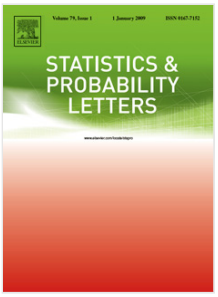
Pseudo-likelihood methodology for partitioned large and complex samples

Geert Molenberghs, Geert Verbeke, Samuel Iddi

Accepted Manuscript



Pseudo-likelihood methodology for partitioned large and complex samples

Geert Molenberghs, Geert Verbeke, Samuel Iddi

PII: S0167-7152(11)00019-8
DOI: 10.1016/j.spl.2011.01.012
Reference: STAPRO 5887

To appear in: *Statistics and Probability Letters*

Please cite this article as: Molenberghs, G., Verbeke, G., Iddi, S., Pseudo-likelihood methodology for partitioned large and complex samples. *Statistics and Probability Letters* (2011), doi:10.1016/j.spl.2011.01.012

Pseudo-likelihood Methodology for Partitioned Large and Complex Samples

Geert Molenberghs[1,2], **Geert Verbeke**[2,1], **and Samuel Iddi**[2]

Interuniversity Insitute for Biostatistics and statistical Bioinformatics

[1] Universiteit Hasselt, Agoralaan 1, 3590 Diepenbeek, Belgium

[2] Katholieke Universiteit Leuven, Kapucijnenvoer 35, 3000 Leuven, Belgium

Abstract

Large data sets, either coming from a large number of independent replications, or because of hierarchies in the data with large numbers of within-unit replication, may pose challenges to the data analyst up to the point of making conventional inferential methods, such as maximum likelihood, prohibitive. Based on general pseudo-likelihood concepts, we propose a method to partition such a set of data, analyze each partition member, and properly combine the inferences into a single one. It is shown that the method is fully efficient for independent partitions, while with dependent sub-samples efficiency is sometimes but not always equal to one. It is argued that, for important realistic settings, efficiency is often very high. Illustrative examples enhance insight in the method's operation, while real-data analysis underscores its power for practice.

Keywords: Asymptotic relative efficiency; Compound-symmetry; Small-sample relative efficiency.

1 Introduction

Contemporary statistical analysis is confronted with data sets of ever increasing dimensions, not only because the number of independent units may become very large, but also owing to hierarchies in the data with large numbers of within-unit replication. While computational and data-analytic resources have recently progressed tremendously, sheer size may be a limiting factor. Our method allows for data analysis that otherwise would be prohibitive, either due to extremely large sample size or a large amount of dependent replication.

Using pseudo-likelihood technology, we propose a method whereby a dataset can be chopped into portions, each of which is then analyzed separately, followed by combining inferences into a single set. The method is inspired by work of Fieuws and Verbeke (2006) and Fieuws *et al* (2006), reviewed in Molenberghs and Verbeke (2005, Ch. 24), who constructed a tool for the analysis of high-dimensional longitudinal data.

Our focus is on proper partitions, where every data point is a member of one and exactly one of the, say M, subsamples. Our method could be generalized to situations with overlapping subsamples as well. However, proper partitions allow us to use maximum likelihood on each subsample, thus simplifying the combination of M inferences into a single one considerably.

A particularly interesting special case is when partitioning is done in terms of independent units (i.e., dependent data are not spread over different members subsample). In that case, the method is fully efficient in the sense that the Cramèr-Rao lower bound is reached. This is true, even though the actual estimator may differ from the one that would result were the data analyzed in full. These results are illustrated using univariate normal and univariate Bernoulli samples.

When correlated data are partitioned by splitting a sequence of dependent data across different subsamples, then full efficiency is sometimes, but not always, reached. An example of this setting is when for each subject in a longitudinal study, Mn repeated measurements are taken, which are subsequently split into M sequences of length n. This results in M separate longitudinal sets of data, each one having shorter length. This situation is useful when the length of a sequence is prohibiting efficient computation. Unlike in the previous case, a dependent partition results. This case is illustrated using the simple, insightful example of a multivariate normal sample with constant mean and compound-symmetry covariance structure. It will be shown that such dependent partitioning can be fully efficient for some but not necessarily for all parameters. The latter remarks begs the questions as to the efficiency of the method. The advantage of our illustrative examples is that the asymptotic relative efficiency, and its small-sample counterparts, can be studied without difficulty.

Of course, the illustrative examples have limited use in practice. The method is devised for settings that genuinely poses computational challenges that can be alleviated by partitioning the data set. We apply our method to a data comprising long binary repeated-measures sequences.

The paper is organized as follows. The motivating data are introduced in Section 2; their analysis is relegated to Section 6. The methodology for our method is presented in Section 3, with attention for general pseudo-likelihood concepts, the method of Fieuws and Verbeke, and our partitioned-sample case, with emphasis on both independent and dependent partitions. The aforementioned illustrations are described in Section 4 and efficiency scrutinized in Section 5.

2 Motivating Case Studies

The motivating case study encompasses data collected by the European Monitoring Centre for Drugs and Drug Addiction (EMCDDA; Mathei *et al* 2006). It consist of annual serological surveys, providing information about the hepatitis C virus (HCV) and human immunodeficiency virus (HIV) status and related risk factors from the 20 Italian regions in the period 1998-2006. Respondents are drug users who sought help in the drug treatment centers. In each year, large numbers of drug users are tested for their HCV and HIV statuses. The resulting dataset is a sequence of binary outcomes for each of the 20 Italian regions over the years. Not all regions are present at all years. We will focus on the HCV data. The maximum number of respondents is 15,401 (average 3866.61), and the maximum of positive tests is 10,875 (average 2578.12). The aim is to investigate the change in HCV over time, i.e., whether year-effects are present in the profiles. Table 3 in the Supplementary Material lists the

overall HCV prevalences for each of the 20 regions. Figure 1 in the Supplementary Material displays the profiles for each of the 20 regions.

3 Methodology

3.1 Standard Pseudo-likelihood

Using Arnold and Strauss (1991), we introduce pseudo-likelihood, the principal idea of which is to replace a numerically challenging joint density by a simpler function assembled from suitable factors.

Consider a sample of size N with repeated measures sequences of length n. Let S be 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 subvector of $\boldsymbol{y}_i$ corresponding to the components of s that are non-zero. The associated joint density is $f_s(\boldsymbol{y}_i^{(s)}; \boldsymbol{\beta})$. 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. This is a general definition and the precise choice depends on inferential goals and the specific form of the model being considered. A common choice is pairwise likelihood where $\delta_s = 1$ for sequences s with exactly two components equal to 1, and zero otherwise. Other choices include the set of full conditionals, ensuring every outcome in a sequence is conditioned upon all other outcomes in the sequence of repeated measures. Details can be found in Molenberghs and Verbeke (2005, Ch. 9, 12, 21, 22, 24, and 25). The log of the pseudo-likelihood is then

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

Adequate regularity conditions have to be invoked to ensure that (1) can be maximized by solving the pseudo-likelihood (score) equations, the latter obtained by differentiating the logarithmic pseudo-likelihood and equating its derivative to zero. These regularity conditions are spelled out in the Supplementary Material. In particular, when the components in (1) result from a combination of marginal and conditional distributions of the original distribution, then a valid pseudo-likelihood function results. In particular, the classical log-likelihood function is found by setting $\delta_s = 1$ if s is the vector consisting solely of ones, and 0 otherwise. Broadly, a pseudo-likelihood is valid if composed of marginal and conditional densities, derived from the full density describing the entire sequence of measurements, thereby allowing for repetition and weighting. More details can be found in Varin (2008), Lindsay (1988), and Joe and Lee (2008). Note that Joe and Lee (2008) use weighting for reasons of efficiency in pairwise likelihood, similar in spirit to Geys, Molenberghs, and Lipsitz (1998), but differently from its use here, which focuses on bias correction when data are incomplete. Another important reference is Cox and Reid (2004).

Let $\boldsymbol{\beta}_0$ the true parameter. Under suitable regularity conditions (Arnold and Strauss 1991, Geys, Molenberghs, and Ryan 1999, Aerts *et al.* 2002), it can be shown that maximizing (1) gives a

consistent, asymptotically normal estimator $\widetilde{\boldsymbol{\beta}}_N$ so that $\sqrt{N}(\tilde{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0)$ converges in distribution to

$$N_p[\mathbf{0}, I_0(\boldsymbol{\beta}_0)^{-1} I_1(\boldsymbol{\beta}_0) I_0(\boldsymbol{\beta}_0)^{-1}]. \tag{2}$$

Precise statements and additional discussion are given in Appendix A of the Supplementary Material section.

3.2 Pseudo-likelihood for Partitioned Samples

3.2.1 Background

Fieuws and Verbeke (2006) and Fieuws *et al.* (2006) proposed a pseudo-likelihood-based method to fit mixed models to high-dimensional longitudinal data. Their method is reviewed in Molenberghs and Verbeke (2005, pp. 470ff). Precisely, when a large number of longitudinal sequences are modeled simultaneously, standard (restricted) maximum likelihood becomes prohibitive. As an alternative, they propose fitting corresponding mixed models to each pair of outcomes. Hence, if there are M longitudinal sequences per subject, $M(M-1)/2$ pairs ensue. The difference with standard pseudo-likelihood, as reviewed in Section 3.1, is that the models are fitted to each pair separately, whereas (1) assembles all contributions into a single pseudo-likelihood function, ensuring that all parameters are estimated only once. In the pairwise approach, the mean parameters governing, for example, the first longitudinal sequence, are estimated $M-1$ times, because the first second is paired with the second, the third, and so on, up to the Mth. Nevertheless, these authors are able to cast their method in the general pseudo-likelihood context. To see this, they first observe that fitting all bivariate models is equivalent to maximizing the function

$$p\ell(\boldsymbol{\theta}) \equiv p\ell(\boldsymbol{y}_{1i}, \boldsymbol{y}_{2i}, \ldots, \boldsymbol{y}_{Mi}|\boldsymbol{\theta}) = \sum_{r<s} \ell(\boldsymbol{y}_{ir}, \boldsymbol{y}_{is}|\boldsymbol{\theta}_{rs}), \tag{3}$$

where $\boldsymbol{y}_{mi}$ is sequence $m = 1, \ldots, M$ for subject $i = 1, \ldots, N$, $\boldsymbol{\theta}$ is the overall parameter vector and $\boldsymbol{\theta}_{rs}$ is the parameter vector pertaining to pair (r, s). To proceed, one temporarily ignores that some of the vectors $\boldsymbol{\theta}_{rs}$ have common elements, i.e., assuming that all vectors $\boldsymbol{\theta}_{\boldsymbol{rs}}$ are completely distinct. In (3), $\boldsymbol{\theta}$ 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 $\boldsymbol{\theta}$.

Evidently, (3), is of the form (1) and hence their pairwise fitting procedure fits within the general framework of pseudo-likelihood. Conveniently, the set of parameters in $\boldsymbol{\theta}_{rs}$ is treated pair-specific, which allows separate maximization of each term in the pseudo log-likelihood function (3).

Because the pairwise approach fits within the pseudo-likelihood framework, an asymptotic multivariate normal distribution for $\widehat{\boldsymbol{\theta}}$ can be derived, using the general pseudo-likelihood theory presented in Section 3.1: $\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \overset{\text{approx.}}{\sim} N(\mathbf{0}, I_0^{-1} I_1 I_0^{-1})$. To pass from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}^*$, 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. Hence, inference for the elements in $\widehat{\boldsymbol{\theta}}^*$ will be based on

$$\sqrt{N}(\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*) = \sqrt{N}(A'\widehat{\boldsymbol{\theta}} - A'\boldsymbol{\theta}) \overset{\text{approx.}}{\sim} N(\mathbf{0}, A' I_0^{-1} I_1 I_0^{-1} A). \tag{4}$$

We will make use of these ideas to apply pseudo-likelihood estimation to partitioned samples, partioned pseudo-likelihood (PPL) for short.

3.2.2 Independent Subsamples

Let us modify the ideas, reviewed in Section 3.2.1, to the case where a given sample, deemed too large, is broken into $m = 1, \ldots, M$ subsamples, each of size n, such that $N = M \cdot n$. While it is possible to let the subsample size vary (with then $N = \sum_{m=1}^{M} n_m$), to avoid cluttering notation we focus on the equal subsample size case. The likelihood for sample m takes the form

$$p\ell_m(\boldsymbol{\theta}_m) = \sum_{i=1}^{n} \ell(\boldsymbol{y}_{mi}|\boldsymbol{\theta}_m), \tag{5}$$

where $\ell(\cdot)$ refers to the likelihood one would consider were the mth subsample the entire set of data. Further, $\boldsymbol{y}_{mi}$ is the ith subject in sample m. Note that all $\boldsymbol{\theta}_m$ are equal to $\boldsymbol{\theta}^*$, the parameter vector of interest. Note that $\boldsymbol{\theta}$ in Section 3.2.1 now takes the form $(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \ldots, \boldsymbol{\theta}^*)$. The overall estimator is naturally defined as:

$$\widehat{\boldsymbol{\theta}}^* = \frac{1}{M} \sum_{m=1}^{M} \widehat{\boldsymbol{\theta}}_m. \tag{6}$$

Owing to the independence of the subsamples, the off-diagonal blocks in the matrices I_0 and I_1 are all zero. Because each of the M contributions to (5) pertains to a genuine likelihood, the diagonal blocks in I_0 and I_1 are identical up to the sign, because

$$H_\theta \equiv I_0(\boldsymbol{\theta}_m, \boldsymbol{\theta}_m) = E\left[\frac{\partial^2 \ell_m(\boldsymbol{\theta}_m)}{\partial \boldsymbol{\theta}_m^T \partial \boldsymbol{\theta}_m}\right] = -E\left[\left(\frac{\partial \ell_m(\boldsymbol{\theta}_m)}{\partial \boldsymbol{\theta}_m}\right)^T . \frac{\partial \ell_m(\boldsymbol{\theta}_m)}{\partial \boldsymbol{\theta}_m}\right] = -I_1(\boldsymbol{\theta}_m, \boldsymbol{\theta}_m).$$

Using the above results, the asymptotic variance (4) for (6) takes the form

$$\text{var}(\widehat{\boldsymbol{\theta}}^*) = \frac{1}{M} H_{\widehat{\theta}}^{-1}. \tag{7}$$

While (7) uses the expected information, one can also employ the observed information matrices, denoted by $\mathcal{H}_{\widehat{\theta},m}$ and derived as the second derivative of the pseudo-likelihood function, for each of the subsamples:

$$\frac{1}{M^2} \sum_{m=1}^{M} \mathcal{H}_{\widehat{\theta},m}^{-1}. \tag{8}$$

These derivations are based upon using

$$A = \frac{1}{M}(I, \ldots, I) \tag{9}$$

in (4), where I is an identity matrix with dimensions equal to the length of the vector $\boldsymbol{\theta}^*$.

3.2.3 Dependent Subsamples

In the above, the overall sample was partitioned into M independent parts, consisting each of n subjects. A different, important area of application is when, say longitudinal or otherwise hierarchical samples would be partitioned by splitting each outcome vector $\boldsymbol{Y}_i$ into M sub-vectors: $(\boldsymbol{Y}_{1i}, \ldots, \boldsymbol{Y}_{Mi})'$, each consisting of n measurements. Three remarks are in order. First, as before, it is perfectly possible to let n vary with subsample, and even from subject to subject, but for ease of exposition we will keep it constant throughout this section. Second, partitioning can be generalized to that of the previous and the current section combined. Third, the difference with Section 3.2.1 is that we do not make multiple use of the same portions within an outcome sequence; should this be desirable, this aspect can then be brought into the picture without trouble. The methodology in the dependent case is entirely similar to that of the dependent case with the sole but important exception that the off-diagonal blocks in I_1 are generally non-zero. Indeed, the (m, m') block of I_1 takes the form

$$I_1(\boldsymbol{\theta}_m, \boldsymbol{\theta}_{m'}) = E\left[\left(\frac{\partial \ell_m(\boldsymbol{\theta}_m)}{\partial \boldsymbol{\theta}_m}\right)^T . \frac{\partial \ell_{m'}(\boldsymbol{\theta}_{m'})}{\partial \boldsymbol{\theta}_{m'}}\right].$$

As a result, the straightforward expressions (7) and (8) no longer apply. Rather, general expression (4) should be used, with A as in (9).

4 Illustrations

In this section, we provide some simple but insightful illustrations in the form of a univariate normal sample, Bernoulli sampling, and a compound-symmetry multivariate normal sample. While a genuine data analysis is relegated to Section 6, these three simple cases have the advantage of establishing three qualitatively distinct situations: (1) **No difference between likelihood (ML) and PPL.** In the normal sample as well as in the Bernoulli sample, parameterized using a probability, there is no difference between either. The same is true for the mean parameter in the compound-symmetry case; (2) **Different estimator, same precision.** When the Bernoulli experiment is characterized by the logit instead, a different estimator emerges, with nevertheless the same asymptotic variance; **Different estimator, precision loss.** ML and PPL differ for both variance components (measurement error, random-intercept variance) in the compound-symmetry case, in terms of both the estimators as well as their asymptotic variance. Each of these three cases will be sketched next, with details relegated to the Supplementary Material, where appropriate.

4.1 A Univariate Normal Sample

Assume $Y_{mi} \sim N(\mu, \sigma^2)$. The PPL takes the form:

$$p\ell(\mu_1, \ldots, \mu_M, \sigma_1^2, \ldots, \sigma_M^2) \propto \sum_{m=1}^{M} \sum_{i=1}^{n} \left\{ -\frac{1}{2}\ln(\sigma_m) - \frac{1}{2}\left(\frac{y_{mi} - \mu_m}{\sigma_m}\right)^2 \right\}. \tag{10}$$

Recall that all μ_m and all σ_m are equal across m, but in the process of deriving the estimator, each subsample is given its own parameter vector. Focusing on the mean parameter, we obtain $\widehat{\mu}_m = \overline{y}_{mi}$. The mean estimator becomes

$$\widetilde{\mu} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{n}\sum_{i=1}^{n} y_{mi} = \widehat{\mu}, \tag{11}$$

where $\widetilde{\mu}$ refers to the PPL estimator, as opposed to the ML estimator $\widehat{\mu}$. The second derivative of (10) is $-n/\sigma_m^2$. Given equality of the parameters, we can write $I_0(\mu_m, \mu_m) = -I_1(\mu_m, \mu_m) = n\sigma^2$ and, with now $A = \frac{1}{M}(1, \ldots, 1)'$, we obtain var$(\widetilde{\mu}) = \sigma^2/(Mn) =$ var$(\widehat{\mu})$. Based upon similar calculations, the asymptotic variance for σ^2 becomes: var$(\widetilde{\sigma}^2) = 2\sigma^4/(Mn) =$ var$(\widehat{\sigma}^2)$. Note that we have treated both estimators independent rather than joint, because μ and σ^2 are independent, a property carrying over onto both I_0 and I_1, so that also in the PPL case these are effectively two scalar-parameter problems.

Hence, as stated earlier, for the univariate normal case, the PPL method produces the same estimators and the same asymptotic variances as the classical one obtained with ML. Given the linearity of all expressions involved, this result is also true for the observed information. When the sample sizes would vary across sub-samples, then (11) has to be weighted appropriate to maintain equality.

4.2 A Univariate Bernoulli Sample

Assume that Y_{mi} is Bernoulli with parameter π or, equivalently, logit $\alpha = \ln[\pi/(1-\pi)]$. The PPL assumes the form:

$$p\ell(\pi_1, \ldots, \pi_M) = \sum_{m=1}^{M}\left[z_m \ln \pi_m + (n - z_m)\ln(1-\pi_m)\right], \tag{12}$$

with $z_m = \sum_{i=1}^{n} y_{mi}$, the number of successes in the mth subsample, for which π_m is the corresponding copy of the success probability π. It then follows that $\widehat{\pi}_m = z_m/n$ and hence

$$\widetilde{\pi} = \frac{1}{Mn}\sum_{m=1}^{M} z_m = \widehat{\pi},$$

where $\widehat{\pi}$ refers to the ML estimator and $\widetilde{\pi}$ is the PPL counterpart. Further

$$E\left(\frac{\partial^2 \ell_m}{\partial \pi_m^2}\right) = -\frac{n}{\pi_m(1-\pi_m)},$$

leading to var$(\widetilde{\pi}) = [\pi(1-\pi)]/(Mn) =$ var$(\widehat{\pi})$. Again, in analogy with the univariate normal case above, there is no difference between the PPL and ML cases.

Switching to the logit parameterization, (12) is replaced with

$$p\ell(\alpha_1, \ldots, \alpha_M) = \sum_{m=1}^{M}\left[z_m\alpha_m - n\ln\left(1 + e^{\alpha_m}\right)\right], \tag{13}$$

leading to the asymptotic variances

$$\widehat{\alpha}_m = \ln\left(\frac{z_m}{n - z_m}\right), \qquad \text{and} \qquad E\left(\frac{\partial^2 \ell_m}{\partial \alpha_m^2}\right) = -n\frac{e^{\alpha_m}}{(1+e^{\alpha_m})^2},$$

and hence, the variance for the logit becomes:

$$\text{var}(\widetilde{\alpha}) = \frac{1}{M^2 n}\sum_{m=1}^{M}\frac{1}{\widehat{\pi}_m(1-\widehat{\pi}_m)}. \tag{14}$$

In case we use the expected value, we obtain:

$$\text{var}(\widetilde{\alpha}) = \frac{1}{Mn}\frac{1}{\pi(1-\pi)},$$

exactly the same as what would be obtained if the sample were not partitioned but rather conventional ML were used. This is interesting, because the point estimator $\widetilde{\alpha} = \frac{1}{M}\sum_{m=1}^{M}\widehat{\alpha}_m$ is different from the ML version $\widehat{\alpha} = \ln[z/(Mn - z)]$, where z is the overall number of successes out of a total of Mn Bernoulli trials.

4.3 Multivariate Normal Compound Symmetry

Consider the compound symmetry model

$$\boldsymbol{Y}_i = (\boldsymbol{Y}_{1i}, \ldots, \boldsymbol{Y}_{Mi})' \sim N(\mathbf{1}\mu, \sigma^2 I + dJ), \tag{15}$$

where $\mathbf{1}$ is a vector of ones of appropriate length, I is the identity matrix, and J is a matrix of ones. The dimensions of $\mathbf{1}$, I, and J will be suppressed from notation when no confusion can arise. We will present the ML and PPL point and precision estimators. Details are relegated to Appendix B in the Supplementary Material.

To simplify notation, let us consider the pseudo-likelihood contribution for a given arbitrary subsample, m, which comes down to the log-likelihood for a sample of N subjects with sequences of length n. Recall that the full sequence length is Mn and that the sequences have been split into M equal portions. The log-likelihood takes the form:

$$\begin{aligned}\ell(\mu_m, \sigma_m^2, d_m) \;=\; & -\frac{1}{2}\sum_{i=1}^{N}\Big\{\ln\left[\sigma^{2n} + n\sigma^{2(n-1)}d_m\right] \\ & +(\boldsymbol{Y}_{mi} - \mu_m\mathbf{1})'\frac{1}{\sigma_m^2}\left(I - \frac{d_m}{\sigma_m^2 + nd_m}\right)(\boldsymbol{Y}_{mi} - \mu_m\mathbf{1})\Big\}.\end{aligned} \tag{16}$$

As in the univariate normal case, the mean μ_m on the one hand and the variance components σ_m^2 and d_m on the other can be treated separately, owing to their functional and statistical independence.

Considering μ_m first and setting the first derivative of (16) equal to zero leads to the conventional:

$$\widehat{\mu}_m = \frac{1}{Nn}\sum_{i=1}^{N}\sum_{j=1}^{n} Y_{ij}. \tag{17}$$

From (17) we immediately deduce:

$$\widetilde{\mu} = \frac{1}{Nnm} \sum_{m=1}^{M} \sum_{i=1}^{N} \sum_{j=1}^{n} Y_{mij}. \tag{18}$$

which is the overall mean. Evidently, this is also the estimator obtained, should one analyze the sample as a whole, i.e., $\widetilde{\mu} = \widehat{\mu}$. Similarly, turning to the variance components, it follows that

$$\widehat{\sigma}_m^2 = \frac{1}{Nn(n-1)} \left(n \sum_{i=1}^{N} \boldsymbol{Z}'_{mi} \boldsymbol{Z}_{mi} - \sum_{i=1}^{N} \boldsymbol{Z}'_{mi} J_n \boldsymbol{Z}_{mi} \right), \tag{19}$$

$$\widehat{d}_m = \frac{1}{Nn(n-1)} \left(\sum_{i=1}^{N} \boldsymbol{Z}'_{mi} J_n \boldsymbol{Z}_{mi} - \sum_{i=1}^{N} \boldsymbol{Z}'_{mi} \boldsymbol{Z}_{mi} \right), \tag{20}$$

where $\boldsymbol{Z}_{mi} = (\boldsymbol{Y}_{mi} - \mu \mathbf{1}_n)$ and J_n is an $n \times n$ matrix of ones. From (19) and (20) it follows that:

$$\widetilde{\sigma}^2 = \frac{1}{MNn(n-1)} \left(n \sum_{m=1}^{M} \sum_{i=1}^{N} \boldsymbol{Z}'_{mi} \boldsymbol{Z}_{mi} - \sum_{m=1}^{M} \sum_{i=1}^{N} \boldsymbol{Z}'_{mi} J_n \boldsymbol{Z}_{mi} \right), \tag{21}$$

$$\widetilde{d} = \frac{1}{MNn(n-1)} \left(\sum_{m=1}^{M} \sum_{i=1}^{N} \boldsymbol{Z}'_{mi} J_n \boldsymbol{Z}_{mi} - \sum_{m=1}^{M} \sum_{i=1}^{N} \boldsymbol{Z}'_{mi} \boldsymbol{Z}_{mi} \right). \tag{22}$$

Evaluating (19) and (20) for the case that the sample is analyzed at once, but then evidently for sequences of length Mn, leads to:

$$\widehat{\sigma}^2 = \frac{1}{MNn(Mn-1)} \left(Mn \sum_{i=1}^{N} \boldsymbol{Z}'_i \boldsymbol{Z}_i - \sum_{i=1}^{N} \boldsymbol{Z}'_i J_{Mn} \boldsymbol{Z}_i \right), \tag{23}$$

$$\widehat{d} = \frac{1}{MNn(Mn-1)} \left(\sum_{i=1}^{N} \boldsymbol{Z}'_i J_{Mn} \boldsymbol{Z}_i - \sum_{i=1}^{N} \boldsymbol{Z}'_m \boldsymbol{Z}_m \right), \tag{24}$$

where $\boldsymbol{Z}_i = (\boldsymbol{Y}_i - \mu \mathbf{1}_{Mn})$ and J_{Mn} is an $Mn \times Mn$ matrix of ones. Thus, $\widetilde{\sigma}^2 \neq \widehat{\sigma}^2$ and $\widetilde{d} \neq \widehat{d}$.

Consider precision estimation. For this, the (block-diagonal) matrix I_0 and the (non-block-diagonal) I_1 need to be derived. Details are found in Supplementary Material B. As stated earlier, the mean can be treated separately from the variance components. Turning to μ, given that $\widetilde{\mu} = \widehat{\mu}$, it is not surprising that both asymptotic variances are equal:

$$\text{var}(\widetilde{\mu}) = \text{var}(\widehat{\mu}) = \frac{\sigma^2 + Mnd}{MNm}. \tag{25}$$

Focusing on the variance components, we find:

$$\text{var}(\widetilde{\sigma}^2) = \frac{2\sigma^4}{MNn - MN}, \tag{26}$$

$$\text{var}(\widehat{\sigma}^2) = \frac{2\sigma^4}{MNn - N}, \tag{27}$$

$$\text{var}(\widetilde{d}) = \frac{2}{MNn}\left(\frac{\sigma^4}{Mn-1} + 2d\sigma^2 + Mnd^2\right), \tag{28}$$

$$\text{var}(\widehat{d}) = \frac{2}{MNn}\left(\frac{\sigma^4}{n-1} + 2d\sigma^2 + Mnd^2\right). \tag{29}$$

5 Efficiency

5.1 Asymptotic Relative Efficiency

In Section 4, three distinct illustrations were used. In the univariate normal and univariate Bernoulli cases (probability and logit scale), and for the mean parameter in the compound-symmetry case, the asymptotic relative efficiency (ARE), defined as the variance ratio of the ML over the PPL estimator, when the overall sample size tends to infinity, equals one. For the logit case, this may not be the case for small samples, a point to which we will in Section 5.2. This is no longer the case for the variance components in the compound-symmetry case. Indeed, for σ^2, we find:

$$\text{ARE}(\sigma^2) = \frac{Mn-M}{Mn-1}. \tag{30}$$

Three remarks are in place. First, if $M = 1$, there is no partitioning, and evidently ARE reduces to 1. Second, if $n = 1$, the ARE ánd the variances are problematic. In particular, if $n = 1$ and $M > 1$, the ML variance (27) is well defined, whereas PPL variance (26) is undefined, because for the PPL each of the subsamples is univariate, from which σ^2 and d cannot be disentangled. Third, if n approaches infinity, with M bounded, the ARE approaches one. The above remarks also hold for the random-intercepts variance d, even though the ARE is slightly more tedious:

$$\text{ARE}(d) = \frac{\frac{1}{Mn-1} + 2\left(\frac{d}{\sigma^2}\right) + Mn\left(\frac{d}{\sigma^2}\right)^2}{\frac{1}{n-1} + 2\left(\frac{d}{\sigma^2}\right) + Mn\left(\frac{d}{\sigma^2}\right)^2} = \frac{\frac{(1-\rho)^2}{Mn-1} + 2\rho(1-\rho) + Mn\rho^2}{\frac{(1-\rho)^2}{n-1} + 2\rho(1-\rho) + Mn\rho^2}, \tag{31}$$

where $\rho = d/(\sigma^2 + d)$, the intraclass correlation. We will study these expressions in what follows.

The ARE for σ^2 has a simple structure, and nicely shows the efficiency loss related to increasing the number of partitions of the original sample. When the sequence length per component, n, is large relative to the number of sub-samples, efficiency loss will be modest. For example, with $M = 5$ pieces and $n = 20$, the ARE is around 95%. The sequence length is realistic because, for very sort sequences, the need to revert to sub-divided samples is usually not present.

ARE expression (31), shows that the two key quantities are M, the number of subsamples in the partition, on the one hand, and the intraclass correlation ρ on the other. Evidently, ARE(d) approaches 1 when M approaches 1; for large M, there is a 'partitioning penalty.' When ρ is small, the penalty is large, because then (31) approaches ARE$(d|\rho = 0) = (n-1)/(Mn-1)$. For example, when the correlation is close to 0 and $M = 5$, estimating the random-intercept variance is effectuated at a

mere 20% of the ML efficiency. This is qualitatively among the worst scenarios. Of course, it actually corresponds to the situation where the component d is unimportant in the first place. Luckily, on the other hand, the penalty vanishes for d large, relative to σ^2, because ARE$(d|\rho = 1) = 1$.

5.2 Small-sample Relative Efficiency for the Logit Parameter

To assess the small-sample impact when there is asymptotic equivalence, but not equality of the estimators, we conduct limited simulations for a logit estimated from univariate Bernoulli samples, as in Section 4.2. We compare partition-based variance (14) with its single-sample counterpart, when the true logit is 0 (probability of 0.5). The number of partition components is $M \in \{1, 5, 10, 20, 50\}$ and the sample size per component is $n \in \{20, 50, 100, 500\}$. Results are in Table 4 of the Supplementary Material. We see that the SSRE is very high. Given that the ARE has been shown to equal 1, it is not a surprise that the SSRE is very close to 1 as well when n is large relative to M. Only when the number of partition components is large relative to the sample size per partition component, does the SSRE shrink somewhat, but the worst case remains at a comfortably high 90%.

6 Analysis of Case Study

We analyze the data of Section 2. For the ith region in year j, let Z_{ij} be the number of independent reported cases of HCV out of n_{ij}. Assume a binomial model with success probability:

$$\text{logit}(\pi_{ij}) = \alpha_0 + \sum_{j=1}^{8} \alpha_j T_{ij} + b_i, \tag{32}$$

with the indicator for year defined as $T_{ij} = 1$ if the year during which the measurement is taken equals $k = 1, \ldots, 9$, and 0 otherwise. The study spans 9 years, with the last year as reference. The model is of the generalized linear mixed model type (Molenberghs and Verbeke 2005), with a region-specific effect $b_i \sim N(0, \sigma^2)$. To study the effect of parameterization on efficiency, consider the equivalent formulation:

$$\text{logit}(\pi_{ij}) = \sum_{j=1}^{9} \beta_j T_{ij} + b_i. \tag{33}$$

For each of the parameterizations, we consider three approaches.

In the first analysis, the data are analyzed as if measurements are independent (i.e., omitting the random effect). Results are summarized in Table 5 of the Supplementary Material. This assumption is unrealistic; the analysis is included to make a few points about the methodology only. It confirm the results for the (independent) Bernoulli case. The difference with Section 4.2 is that here we assume logistic regression models. Not unexpectedly, the parameter estimates and standard errors are the same, up to four decimal places for all of the values of $M = 1, 2, 5, 10, 15$. As a second analysis, the data are split into M independent subsamples, for $M = 1, 2, 4$. Results are in Table 1.

Table 1: *HCV data. Parameter estimates and standard errors for independent partitioning. For $M = 2, 4, 15$ the proper empirically corrected standard errors are followed by their (inappropriate) purely model-based counterparts.*

Par.	$M = 1$ (ML)	$M = 2$	$M = 4$
		Parameterization (32)	
α_0	0.592(0.112)	0.598(0.111)	0.593(0.108)
α_1	0.223(0.011)	0.213(0.011)	0.243(0.012)
α_2	0.209(0.011)	0.202(0.011)	0.215(0.011)
α_3	0.288(0.011)	0.287(0.011)	0.300(0.012)
α_4	0.179(0.011)	0.175(0.011)	0.170(0.011)
α_5	0.106(0.011)	0.099(0.011)	0.095(0.011)
α_6	0.114(0.011)	0.104(0.011)	0.106(0.011)
α_7	0.072(0.011)	0.062(0.011)	0.068(0.011)
α_8	-0.037(0.011)	-0.043(0.011)	-0.049(0.011)
σ	0.501(0.079)	0.493(0.078)	0.459(0.076)
		Parameterization (33)	
β_1	0.815(0.113)	0.811(0.111)	0.836(0.108)
β_2	0.801(0.113)	0.800(0.111)	0.808(0.108)
β_3	0.880(0.113)	0.886(0.111)	0.894(0.108)
β_4	0.771(0.113)	0.773(0.111)	0.763(0.108)
β_5	0.698(0.112)	0.697(0.111)	0.689(0.108)
β_6	0.706(0.112)	0.702(0.111)	0.699(0.108)
β_7	0.664(0.112)	0.660(0.111)	0.662(0.108)
β_8	0.555(0.113)	0.556(0.111)	0.544(0.108)
β_9	0.592(0.112)	0.598(0.111)	0.593(0.108)
σ	0.501(0.079)	0.493(0.078)	0.459(0.076)

In line with Section 3.2.2, results are virtually identical across values of M. In the final analysis, dependent samples are created by sub-dividing the sequences into $M = 1, 2, 5, 10, 15$ parts. Results are in Table 2. Several observations can be made. First, sub-samples are not all of equal size. This is a trivial extension of the methodology presented so far. For convenience, we ensured that portions were roughly equal, because exact equality could not be achieved. Now, weighting matrix A to take the form (9) may not be fully optimal; although this does not affect the validity of the method, efficiency may be affected slightly. Second, the said efficiency in the dependent case is more affected when parameterization (32) is used, for the parameters α_1–α_8, than with parameterization

Table 2: *HCV data. Parameter estimates and standard errors for dependent partitioning. For $M = 1, 2, 5, 10, 15$, the proper empirically corrected standard errors are followed by their (inappropriate) purely model-based counterparts.*

Par.	$M = 1$ (ML)	$M = 2$	$M = 5$	$M = 10$	$M = 15$
			Parameterization (32)		
α_0	0.592(0.112)	0.592(0.119;0.080)	0.592(0.119)	0.593(0.119)	0.593(0.119;0.030)
α_1	0.223(0.011)	0.223(0.077;0.011)	0.223(0.077)	0.223(0.077)	0.223(0.077;0.011)
α_2	0.209(0.011)	0.209(0.070;0.011)	0.209(0.070)	0.209(0.070)	0.209(0.070;0.011)
α_3	0.288(0.011)	0.288(0.063;0.011)	0.288(0.063)	0.288(0.063)	0.288(0.063;0.011)
α_4	0.179(0.011)	0.179(0.061;0.011)	0.179(0.061)	0.179(0.061)	0.179(0.061;0.011)
α_5	0.106(0.011)	0.106(0.055;0.011)	0.106(0.055)	0.106(0.055)	0.106(0.055;0.011)
α_6	0.114(0.011)	0.114(0.051;0.011)	0.114(0.051)	0.114(0.051)	0.114(0.051;0.011)
α_7	0.072(0.011)	0.072(0.054;0.011)	0.072(0.054)	0.072(0.054)	0.072(0.054;0.011)
α_8	-0.037(0.011)	-0.037(0.033;0.011)	-0.037(0.033)	-0.037(0.033)	-0.037(0.033;0.011)
σ	0.501(0.079)	0.501(0.079;0.056)	0.500(0.079)	0.498(0.079)	0.496(0.079;0.021)
			Parameterization (33)		
β_1	0.815(0.113)	0.815(0.096;0.080)	0.815(0.096)	0.815(0.096)	0.816(0.096;0.030)
β_2	0.801(0.113)	0.801(0.116;0.080)	0.802(0.116)	0.802(0.116)	0.802(0.116;0.030)
β_3	0.880(0.113)	0.880(0.127;0.080)	0.880(0.127)	0.881(0.127)	0.881(0.127;0.030)
β_4	0.771(0.113)	0.771(0.112;0.080)	0.772(0.112)	0.771(0.113)	0.771(0.113;0.030)
β_5	0.698(0.112)	0.698(0.121;0.080)	0.699(0.121)	0.699(0.121)	0.699(0.121;0.030)
β_6	0.706(0.112)	0.706(0.119;0.080)	0.706(0.119)	0.707(0.119)	0.707(0.119;0.030)
β_7	0.664(0.112)	0.664(0.131;0.080)	0.665(0.131)	0.666(0.131)	0.666(0.131;0.030)
β_8	0.555(0.113)	0.555(0.118;0.080)	0.556(0.118)	0.556(0.118)	0.557(0.118;0.030)
β_9	0.592(0.112)	0.592(0.119;0.080)	0.593(0.119)	0.593(0.119)	0.593(0.119;0.030)
σ	0.501(0.079)	0.501(0.079;0.079)	0.500(0.079)	0.498(0.079)	0.496(0.079;0.021)

(33), with efficiency largely unaffected. Also, under (32), the overall intercept parameter α_0 is unaffected by partitioning. Note that parameters β_1–β_9 have the meaning of an intercept parameter as well, be it for a specific year, whereas α_1–α_8 are contrasts between a given year and the last one. Hence, in some parameterizations, the efficiency loss may be spread out relatively evenly across parameters, whereas others may segregate parameters that are largely unaffected from others that are more severely impacted. Even for strongly affected parameters, α_1–α_8, it seems to be more a consequence of partitioning as such, rather than of the number of components, M, in the partition. This is not general; counterexamples are given by (31) and (30), which are smooth, continuous

functions of M. Arguably, efficiency will be more at risk for parameters that have a within-subject meaning. This agrees with the fact that, in the compound-symmetry case (4.3), μ is unaffected by partitioning, whereas the variance components, especially d, can be affected considerably stronger. Third, it is unwise to replace the empirically corrected standard errors, based on $I_0^{-1} I_1 I_0^{-1}$ with their purely model-based counterparts, using I_0^{-1} only, as can be seen when comparing the appropriate empirically corrected standard errors with their model-based counterparts, in Tables 1 and 2. This is in line with other uses of pseudo-likelihood (Molenberghs and Verbeke 2005).

7 Concluding Remarks

We have presented a convenient, simple, pseudo-likelihood based method to partition large data sets to facilitate estimation. Partitions into dependent as well as independent subsamples have been studied. In the independent subsample case, full efficiency can be reached. This is not always so for dependent samples, but in important realistic settings high to very high efficiency can be obtained. Especially in such cases, it is recommendable to ensure the sub-sample size is not too small.

Needless to say that large sets of data are extremely common in current-day empirical research, including but not limited to large surveys, microarray experiments, consumer databases, etc. When confronted with choice, independent sub-samples are preferred, because of efficiency. Dependent sub-samples become indispensable whenever analyzing the sequences in full is practically prohibitive. Finally, whether splitting in sub-sampling is done and, if so, in how many components, strongly depends on what is practically and numerically feasible. Indeed, given the potential loss in efficiency, fewer components is better than more.

Apart from illustrative examples, real-data analysis has been employed to underscore the practical power of the method. While the illustrative examples allow for closed-form derivation of point and precision estimators, this is relevant from theoretical and illustrative perspectives only. In particular, it is is clear that efficiency loss in the dependent case is relative to a particular parameter, which is why the mean, variance, and correlation parameters all exhibit different behavior. Therefore, the data analyst is advised to study efficiency loss, perhaps using a simulation study designed after the real application at hand. Of course, there are cases where maximum likelihood is not even feasible, underscoring our method's use.

For data analysis, such as that based on the generalized linear mixed model in Section 6, closed-form derivations are neither feasible (because even conventional ML defies closed forms) nor necessary (because our methodology is presented in generic terms and can be applied whenever score-vector contributions are available, either in analytical or numerical form). SAS code used for the purpose of this article are available from the authors' web site.

Acknowledgment

We acknowledge financial support from IAP research Network P6/03 of the Belgian Government (Belgian Science Policy).

References

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

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

Cox, D. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**, 729–737.

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.

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. and Ryan, L. (1999). Pseudo-likelihood modelling of multivariate outcomes in developmental toxicology. *Journal of the American Statistical Association* **94**, 34–745.

Joe, H. and Lee, Y. (2008). On weighting of bivariate margins in pairwise likelihood. *Journal of Multivariate Analysis* **100**, 670–685.

Lindsay, B.G. (1988). Composite likelihood methods. *Contemporary Mathematics* **80**, 221–239.

Mathei, C., Shkedy, Z., Denis, B., Kabali, C., Aerts, M., Molenberghs, G., Van Damme, P., and Buntinx, F. (2006). Evidence for a substantial role of sharing of injection paraphernalia other than syringes/needles to the spread of hepatitis C among injecting drug users. *Journal of Viral Hepatitis*, **13**, 560–570.

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

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis* **92**, 1–28.

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