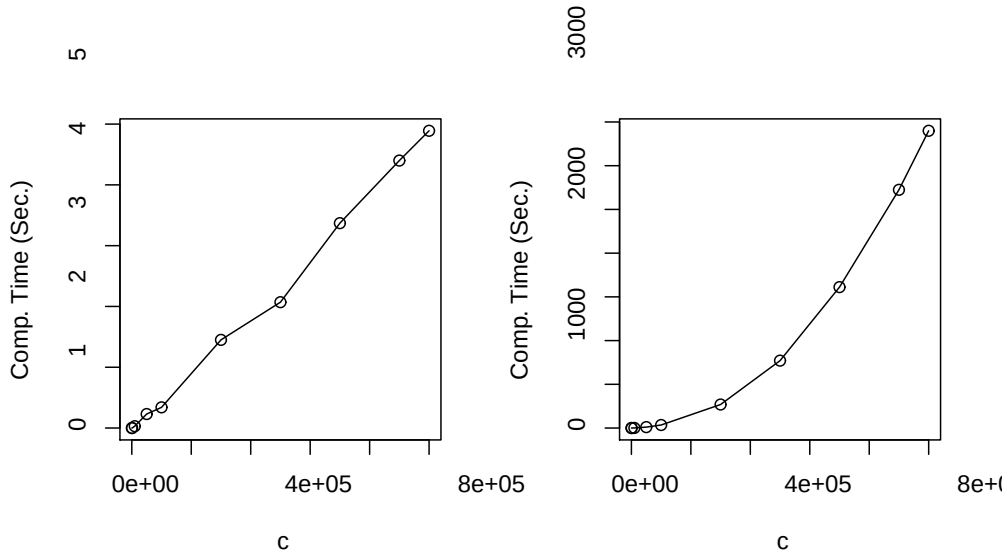

Figure C.10: *Simulation study. Comparing computation time using closed form (left) and numerical (right) solutions. The horizontal axis shows number of clusters (c) and the vertical axis shows the computation time in seconds.*

```
v22=2*(sigma2^2)*(1+(rho^2))
v12=2*sigma2*(1-(rho^2))
v11=(1-rho^2)^2
var.varcomp1=matrix(c(v11,v12,v12,v22),2,2)
varcomp.coef=1/(ck*(nk-((nk-2)*(rho^2))))
var.varcomp=outer(var.varcomp1,varcomp.coef)

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

return(list(var.mu=var.mu,var.varcomp=var.varcomp
```

Table C.6: *PANSS data. Number of clusters in each trial for each cluster pattern.*

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

```
,w.mu=w.mu,w.varcomp=w.varcomp))
}
```

Computing iterated optimal weights

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

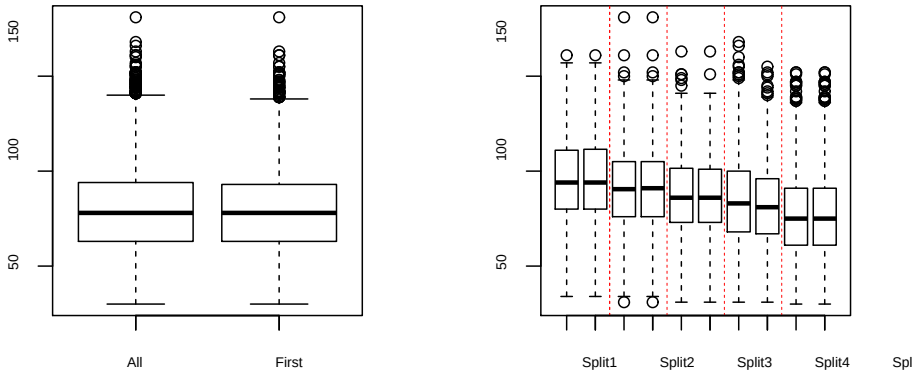

Figure C.11: *PANSS data. Boxplots for the entire set of data, for the subject from the first pattern only, and for various split samples.*

Table C.7: *PANSS data. Comparing different error covariance structures using three model comparison criteria for model (6.26) (residual log-likelihood value; AIC; BIC). Three R structures: Ind. : independence structure ($R = \sigma^2 I$), CS: compound-symmetry structure ($R = \sigma^2 I + dJ$), AR(1): AR(1) structure ($R_{ij} = \sigma^2 \rho^{|i-j|}$), UN: unstructured ($R_{ij} = \sigma^2_{ij}$).*

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

```
num.split=length(ck)
W.size.prop=(nk*ck)/sum(nk*ck)
rho.hat= sum(var.comp.split[1,]*W.size.prop)
sigma2.hat=sum(var.comp.split[2,]*W.size.prop)

diff=10
var.comp.hat=matrix(c(rho.hat,sigma2.hat),2,1)
count=0
while (diff>tol){
WW=cov.ar1 (ck,nk,var.comp.hat[1,1],var.comp.hat[2,1])
W.mu=WW$w.mu
```

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

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

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

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

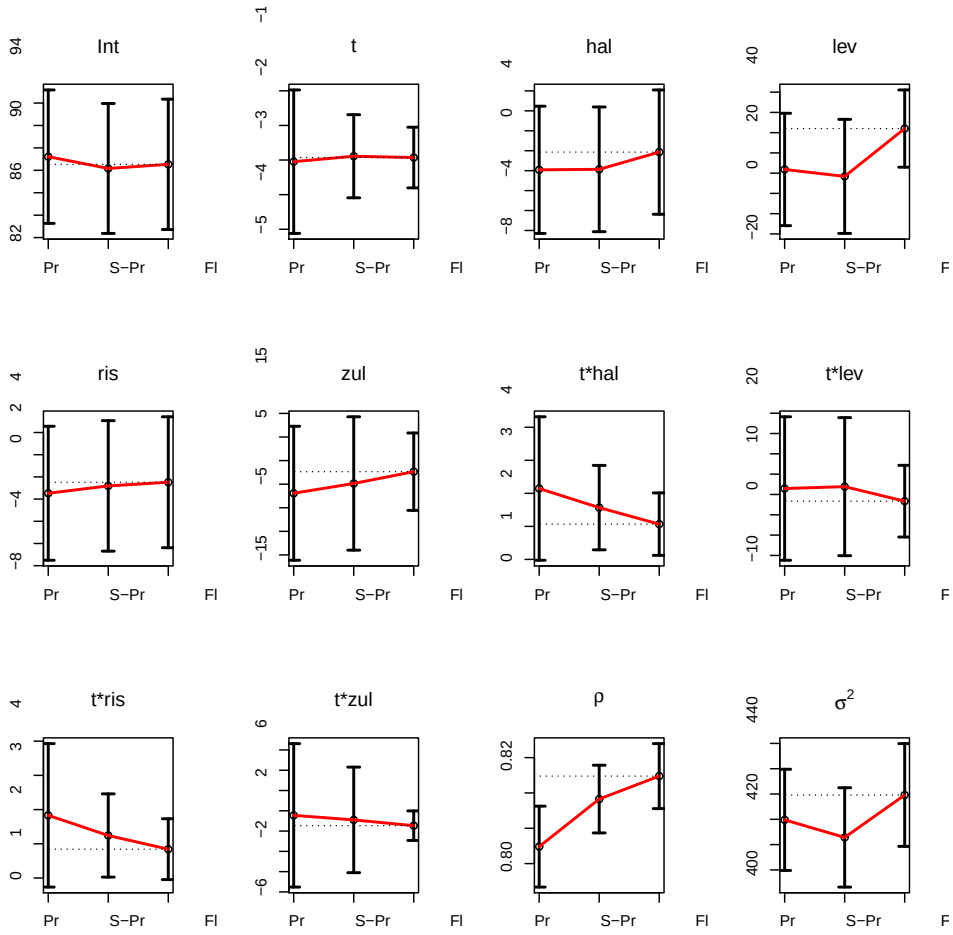

Figure C.12: *PANSS data. 95% confidence intervals for fixed effects and variance components estimates and the standard deviations of these estimates using sample splitting, combined with proportional (Pr - first) and size-proportional (S-Pr - second) weights, and full likelihood (Fl - third). The dashed horizontal line shows the full likelihood estimate. The model used in here is without trial effect (6.26).*

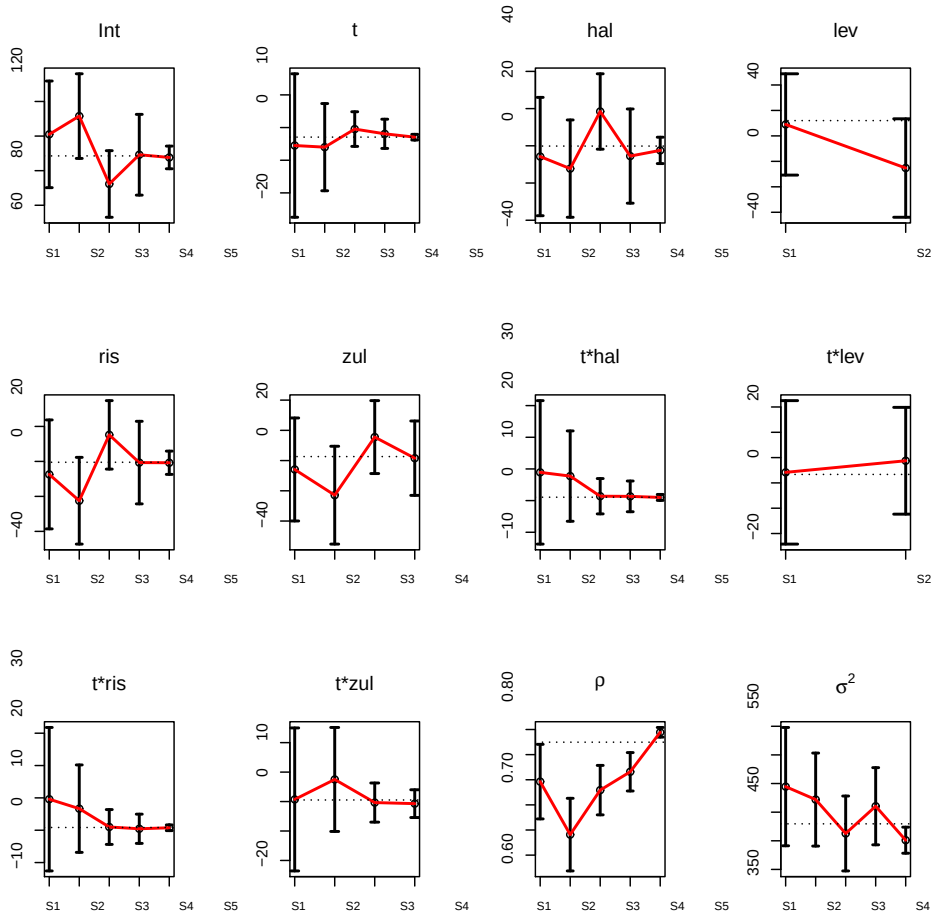

Figure C.13: *PANSS data.* 95% *confidence intervals for fixed effects and variance components estimates and the standard deviations of these estimates within each split. The dashed horizontal line shows the full likelihood estimate. The model used in here is without trial effect (6.27).*

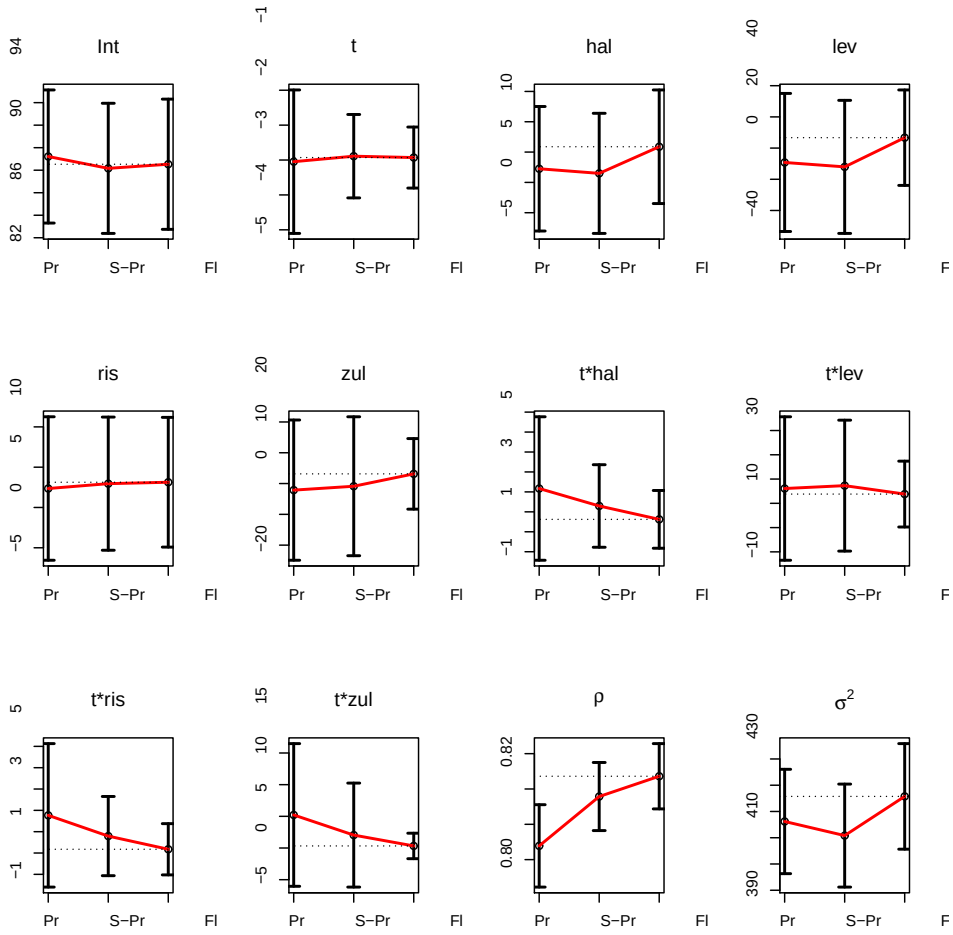

Figure C.14: *PANSS data.* 95% *confidence intervals for fixed effects and variance components estimates and the standard deviations of these estimates using sample splitting, combined with proportional (Pr - first) and size-proportional (S-Pr - second) weights, and full likelihood (Fl - third). The dashed horizontal line shows the full likelihood estimate. The model used in here is with trial effect (6.27).*

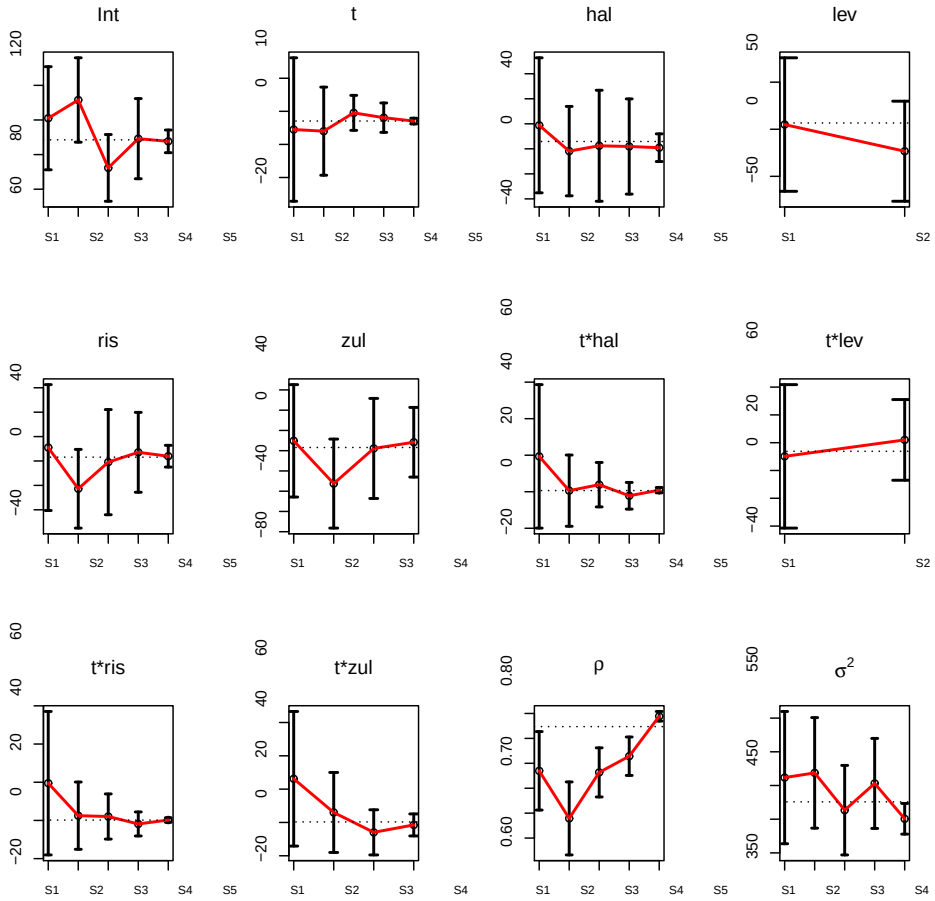

Figure C.15: *PANSS data.* 95% *confidence intervals for fixed effects and variance components estimates and the standard deviations of these estimates within each split. The dashed horizontal line shows the full likelihood estimate. The model used in here is with trial effect (6.27).*

Appendix D

Appendix for Chapter 7

D.1 Calculations of the Optimal Weights

Obtaining optimal weights is standard, but nevertheless useful to briefly review here for our purposes. We seek to minimize the variance, $\text{var}(\tilde{\psi}) = \sum_{i=1}^{N} \alpha_i^2 \text{var}(\psi_i)$. Write $v_i = \text{var}(\psi_i)$ so $\text{var}(\tilde{\psi}) = \sum_{i=1}^{N} \alpha_i^2 v_i$, and define the objective function:

$$Q = \sum_{i=1}^{N} \alpha_i^2 v_i - \lambda \left(\sum_{i=1}^{N} \alpha_i - 1 \right), \tag{D.1}$$

with λ a Lagrange multiplier.

To properly calculate the weights, the first derivative of Q with respect to weights α_i are taken and equated to zero:

$$\begin{aligned} \frac{\partial Q}{\partial \alpha_i} &= 2\alpha_i v_i - \lambda = 0, \\ 2\alpha_i v_i = \lambda \quad &\Rightarrow \quad \alpha_i = \lambda \frac{1}{2v_i}. \end{aligned} \tag{D.2}$$

By summing both sides, the left hand side is equal to 1 and an expression for λ is obtained:

$$1 = \sum_i \alpha_i = \frac{\lambda}{2} \sum_{i=1}^{N} \frac{1}{v_i} \Rightarrow \lambda = \frac{2}{\sum_{i=1}^{N} \frac{1}{v_i}}. \tag{D.3}$$

Plugging (D.3) into (D.2), the weights are:

$$\alpha_i = \frac{v_i^{-1}}{\sum_j v_j^{-1}}. \tag{D.4}$$

Appendix E

Appendix for Chapter 8

Section E.1 presents the estimating equations under exchangeability. Detailed calculations for Section 8.2 can be found in Section E.2. Section E.3 gives all details about the implementation in SAS.

E.1 Pairwise Estimating Equations Under Exchangeability

Molenberghs et al. (2011) proved that under exchangeability, meaning that the distribution of any sub-vector of $\boldsymbol{Y}_i$ is that of any other sub-vector of equal length or a permutation thereof, the estimating equations simplify considerably.

In the general formula

$$\boldsymbol{U}_{\mathsf{CS,dr}} = \sum_{i=1}^{N} \sum_{s \in S} \left[\frac{R_{i,s}}{\pi_{i,s}} \cdot \delta_s \boldsymbol{U}_s(\boldsymbol{Y}_i^{(s)o}) + \left(1 - \frac{R_{i,s}}{\pi_{i,s}} \right) \cdot \delta_s E_{\boldsymbol{Y}_i^m | \boldsymbol{Y}_i^o} \boldsymbol{U}_s(\boldsymbol{Y}_i^{(s)}) \right], \quad \text{(E.1)}$$

the term $E_{\boldsymbol{Y}_i^m | \boldsymbol{Y}_i^o} \boldsymbol{U}_s(\boldsymbol{Y}_i^{(s)})$ equals $E_{\boldsymbol{Y}_i^m | \boldsymbol{Y}_i^o}[\boldsymbol{U}_s(\boldsymbol{Y}_i^{(s)o}) + \boldsymbol{U}_s(\boldsymbol{Y}_i^{(s)m}) | \boldsymbol{Y}_i^{(s)o}]$. Now, the expectation over the second term can be replaced by $E_{\boldsymbol{Y}_i^{(s)m} | \boldsymbol{Y}_i^{(s)o}} \boldsymbol{U}_s(\boldsymbol{Y}_i^{(s)m}) | \boldsymbol{Y}_i^{(s)o}$, due to the full exchangeability and the fact that the score contributions arise from derivatives of sub-vectors of $\boldsymbol{Y}_i$. Under this the conditional expectations vanishes and (E.1) reduces to

$$\boldsymbol{U}_{\mathsf{CS,dr}} = \sum_{i=1}^{N} \sum_{s \in S} \delta_s \boldsymbol{U}_s(\boldsymbol{Y}_i^{(s)o}). \quad \text{(E.2)}$$

This makes the naive available case version not only valid, but actually doubly robust. Of course, this is the case only under exchangeability.

E.2 Detailed Calculations for Section 8.2

The general formulation is outlined at the beginning of Section 8.2 resulting in the estimating equations in Table 8.1. In this part of the appendix some more detailed calculations on parts of Section 8.2 can be found.

Let us consider precision estimation. In the singly robust case we must also take into account the uncertainty coming from the parameters of the weight model. The asymptotic variance-covariance matrix can than be estimated using the sandwich estimator in Eq. (8.11). The logistic form of the missingness model equals

$$\pi_i = \prod_{j=2}^{n_i} \left(1 + e^{\boldsymbol{z}'_{ij}\boldsymbol{\psi}}\right)^{-1}$$

and

$$\begin{aligned}
\frac{\partial \pi_i}{\boldsymbol{\psi}} &= -\sum_{k=2}^{n_i} \left(\prod_{j=2, j\neq k}^{n_i} (1 + e^{\boldsymbol{z}'_{ij}\boldsymbol{\psi}})^{-1} \right) \cdot \frac{e^{\boldsymbol{z}'_{ik}\boldsymbol{\psi}}}{\left(1 + e^{\boldsymbol{z}'_{ik}\boldsymbol{\psi}}\right)^2} \cdot \boldsymbol{z}'_{ik} \\
&= -\sum_{k=2}^{n_i} \boldsymbol{z}'_{ik} \cdot e^{\boldsymbol{z}'_{ij}\boldsymbol{\psi}} \cdot \left(1 + e^{\boldsymbol{z}'_{ik}\boldsymbol{\psi}}\right)^{-1} \cdot \prod_{j=2}^{n_i} \left(1 + e^{\boldsymbol{z}'_{ij}\boldsymbol{\psi}}\right)^{-1} \\
&= -\pi_i \cdot \left(\sum_{k=2}^{n_i} \boldsymbol{z}'_{ik} \cdot p_{ik} \right)
\end{aligned}$$

with $\boldsymbol{\psi}$ the missingness parameter and $\boldsymbol{z}_{ij}$ stacked with covariates prior to the jth moment.

The parameter $\boldsymbol{\psi}$ is estimated from the weight model, a logistic regression with

$$\begin{aligned}
L_i &= \prod_{j=2}^{n_i} \frac{e^{R_{ij}\boldsymbol{z}'_{ij}\boldsymbol{\psi}}}{1 + e^{\boldsymbol{z}'_{ij}\boldsymbol{\psi}}} \\
\ell_i &= \sum_{j=2}^{n_i} \left[R_{ij}\boldsymbol{z}'_{ij}\boldsymbol{\psi} - \ln\left(1 + e^{\boldsymbol{z}'_{ij}\boldsymbol{\psi}}\right) \right] \\
\boldsymbol{W}_i &= \sum_{j=2}^{n_i} \boldsymbol{z}_{ij}(R_{ij} - p_{il}) \\
R_{ij} &= \begin{cases} 0 & if \quad j < d_i \\ 1 & if \quad j = d_i \end{cases}
\end{aligned}$$

The first block of Eq. (8.11), $\frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\theta}}$, can be straightforwardly calculated from the Bahadur model with the ψ-parameter kept fixed. The same for the weight model, $\frac{\partial \boldsymbol{W}_i}{\partial \boldsymbol{\psi}}$ can

be deducted from is logistic structure. More attention and thorough calculations should go to the third part, $\frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\psi}}$. In contrast to the other two, this one is not directly computable by software and needs some extra programming. For the three cases respectively:

- CC, sr: $\boldsymbol{V}_i = \frac{\tilde{R}_i}{\pi_i} \sum_{j<k} \boldsymbol{U}_i(y_{ij}, y_{ik})$ and $\boldsymbol{W}_i = \sum_{j=2}^{n_i} \boldsymbol{z}_{ij}(R_{ij} - p_{ij})$

$$\frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\theta}} = \frac{\tilde{R}_i}{\pi_i} \sum_{j<k} \frac{\partial \boldsymbol{U}_i(y_{ij}, y_{ik})}{\partial \boldsymbol{\theta}}$$

$$\frac{\partial \boldsymbol{W}_i}{\partial \boldsymbol{\psi}} = -\sum_{j=2}^{n_i} (\boldsymbol{z}_{ij}\boldsymbol{z}'_{ij}) p_{ij}(1 - p_{ij})$$

$$\frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\psi}} = \frac{\tilde{R}_i}{\pi_i} \sum_{j<k} \boldsymbol{U}_i(y_{ij}, y_{ik}) \left(\sum_{l=2}^{n_i} \boldsymbol{z}'_{il} p_{il} \right)$$

- CP, sr: $\boldsymbol{V}_i = \sum_{j<k<d_i} \frac{R_{ik}}{\pi_{ik}} \boldsymbol{U}_i(y_{ij}, y_{ik})$ and $\boldsymbol{W}_i = \sum_{j=2}^{n_i} \boldsymbol{z}_{ij}(R_{ij} - p_{ij})$

$$\frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\theta}} = \sum_{j<k<d_i} \frac{R_{ik}}{\pi_{ik}} \frac{\partial \boldsymbol{U}_i(y_{ij}, y_{ik})}{\partial \boldsymbol{\theta}}$$

$$\frac{\partial \boldsymbol{W}_i}{\partial \boldsymbol{\psi}} = -\sum_{j=2}^{n_i} (\boldsymbol{z}_{ij}\boldsymbol{z}'_{ij}) p_{ij}(1 - p_{ij})$$

$$\frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\psi}} = \sum_{j<k<d_i} \frac{R_{ik}}{\pi_{ik}} \boldsymbol{U}_i(y_{ij}, y_{ik}) \left(\sum_{l=2}^{k} \boldsymbol{z}'_{il} p_{il} \right)$$

- AC, sr: $\boldsymbol{V}_i = \sum_{j<k} \left[\frac{R_{ij}}{\pi_{ij}} \boldsymbol{U}_i(y_{ij}) + \frac{R_{ik}}{\pi_{ik}} \boldsymbol{U}_i(y_{ik}|y_{ij}) \right]$ and $\boldsymbol{W}_i = \sum_{j=2}^{n_i} \boldsymbol{z}_{ij}(R_{ij} - p_{ij})$

$$\frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\theta}} = \sum_{j<k} \left[\frac{R_{ij}}{\pi_{ij}} \frac{\partial \boldsymbol{U}_i(y_{ij})}{\partial \boldsymbol{\theta}} + \frac{R_{ik}}{\pi_{ik}} \frac{\partial \boldsymbol{U}_i(y_{ik}|y_{ij})}{\partial \boldsymbol{\theta}} \right]$$

$$\frac{\partial \boldsymbol{W}_i}{\partial \boldsymbol{\psi}} = -\sum_{j=2}^{n_i} (\boldsymbol{z}_{ij}\boldsymbol{z}'_{ij}) p_{ij}(1 - p_{ij})$$

$$\frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\psi}} = \sum_{j<k} \left[\frac{R_{ij}}{\pi_{ij}} \boldsymbol{U}_i(y_{ij}) + \frac{R_{ik}}{\pi_{ik}} \boldsymbol{U}_i(y_{ik}|y_{ij}) \right] \left(\sum_{l=2}^{k} \boldsymbol{z}'_{il} p_{il} \right)$$

The doubly robust version is extend with a predictive model for the unobserved responses, given the observed ones. The asymptotic variance-covariance estimator is than outlined in Eq. (8.12), given we use the full expressions in Table 8.1. For the predictive model a separate Bahadur model is implemented, meaning that a third score equation $T(\boldsymbol{\phi})$ representing the conditional Bahadur model joins the entire score.

Here we focus on the first row of Eq. (8.12), as $\frac{\partial \boldsymbol{W}_i}{\partial \boldsymbol{\psi}}$ is identical as in the single robust case and $\frac{\partial \boldsymbol{T}_i}{\partial \boldsymbol{\phi}}$ follows directly form the conditional Bahadur model. For the expectations in the formulas Eq. (8.8)-(8.9) are used.

Also here for the three cases respectively:

- CC, dr: $\boldsymbol{V}_i = \sum_{j<k} \left[\frac{\widetilde{R}_i}{\pi_i} \boldsymbol{U}_i(y_{ij}, y_{ik}) + \left(1 - \frac{\widetilde{R}_i}{\pi_i}\right) E_{\boldsymbol{Y}^m|\boldsymbol{y}^o} \boldsymbol{U}_i(y_{ij}, y_{ik}) \right]$

$$\frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\theta}} = \sum_{j<k} \left[\frac{\widetilde{R}_i}{\pi_i} \frac{\partial \boldsymbol{U}_i(y_{ij}, y_{ik})}{\partial \boldsymbol{\theta}} + \left(1 - \frac{\widetilde{R}_i}{\pi_i}\right) \sum_{y_{ij}=0}^{1} \sum_{y_{ik}=0}^{1} \frac{\partial \boldsymbol{U}_i(y_{ij}, y_{ik})}{\partial \boldsymbol{\theta}} q(y_{ij}, y_{ik}) \right]$$

$$\frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\psi}} = \sum_{j<k} \left[\left(\frac{\widetilde{R}_i}{\pi_i} \boldsymbol{U}_i(y_{ij}, y_{ik}) - \frac{\widetilde{R}_i}{\pi_i} \sum_{y_{ij}=0}^{1} \sum_{y_{ik}=0}^{1} \boldsymbol{U}_i(y_{ij}, y_{ik}) q(y_{ij}, y_{ik}) \right) \left(\sum_{l=2}^{n_i} \boldsymbol{z}'_{il} p_{il} \right) \right]$$

$$\frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\phi}} = \sum_{j<k} \left[\left(1 - \frac{\widetilde{R}_i}{\pi_i}\right) \sum_{y_{ij}=0}^{1} \sum_{y_{ik}=0}^{1} \boldsymbol{U}_i(y_{ij}, y_{ik}) \left(\frac{\partial q(y_{ij}, y_{ik})}{\partial \boldsymbol{\phi}} \right)' \right]$$

- CP, dr: $\boldsymbol{V}_i = \sum_{j<k<n_i} \left[\frac{R_{ijk}}{\pi_{ijk}} \boldsymbol{U}_i(y_{ij}, y_{ik}) + \left(1 - \frac{R_{ijk}}{\pi_{ijk}}\right) E_{\boldsymbol{Y}^m|\boldsymbol{y}^o} \boldsymbol{U}_i(y_{ij}, y_{ik}) \right]$

$$\frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\theta}} = \sum_{j<k<n_i} \left[\frac{R_{ik}}{\pi_{ik}} \frac{\partial \boldsymbol{U}_i(y_{ij}, y_{ik})}{\partial \boldsymbol{\theta}} + \left(1 - \frac{R_{ik}}{\pi_{ik}}\right) \sum_{y_{ij}=0}^{1} \sum_{y_{ik}=0}^{1} \frac{\partial \boldsymbol{U}_i(y_{ij}, y_{ik})}{\partial \boldsymbol{\theta}} q(y_{ij}, y_{ik}) \right]$$

$$\frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\psi}} = \sum_{j<k<n_i} \left[\left(\frac{R_{ik}}{\pi_{ik}} \boldsymbol{U}_i(y_{ij}, y_{ik}) - \frac{R_{ik}}{\pi_{ik}} \sum_{y_{ij}=0}^{1} \sum_{y_{ik}=0}^{1} \boldsymbol{U}_i(y_{ij}, y_{ik}) q(y_{ij}, y_{ik}) \right) \left(\sum_{l=2}^{k} \boldsymbol{z}'_{il} p_{il} \right) \right]$$

$$\frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\phi}} = \sum_{j<k<n_i} \left[\left(1 - \frac{R_{ik}}{\pi_{ik}}\right) \sum_{y_{ij}=0}^{1} \sum_{y_{ik}=0}^{1} \boldsymbol{U}_i(y_{ij}, y_{ik}) \left(\frac{\partial q(y_{ij}, y_{ik})}{\partial \boldsymbol{\phi}} \right)' \right]$$

- AC, dr: $\boldsymbol{V}_i = \sum_{j<k} \frac{R_{ik}}{\pi_{ik}} \boldsymbol{U}_i(y_{ik}|y_{ij}) + \sum_{j=1}^{d_i-1} \frac{R_{ij}}{\pi_{ij}} \boldsymbol{U}_i(y_{ij})$
  $+ \sum_{j<k} \left(1 - \frac{R_{ik}}{\pi_{ik}}\right) E_{\boldsymbol{Y}^m|\boldsymbol{y}^o} \boldsymbol{U}_i(y_{ik}|y_{ij}) + \sum_{j=1}^{d_i-1} \left(1 - \frac{R_{ij}}{\pi_{ij}}\right) E_{\boldsymbol{Y}^m|\boldsymbol{y}^o} \boldsymbol{U}_i(y_{ij})$

$$\frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\theta}} = \sum_{j<k} \frac{R_{ik}}{\pi_{ik}} \frac{\partial \boldsymbol{U}_i(y_{ik}|y_{ij})}{\partial \boldsymbol{\theta}} + \sum_{j=1}^{d_i-1} \frac{R_{ij}}{\pi_{ij}} \frac{\partial \boldsymbol{U}_i(y_{ij})}{\partial \boldsymbol{\theta}}$$
$$+ \sum_{j<k} \left(1 - \frac{R_{ik}}{\pi_{ik}}\right) \sum_{y_{ik}=0}^{1} \frac{\partial \boldsymbol{U}_i(y_{ik}|y_{ij})}{\partial \boldsymbol{\theta}} q(y_{ik}|y_{ij}) + \sum_{j=1}^{d_i-1} \left(1 - \frac{R_{ij}}{\pi_{ij}}\right) \sum_{y_{ij}=0}^{1} \frac{\partial \boldsymbol{U}_i(y_{ij})}{\boldsymbol{\theta}} q(y_{ij})$$
$$\frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\psi}} = \sum_{j<k} \left[\left(\frac{R_{ik}}{\pi_{ik}} \boldsymbol{U}_i(y_{ik}|y_{ij}) - \frac{R_{ik}}{\pi_{ik}} \sum_{y_{ik}=0}^{1} \boldsymbol{U}_i(y_{ik}|y_{ij}) q(y_{ik}|y_{ij}) \right) \left(\sum_{l=2}^{k} \boldsymbol{z}'_{il} p_{il} \right) \right]$$
$$+ \sum_{j=1}^{d_i-1} \left[\left(\frac{R_{ij}}{\pi_{ij}} \boldsymbol{U}_i(y_{ij}) - \frac{R_{ij}}{\pi_{ij}} \sum_{y_{ij}=0}^{1} \boldsymbol{U}_i(y_{ij}) q(y_{ij}) \right) \left(\sum_{l=2}^{j} \boldsymbol{z}'_{il} p_{il} \right) \right]$$
$$\frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\phi}} = \sum_{j<k} \left(1 - \frac{R_{ik}}{\pi_{ik}}\right) \boldsymbol{U}_i(y_{ik}|y_{ij}) \left(\frac{\partial q(y_{ik}|y_{ij})}{\partial \boldsymbol{\phi}} \right)' + \sum_{j=1}^{d_i-1} \left(1 - \frac{R_{ij}}{\pi_{ij}}\right) \boldsymbol{U}_i(y_{ij}) \left(\frac{\partial q(y_{ij})}{\partial \boldsymbol{\phi}} \right)'$$

All three expressions coincide as expressed in Eq. (8.1). In this case it is not needed to explicitly model the missingness and the entire score reduces to $\boldsymbol{S}_i = (\boldsymbol{V}_i, \boldsymbol{T}_i)$. The asymptotic variance-covariance matrix can be estimated using

$$I_0 = \frac{1}{N} \sum_{i=1}^{N} \begin{pmatrix} \frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\theta}} & \frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\phi}} \\ \mathbf{0} & \frac{\partial \boldsymbol{T}_i}{\partial \boldsymbol{\phi}} \end{pmatrix} \quad \text{and} \quad I_1 = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{S}_i(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\phi}}) \boldsymbol{S}'_i(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\phi}}). \quad \text{(E.3)}$$

Notice that to avoid complex and too long formulas for the derivatives in this section of the appendix, factors representing a deviation with the amount of pairs or pairs with a single observation are omitted. Although these factors most of time cancel when solving the estimating equations for the parameters, they are necessary for the calculation of the standard errors as they get estimated to big otherwise.

E.3 Implementation with SAS

E.3.1 Full Likelihood Based Bahadur Model

To apply the full likelihood Bahadur model, the original format of the GSA data was used:

PATID	gsabin1	gsabin2	gsabin3	gsabin4	pca0	intercept
1	0	.	.	.	3	1
2	1	.	.	.	5	1

```
     3          1          1          1          0     1          1
     4          1          0          .          .     4          1
     6          0          .          .          .     4          1
...
```

Further, the NLMIXED procedure of SAS (SAS 9.4) was applied, when all possible profiles were considered. For a theoretical background on the Bahadur model, see Section 2.2.2. Note, that variable *intercept* is used here only to make it possible to apply the general(.) statement of proc nlmixed. For the first-order correlations, a Toepliz structure was assumed, for the higher-order correlations a general form of the correlations.

```
proc nlmixed data=gsa;
parms beta0=0 beta1=0 beta2=0 beta3=0 rho_1 rho_2=0 rho_3=0 rho_123=0
rho_124=0 rho_134=0 rho_234=0 rho_1234=0;

eta_1 = beta0 + beta1*1 + beta2*1+ beta3*pca0 ;
eta_2 = beta0 + beta1*2 + beta2*4+ beta3*pca0 ;
eta_3 = beta0 + beta1*3 + beta2*9+ beta3*pca0 ;
eta_4 = beta0 + beta1*4 + beta2*16+ beta3*pca0 ;

nu_1=exp(eta_1)/(1+exp(eta_1));
nu_2=exp(eta_2)/(1+exp(eta_2));
nu_3=exp(eta_3)/(1+exp(eta_3));
nu_4=exp(eta_4)/(1+exp(eta_4));

e_1=(gsabin1-nu_1)/sqrt(nu_1*(1-nu_1));
e_2=(gsabin2-nu_2)/sqrt(nu_2*(1-nu_2));
e_3=(gsabin3-nu_3)/sqrt(nu_3*(1-nu_3));
e_4=(gsabin4-nu_4)/sqrt(nu_4*(1-nu_4));

if (gsabin2 =. and gsabin3 =. and gsabin4 =.)then do;
c=1;
loglik=log(c)+gsabin1*log(nu_1)+(1-gsabin1)*log(1-nu_1);
end; else
if (gsabin2 ne . and gsabin3 =. and gsabin4 =.) then do;
c=1+rho_1*e_1*e_2;
loglik=log(c)+gsabin1*log(nu_1)+(1-gsabin1)*log(1-nu_1)+
gsabin2*log(nu_2)+(1-gsabin2)*log(1-nu_2);
end; else
if (gsabin2 ne . and gsabin3 ne . and gsabin4 =.) then do;
c=1+rho_1*e_1*e_2+rho_2*e_1*e_3+rho_1*e_2*e_3+rho_123*e_1*e_2*e_3;
loglik=log(c)+gsabin1*log(nu_1)+(1-gsabin1)*log(1-nu_1)+
gsabin2*log(nu_2)+(1-gsabin2)*log(1-nu_2)+gsabin3*log(nu_3)+
(1-gsabin3)*log(1-nu_3);
end; else
if (gsabin2 ne . and gsabin3 ne . and gsabin4 ne .) then do;
```

```
c=1+rho_1*e_1*e_2+rho_2*e_1*e_3+rho_3*e_1*e_4+rho_1*e_2*e_3+rho_2*e_2*e_4+
rho_1*e_3*e_4+rho_123*e_1*e_2*e_3+rho_124*e_1*e_2*e_4+rho_134*e_1*e_3*e_4
+rho_234*e_2*e_3*e_4+rho_1234*e_1*e_2*e_3*e_4;
loglik=log(c)+gsabin1*log(nu_1)+(1-gsabin1)*log(1-nu_1)+gsabin2*log(nu_2)+
(1-gsabin2)*log(1-nu_2)+gsabin3*log(nu_3)+(1-gsabin3)*log(1-nu_3)+
gsabin4*log(nu_4)+(1-gsabin4)*log(1-nu_4);
end;
model intercept ~ general(loglik);
run;
```

E.3.2 Pairwise Likelihood Based Model

In this section we list some examples of data formatting and the code to fit the model to different cases of the estimating equation. In these examples, some cases are omitted due to their mutual similarity. Complete cases and complete pairs are different in their inverse probability weights: for complete cases, the weights are calculated based on the probability of a patient to be completely observed, for complete pairs the probability to be observed until a certain time-point. After the model is fitted, the principle of the sandwich-type robust variance estimation will be applied to obtain precision estimates.

E.3.2.1 Naive Case

For modeling using naive estimating equations, all possible pairs from the sequence of GSA measurement per patient were considered. If the response measurement of a pair was missing, the 999 code was used instead. As result, the data was re-formated as follows:

PATID	pca0	intercept	responsej	timej	timej_2	responsek	timek	timek_2
1	3	1	0	1	1	999	2	4
1	3	1	0	1	1	999	3	9
1	3	1	0	1	1	999	4	16
2	5	1	1	1	1	999	2	4
2	5	1	1	1	1	999	3	9
2	5	1	1	1	1	999	4	16
3	1	1	1	1	1	1	2	4
3	1	1	1	1	1	1	3	9
3	1	1	1	1	1	0	4	16
3	1	1	1	2	4	1	3	9
3	1	1	1	2	4	0	4	16
3	1	1	1	3	9	0	4	16
4	4	1	1	1	1	0	2	4
4	4	1	1	1	1	999	3	9
4	4	1	1	1	1	999	4	16
4	4	1	0	2	4	999	3	9

```
    4     4        1          0        2        4        999       4        16
    6     4        1          0        1        1        999       2         4
    6     4        1          0        1        1        999       3         9
    6     4        1          0        1        1        999       4        16
 ...
```

For the available cases the whole data was used, for the complete pairs and complete cases the corresponding subject selection. Herewith, the code for the available cases:

```
proc nlmixed data=model_gsa_ac_naive;
parms beta0=0 beta1=0 beta2=0 beta3=0 rho_1=0 rho_2=0 rho_3=0;
eta_j = beta0 + beta1*timej + beta2*timej_2+ beta3*pca0 ;
eta_k = beta0 + beta1*timek + beta2*timek_2+ beta3*pca0 ;
nu_j = exp(eta_j)/(1+exp(eta_j));
nu_k = exp(eta_k)/(1+exp(eta_k));

if (timej=1 and timek=2) or (timej=2 and timek=3) or (timej=3 and timek=4)
then mu11 = nu_j*nu_k + rho_1*sqrt(nu_j*(1-nu_j)*nu_k*(1-nu_k)); else
if (timej=1 and timek=3) or (timej=2 and timek=4)
then mu11 = nu_j*nu_k + rho_2*sqrt(nu_j*(1-nu_j)*nu_k*(1-nu_k)); else
if (timej=1 and timek=4)
then mu11 = nu_j*nu_k + rho_3*sqrt(nu_j*(1-nu_j)*nu_k*(1-nu_k));

mu10 = nu_j - mu11;
mu01 = nu_k - mu11;
mu00 = 1 - (mu11+mu10+mu01);

if responsej = 1 and responsek = 1 then loglik=log(mu11); else
if responsej = 1 and responsek = 0 then loglik=log(mu10); else
if responsej = 1 and responsek = 999 then loglik=log(mu11 + mu10); else
if responsej = 0 and responsek = 1 then loglik=log(mu01); else
if responsej = 0 and responsek = 0 then loglik=log(mu00); else
if responsej = 0 and responsek = 999 then loglik=log(mu00+mu01);

model intercept ~ general(loglik);
run;
```

E.3.2.2 Singly Robust Case

First, a dropout model with previous measurement was considered, using the data in the following format and the corresponding SAS code:

```
PATID    pca0    ID_AC    time    GSAbin    dropout    prev
    1       3        1       1         0          0       .
    1       3        1       2         .          1       0
    1       3        1       3         .          1       .
```

```
   1      3       1       4       .       1       .
   2      5       2       1       1       0       .
   2      5       2       2       .       1       1
   2      5       2       3       .       1       .
   2      5       2       4       .       1       .
   3      1       3       1       1       0       .
   3      1       3       2       1       0       1
   3      1       3       3       1       0       1
   3      1       3       4       0       0       1

...

proc nlmixed data=model_weights;
parms gamma0=0 gamma1=0 gamma2=0;
eta = gamma0 + gamma1*prev + gamma2*pca0 ;
nu = exp(eta)/(1+exp(eta));
if dropout = 1 then loglik=log(nu); else
if dropout = 0 then loglik=log(1-nu);
model dropout ~ general(loglik);
run;
```

Based on probabilities estimated with the dropout model, the inverse probabilities weights were calculated. So, the data for the analysis of the available cases are as follows:

```
 PATID pca0 intercept responsej   wij   timej timej_2 responsek   wik   timek timek_2
     1   3      1         0     1.00000   1      1       999    1.50253   2       4
     1   3      1         0     1.00000   1      1       999     .        3       9
     1   3      1         0     1.00000   1      1       999     .        4      16
     2   5      1         1     1.00000   1      1       999    1.25292   2       4
     2   5      1         1     1.00000   1      1       999     .        3       9
     2   5      1         1     1.00000   1      1       999     .        4      16
     3   1      1         1     1.00000   1      1         1    1.16671   2       4
     3   1      1         1     1.00000   1      1         1    1.36121   3       9
     3   1      1         1     1.00000   1      1         0    1.58814   4      16
     3   1      1         1     1.16671   2      4         1    1.36121   3       9
     3   1      1         1     1.16671   2      4         0    1.58814   4      16
     3   1      1         1     1.36121   3      9         0    1.58814   4      16
...
```

The code for the available cases is as follows:

```
proc nlmixed data=model_gsa_wi_ac;
parms beta0=0 beta1=0 beta2=0 beta3=0 rho_1=0 rho_2=0 rho_3=0;
eta_j = beta0 + beta1*timej + beta2*timej_2+ beta3*pca0 ;
eta_k = beta0 + beta1*timek + beta2*timek_2+ beta3*pca0 ;
nu_j = exp(eta_j)/(1+exp(eta_j));
nu_k = exp(eta_k)/(1+exp(eta_k));

if (timej=1 and timek=2) or (timej=2 and timek=3) or (timej=3 and timek=4)
```

```
then mu11 = nu_j*nu_k + rho_1*sqrt(nu_j*(1-nu_j)*nu_k*(1-nu_k)); else
if (timej=1 and timek=3) or (timej=2 and timek=4)
then mu11 = nu_j*nu_k + rho_2*sqrt(nu_j*(1-nu_j)*nu_k*(1-nu_k)); else
if (timej=1 and timek=4)
then mu11 = nu_j*nu_k + rho_3*sqrt(nu_j*(1-nu_j)*nu_k*(1-nu_k));

mu10 = nu_j - mu11;
mu01 = nu_k - mu11;
mu00 = 1 - (mu11+mu10+mu01);

if responsek = 1 and responsej = 1 then loglik=wik*log(mu11/(mu11 + mu10)); else
if responsek = 0  and responsej = 1 then loglik=wik*log(mu10/(mu11 + mu10)); else
if responsek = 1 and responsej = 0 then loglik=wik*log(mu01/(mu01 + mu00)); else
if responsek = 0  and responsej = 0 then loglik=wik*log(mu00/(mu01 + mu00)); else

if responsej = 1 and responsek = 999 then loglik= wij*log(mu11 + mu10); else
if responsej = 0 and responsek = 999 then loglik= wij*log(mu00 + mu01);

model intercept ~ general(loglik);
run;
```

As for naive case, the data and the code was adopted for the complete pairs and complete cases.

E.3.2.3 Doubly Robust Case

The inverse probability weights were calculated using the same way as for the singly robust case.

To model the expectations in doubly robust case, a "help" Bahadur model was applied. To model the marginal probabilities, the expression for the linear predictors were defined at different time-points. To model pairs, assuming that all previous measurements are available, the "history"parameters were used taking as reference for the history the response with the lower value of the time point (e.g., in pair Y2 - Y3 we used only gsabin1 as the history covariate). If some of the measurements are missing, the corresponding term in the model will be omitted.

```
proc nlmixed data=model_gsa_wi_AC;
parms kappa0=0 kappa1=0 kappa2=0 kappa3=0 omega1=0 omega2=0 omega3=0 tau_1=0 tau_2=0 tau_3=0;

if gsabin2 ne . then do;
if (timej=1 and timek=2) or (timej=1 and timek=3) or (timej=1 and timek=4) then do;
eta_j = kappa0 + kappa1*timej + kappa2*timej_2+ kappa3*pca0 ;
eta_k = kappa0 + kappa1*timek + kappa2*timek_2+ kappa3*pca0 ;
end; else
if timej=2 and timek=3 then do;
```

```
eta_j = kappa0 + kappa1*timej + kappa2*timej_2+ kappa3*pca0 + omega1*gsabin1;
eta_k = kappa0 + kappa1*timek + kappa2*timek_2+ kappa3*pca0 + omega2*gsabin1;
end; else
if timej=2 and timek=4 then do;
eta_j = kappa0 + kappa1*timej + kappa2*timej_2+ kappa3*pca0 + omega1*gsabin1;
eta_k = kappa0 + kappa1*timek + kappa2*timek_2+ kappa3*pca0 + omega3*gsabin1;
end; else
if timej=3 and timek=4 then do;
eta_j = kappa0 + kappa1*timej + kappa2*timej_2+ kappa3*pca0 + omega2*gsabin1 + omega1*gsabin2;
eta_k = kappa0 + kappa1*timek + kappa2*timek_2+ kappa3*pca0 + omega3*gsabin1 + omega2*gsabin2;
end;
end;

if gsabin2 = . then do;
if (timej=1 and timek=2) or (timej=1 and timek=3) or (timej=1 and timek=4) then do;
eta_j = kappa0 + kappa1*timej + kappa2*timej_2+ kappa3*pca0 ;
eta_k = kappa0 + kappa1*timek + kappa2*timek_2+ kappa3*pca0 ;
end; else
if timej=2 and timek=3 then do;
eta_j = kappa0 + kappa1*timej + kappa2*timej_2+ kappa3*pca0 + omega1*gsabin1;
eta_k = kappa0 + kappa1*timek + kappa2*timek_2+ kappa3*pca0 + omega2*gsabin1;
end; else
if timej=2 and timek=4 then do;
eta_j = kappa0 + kappa1*timej + kappa2*timej_2+ kappa3*pca0 + omega1*gsabin1;
eta_k = kappa0 + kappa1*timek + kappa2*timek_2+ kappa3*pca0 + omega3*gsabin1;
end; else
if timej=3 and timek=4 then do;
eta_j = kappa0 + kappa1*timej + kappa2*timej_2+ kappa3*pca0 + omega2*gsabin1;
eta_k = kappa0 + kappa1*timek + kappa2*timek_2+ kappa3*pca0 + omega3*gsabin1;
end;
end;

q_j = exp(eta_j)/(1+exp(eta_j));
q_k = exp(eta_k)/(1+exp(eta_k));

if (timej=1 and timek=2) or (timej=2 and timek=3) or (timej=3 and timek=4)
then mu11 = q_j*q_k + tau_1*sqrt(q_j*(1-q_j)*q_k*(1-q_k)); else
if (timej=1 and timek=3) or (timej=2 and timek=4)
then mu11 = q_j*q_k + tau_2*sqrt(q_j*(1-q_j)*q_k*(1-q_k)); else
if (timej=1 and timek=4)
then mu11 = q_j*q_k + tau_3*sqrt(q_j*(1-q_j)*q_k*(1-q_k));

mu10 = q_j - mu11;
mu01 = q_k - mu11;
mu00 = 1 - (mu11+mu10+mu01);

if responsej = 1 and responsek = 1 then loglik=log(mu11); else
```

```
if responsej = 1 and responsek = 0 then loglik=log(mu10); else
if responsej = 0 and responsek = 1 then loglik=log(mu01); else
if responsej = 0 and responsek = 0 then loglik=log(mu00);

if responsej = 1 and responsek = 999 then loglik= log(mu11 + mu10); else
if responsej = 0 and responsek = 999 then loglik= log(mu00 + mu01);

model intercept ~ general(loglik);
run;
```

Further, the pairs are complemented with expectations. The data for the analysis looks as follows:

```
      i  r         r
      n  e         e
      t  s     t   s       t
      e  p     i   p       i
P     r  o  t  m   o   t   m
A  p  c  n  i  e   n   i   e
T  c  e  s  m  j   s   m   k     w        w        q        q        q        q
I  a  p  e  e  _   e   e   _     i        i        1        1        0        0
D  0  t  j  j  2   k   k   2     j        k        1        0        1        0

1  3  1  0  1  1  999  2   4  1.00000  1.50253  0.65036  0.16425  0.09806  0.08733
1  3  1  0  1  1  999  3   9  1.00000  1.50253  0.65232  0.16229  0.09357  0.09182
1  3  1  0  1  1  999  4  16  1.00000  1.50253  0.70389  0.11072  0.10458  0.08081
2  5  1  1  1  1  999  2   4  1.00000  1.25292  0.55410  0.19757  0.11795  0.13039
2  5  1  1  1  1  999  3   9  1.00000  1.25292  0.55665  0.19501  0.11244  0.13590
2  5  1  1  1  1  999  4  16  1.00000  1.25292  0.61516  0.13650  0.12893  0.11940
3  1  1  1  1  1    1  2   4  1.00000  1.16671  0.73422  0.13026  0.07777  0.05775
3  1  1  1  1  1    1  3   9  1.00000  1.36121  0.73566  0.12882  0.07427  0.06125
3  1  1  1  1  1    0  4  16  1.00000  1.58814  0.77841  0.08607  0.08129  0.05423
3  1  1  1  2  4    1  3   9  1.16671  1.36121  0.77933  0.07623  0.09917  0.04526
3  1  1  1  2  4    0  4  16  1.16671  1.58814  0.79464  0.06092  0.10095  0.04349
3  1  1  1  3  9    0  4  16  1.36121  1.58814  0.86708  0.04133  0.06863  0.02296
...
```

The code for available cases is as follows:

```
proc nlmixed data=model_gsa_wi_qs_ac;
parms beta0=0 beta1=0 beta2=0 beta3=0 rho_1=0 rho_2=0 rho_3=0;
eta_j = beta0 + beta1*timej + beta2*timej_2+ beta3*pca0 ;
eta_k = beta0 + beta1*timek + beta2*timek_2+ beta3*pca0 ;
nu_j = exp(eta_j)/(1+exp(eta_j));
nu_k = exp(eta_k)/(1+exp(eta_k));

if (timej=1 and timek=2) or (timej=2 and timek=3) or (timej=3 and timek=4)
```

```
then mu11 = nu_j*nu_k + rho_1*sqrt(nu_j*(1-nu_j)*nu_k*(1-nu_k)); else
if (timej=1 and timek=3) or (timej=2 and timek=4)
then mu11 = nu_j*nu_k + rho_2*sqrt(nu_j*(1-nu_j)*nu_k*(1-nu_k)); else
if (timej=1 and timek=4)
then mu11 = nu_j*nu_k + rho_3*sqrt(nu_j*(1-nu_j)*nu_k*(1-nu_k));

mu10 = nu_j - mu11;
mu01 = nu_k - mu11;
mu00 = 1 - (mu11+mu10+mu01);

Exp_cond1=q11/(q11+q10)*log(mu11/(mu11+mu10))+q10/(q11+q10)*log(mu10/(mu11+mu10));
/*Responsej=1*/
Exp_cond0=q01/(q01+q00)*log(mu01/(mu01+mu00))+q00/(q01+q00)*log(mu00/(mu01+mu00));
/*Responsej=0*/


if responsek = 1 and responsej = 1 then  loglik=wik*log(mu11/(mu11 + mu10))
+ (1-wik)*Exp_cond1; else
if responsek = 0  and responsej = 1 then loglik=wik*log(mu10/(mu11 + mu10))
+ (1-wik)*Exp_cond1; else
if responsek = 1 and responsej = 0 then loglik=wik*log(mu01/(mu01 + mu00))
+ (1-wik)*Exp_cond0; else
if responsek = 0  and responsej = 0 then loglik=wik*log(mu00/(mu01 + mu00))
+ (1-wik)*Exp_cond0; else

if responsej = 1 and responsek = 999 then loglik= wij*log(mu11 + mu10) + (1-wik)*Exp_cond1
                      + (1-wij)*log(q11+q10); else
if responsej = 0 and responsek = 999 then loglik= wij*log(mu00 + mu01)+ (1-wik)*Exp_cond0
                      + (1-wij)*log(q01+q00);

model intercept ~ general(loglik);
run;
```

After selecting the data for the complete pairs, the model can be fitted using the following program:

```
proc nlmixed data=model_gsa_wi_qs_cp;
parms beta0=0 beta1=0 beta2=0 beta3=0 rho_1=0 rho_2=0 rho_3=0;
eta_j = beta0 + beta1*timej + beta2*timej_2+ beta3*pca0 ;
eta_k = beta0 + beta1*timek + beta2*timek_2+ beta3*pca0 ;
nu_j = exp(eta_j)/(1+exp(eta_j));
nu_k = exp(eta_k)/(1+exp(eta_k));

if (timej=1 and timek=2) or (timej=2 and timek=3) or (timej=3 and timek=4)
then mu11 = nu_j*nu_k + rho_1*sqrt(nu_j*(1-nu_j)*nu_k*(1-nu_k)); else
if (timej=1 and timek=3) or (timej=2 and timek=4)
then mu11 = nu_j*nu_k + rho_2*sqrt(nu_j*(1-nu_j)*nu_k*(1-nu_k)); else
if (timej=1 and timek=4)
```

```
then mu11 = nu_j*nu_k + rho_3*sqrt(nu_j*(1-nu_j)*nu_k*(1-nu_k));

mu10 = nu_j - mu11;
mu01 = nu_k - mu11;
mu00 = 1 - (mu11+mu10+mu01);

Exp_joint=q11*log(mu11)+q10*log(mu10)+q01*log(mu01)+q00*log(mu00);
if responsej = 1 and responsek = 1 then loglik=wik*log(mu11)+(1-wik)*Exp_joint; else
if responsej = 1 and responsek = 0 then loglik=wik*log(mu10)+(1-wik)*Exp_joint; else
if responsej = 0 and responsek = 1 then loglik=wik*log(mu01)+(1-wik)*Exp_joint; else
if responsej = 0 and responsek = 0 then loglik=wik*log(mu00)+(1-wik)*Exp_joint;

model intercept ~ general(loglik);
run;
```

And for the selected complete cases data, similar code will be applied to fit the model.

Summary

A random sample is not always of a fixed, *a priori* determined size. Examples include sequential sampling and stopping rules, missing data, and clusters with random size. Often there then is no complete sufficient statistic. Completeness means that any measurable function of a sufficient statistic that has zero expectation for every value of the parameter indexing the parametric model class, is the zero function almost everywhere. A simple characterization of incompleteness is given for the exponential family in terms of the mapping between the sufficient statistic and the parameter, based upon the implicit function theorem. Essentially this is a comparison of the dimension of the sufficient statistic to the length of the parameter vector. This results in an easy verifiable criterion for incompleteness, clear and simple to use, even for complex settings as is shown for missing data and clusters of random size.

The analysis of hierarchical data that take the form of clusters with random size has received considerable attention in literature. In this work, the focus was on clustered data with unequal cluster sizes, meaning that a joint model of outcome and sample size was not studied. Also, the focus here was 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 become prohibitive. Sample-splitting (Molenberghs, Verbeke, and Iddi, 2011) was proposed as a way to replace iterative optimization of a likelihood that does not admit an analytical solution, with closed-form calculations. Pseudo-likelihood (Molenberghs *et al.*, 2014), consisting of computing weighted averages over solutions obtained from subsamples created according to sample size, was used. As a result, the statistical properties of this approach were investigated. In a first attempt, the compound-symmetry variance structure was used to investigate this modelling framework. In a subsample with only clusters of

the same size, there are closed-from solutions and other useful properties can be obtained. The operational characteristics are studied using simulations. It follows that the proposed non-iterative methods have a strong beneficial impact on computation time.

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

The odds ratio is a frequently used measure to investigate the association between binary variables. Often, such outcomes are measured across strata of different sizes. Mantel and Haenszel (1959) proposed estimators for a common odds ratio, taking into account the stratification. The most common one is among the best known and most used estimators in statistics.

The setting studied by Mantel and Haenszel fits within this framework of sample-splitting and combining with proper weights. The Mantel and Haenszel estimator does not follow from optimality considerations, but nevertheless has properties similar to and often better than the optimal estimator. This was done by comparing it to the optimal estimator, whose existence was demonstrated in spite of the absence of complete sufficient statistics. It is shown, via simulations, that the optimal estimator outperforms the Mantel-Haenszel estimator only in certain settings with huge sample sizes.

Missing data is almost inevitable in correlated-data studies. For non-Gaussian outcomes with moderate to large sequences, direct-likelihood methods can involve complex, hard-to-manipulate likelihoods. Popular alternative approaches, like generalized estimating equations, that are frequently used to circumvent the computational complexity of full likelihood, are less suitable when scientific interest, at least in part, is placed on the association structure; pseudo-likelihood methods are then a viable alternative. When the missing data are missing at random, Molenberghs et al. (2011) proposed a suite of corrections to the standard form of pseudo-likelihood, taking the form of singly and doubly robust estimators. They provided the basis, and exemplified it in insightful yet primarily illustrative examples. The important case of marginal models for hierarchical binary data was considered. Our doubly robust estimator is more convenient than the classical doubly robust estimators. The ideas are illustrated using a marginal model for a binary response, more specifically a Bahadur model.

Samenvatting

Een steekproef is niet steeds van een vaste, vooraf bepaalde grootte. Voorbeelden zijn sequentiële studies, ontbrekende gegevens en ongebalanceerde hiërarchische data. In dit soort settings is er vaak geen *complete sufficient statistic*. Een eenvoudige karakterisering van *completeness* wordt geformuleerd voor de exponentiële familie in termen van de dimensievergelijking tussen de *sufficient statistic* en de parameter, gebaseerd op de impliciete functiestelling. Het is een eenvoudig en makkelijk verifieerbaar criterium, zelfs voor complexe settings met ontbrekende gegevens en ongebalanceerde hiërarchische data.

Ongebalanceerde hiërarchische data werd al vanuit verschillende invalshoeken bestudeerd. In deze thesis ligt de focus op steekproeven die zeer groot zijn, m.a.w. veel clusters of veel metingen per cluster, en die zeer klein zijn (studies van zeldzame ziekten). De *Maximum likelihood estimator* bepalen in middelgrote steekproeven is goed uitvoerbaar, maar in de settings die hier besproken worden, kan dat moeilijkheden met zich meebrengen, zoals geen analystische oplossingen van gesloten vorm en de likelihoodsfunctie kan alleen iteratief geoptimaliseerd worden. Bijgevolg werd de steekproef opgedeeld in stukken naargelang de grootte van de clusters (Molenberghs, Verbeke, and Iddi, 2011). Deze deelsteekproeven werden hierdoor gebalanceerd en resulteren wel in oplossingen van gesloten vorm. Een *pseudo-likelihood* werd gebruikt om de oplossingen van elke deelsteekproef te combineren gebruikmakend van gewichten. De eigenschappen van deze methodologie werden in detail onderzocht op gebalanceerde data die een *compound-symmetry* covariantiestructuur volgen. Via een simulatiestudie werd de toepasbaarheid onderzocht. Hieruit volgt dat deze niet-iteratieve methode slechts een korte berekeningstijd vereist en zeer precies is.

Vervolgens werd deze schattingsmethode verder onderzocht in een ongebalanceerde hiërarchische dataset met een *autoregressive (AR(1))* covariantiestructuur. Ook hier is deze methode bijna even efficiënt als maximum likelihood en de berekeningstijd is veel

lager.

The *odds ratio* is een statistiek die frequent gebruikt wordt om de associatie tussen binaire variabelen te onderzoeken. Ook in dit soort settings kunnen er groeperingen van de gegevens voorkomen van ongelijke grootte. De meeste gekende en gebruikte schatter is deze ontworpen door Mantel and Haenszel (1959).

De schatter combineert de *odds ratio* van subpopulaties in een gewogen schatter, maar volgt niet vanuit optimalisatieberekeningen. The Mantel en Haenszel schatter werd vergeleken met de optimale schatter. Hieruit kan geconcludeerd worden dat de Mantel en Haenszel schatter over zeer goede eigenschappen beschikt. Enkel in settings met zeer grote steekproefgroottes zal de optimale schatter het beter dan doen de Mantel en Haenszel schatter.

Ontbrekende gegevens komen zeer vaak voor in dit soort settings. Voor niet-normaalverdeelde gegevens van een zeer grote steekproef, kunnen de berekeningen van de likelihoodsfunctie zeer complex worden. *Generalized estimating equations* is dan een goed alternatief, maar minder geschikt indien de interesse (gedeeltelijk) gaat naar de correlatiestructuur van de data. Pseudo-likelihoodsfuncties zijn hier beter geschikt. Wanneer de ontbrekende gegevens *missing at random* zijn, maakte Molenberghs et al. (2011) enkelvoudige en dubbelvoudige robuste aanpassingen aan de standaard pseudo-likelihoodsfunctie om correcte inferentie te kunnen doen. Waar dat zij de algemene basis hiervan vormden, focuste dit werk op marginale modellen voor hiërarchische binare data. Een Bahadur model werd hier gekozen als marginaal model.