# Pseudo-likelihood methodology for hierarchical count data

Peer-reviewed author version

KALEMA, George & MOLENBERGHS, Geert (2012) Pseudo-likelihood methodology for hierarchical count data. In: COMMUNICATIONS IN STATISTICS-THEORY AND METHODS 43(22), p. 4790-4805.

DOI: 10.1080/03610926.2012.744053
Handle: http://hdl.handle.net/1942/14834

# PSEUDO-LIKELIHOOD METHODOLOGY FOR HIERARCHICAL COUNT DATA

**George Kalema**[1], **Geert Molenberghs**[1,2]

[1] *I-Biostat, Katholieke Universiteit Leuven, Kapucijnenvoer 35, B3000 Leuven, Belgium*

[2] *I-Biostat, Universiteit Hasselt, Agoralaan 1, B3590 Diepenbeek, Belgium*

**Abstract**

Generalized Estimating Equations (GEE) are a widespread tool for modeling correlated data, based on properly formulating a marginal regression function, combined with working assumptions about the correlation function. Should interest be placed in addition on the correlation function, then, apart from second-order GEE, pseudo-likelihood (PL) also provides an attractive alternative, especially in its pairwise form, where the covariance between each pair of the response vector is modeled as well. An elegant PL approach is formulated in this paper, based on a flexible bivariate Poisson model. The performance of the PL-method is studied, relative to GEE, using simulations. Data on repeated counts of epileptic seizures in a two-arm clinical trial are analyzed. A macro has been developed by the authors and made available on their web pages.

*Key words and phrases:* Bivariate Poisson distribution; Correlated data; Generalized estimating equations; Pseudo-likelihood.

## 1 Introduction

Count data collected repeatedly over time for the same subject are commonly encountered in scientific research. When collected only once per subject or at one time point, one usually assumes the data to be generated from a univariate Poisson distribution. Contemporary studies frequently aim at describing the

evolution of subjects over time or observing more than one response from a single subject. Assuming a univariate Poisson distribution as the parent distribution of such data would ignore correlation and lead to erroneous inferences.

A lot of research has been done to account for correlation in count data. Breslow and Clayton (1993) and Wolfinger and O'Connell (1993) extended the generalized linear modeling (GLM) framework to the so-called generalized linear mixed model (GLMM) in which the correlation is accounted for by use of random effects. Molenberghs et al. (2007) (see also Molenberghs et al. 2010) propose a joint model for clustering and over-dispersion through two separate sets of random effects.

Extensions of the univariate Poisson model to a multivariate version have also been proposed. This has the advantage of a gain in efficiency as long as the model is correctly specified. However, use of a so-called Multivariate Poisson (MP) model is constrained by the complexity of the probability function to be calculated. This is because it involves summations which may increase the computational burden with increase in the number of measurements per subject and/or sample size. Karlis (2003) uses the Expectation-Maximization (EM) algorithm to derive a MP distribution via a multivariate reduction technique. Karlis and Ntzoufras (2003) model sports data using a bivariate Poisson distribution. Kocherlakota and Kocherlakota (2001) apply a bivariate Poisson distribution to longitudinal data but with only two time points. In this paper, we propose a pseudo-likelihood, taking the form of pairwise likelihood, to drastically simplify computational burden while retaining sufficiently high statistical efficiency. For each pair, a bivariate Poisson distribution is specified hence capturing the association between the two measurements. We restrict attention to each subject having at least 2 measurements recorded. We compare our proposal to Generalizing Estimating Equations (GEE, Liang and Zeger 1986) based on a simulation with varying sample sizes ($K$) and number of measurements per subject $i$ ($n_i$). Our proposal allows for $n_i$ to differ between subjects but we assign equal $n_i$ to all subjects in the simulations. We quantify the behavior of the two methods in terms of mean square error (MSE), variance, and the absolute bias of the estimators. Two cases worth investigating are (a) when there is no association in the data and, (b) where there exists association or when data is collected

repeatedly per subject. In Section 2, an overview of GEE, the general idea of pseudo-likelihood and our proposition are given. Section 3 outlines the set-up as well as the results of the simulation study while an application of the proposal to a clinical trial study in epileptic seizures is presented in Section 4.

# 2 Methodology

Inference in a good number of longitudinal studies is primarily based on marginal parameters. Using classical maximum likelihood methodology then necessitates the full specification of the joint distribution for $\mathbf{Y}_i$. In the context of discrete data, one needs to specify the first-order moments as well as all higher-order moments (Molenberghs and Verbeke 2005) which often is computationally restrictive for high-dimensional vectors of correlated data. With primary interest placed on the marginal parameters, however, tools like GEE and pseudo-likelihood (PL, Arnold and Strauss 1991, Le Cessie van Van Houwelingen 1994, Zhao and Joe 2005, Molenberghs and Verbeke 2005, Yi, Zeng, and Cook 2011) have been proposed and implemented in statistical software. These two tools still allow for within-subject dependence but yet are computationally more practical relative to full likelihood.

Assume that there are $K$ independent subjects in a study with subject $i$ having a measurement $Y_{ij}$, $i = 1, \ldots, K$, $j = 1, \ldots, n_i$ and a corresponding $q \times p$ known design matrix $\mathbf{X}_i$. Denote the responses of subject $i$ at any given pair of time points, $s$ and $t$ as $Y_{is}$ and $Y_{it}$, respectively, $1 \leq s < t \leq n_i$.

## 2.1 Generalized Estimating Equations

GEE makes no distributional assumptions apart from the specification of the mean function $\boldsymbol{\mu}_i = \exp(\mathbf{X}_i\boldsymbol{\beta})$ for models with the log link, the variance function $V_i = \left(A_i^{1/2} R_i(\alpha) A_i^{1/2}\right)^{-1}$ where $A_i$ is an $n_i \times n_i$ diagonal matrix with $var(\mu_{ij})$ as the $j^{th}$ diagonal element, and $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is an $n_i \times n_i$ (perhaps incorrect) working correlation matrix to model the dependence between within-subject observations expressed in terms of $\boldsymbol{\alpha}$ a vector of unknown parameters. Liang and Zeger (1986)

solve the estimating equation

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'} V_i(\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{1}$$

where $\boldsymbol{\mu}_i = E(\mathbf{Y}_i)$, $\boldsymbol{\beta}$ is a $p$-dimensional vector of unknown regression parameters. The correlation between measurements can be assumed as, for example, $\text{Corr}(Y_{is}, Y_{it}) = 0$ for independence, $\text{Corr}(Y_{is}, Y_{it}) = \alpha$ for exchangeability or $\text{Corr}(Y_{is}, Y_{it}) = \alpha_{st}$ for unstructured working assumptions ($s \neq t$). The solution to (1) is consistent and asymptotically normally distributed with mean $\boldsymbol{\beta}$ and an asymptotic variance-covariance matrix

$$\text{Var}(\hat{\boldsymbol{\beta}}) = I_0^{-1} I_1 I_0^{-1}, \tag{2}$$

also referred to as the sandwich estimator, where

$$I_0 = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'}, \tag{3a}$$

$$I_1 = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \text{Var}(\mathbf{Y}_i) V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'}, \tag{3b}$$

as long as the marginal mean is correctly specified. Consistent parameter estimates and standard errors are obtained even with miss-specification of the working assumption. Correct specification of the working correlation matrix results in improved efficiency of the parameter estimates while severe miss-specification may compromise efficiency. We refer to Molenberghs and Verbeke (2005) and related references therein for further details on GEE.

GEE however falls short when scientific interest is in drawing inferences on the association parameters or if the estimated correlation matrix is not positive definite leading to a breakdown in the iterative procedure (Sun, Shults, and Leonard 2009). Correct estimation of the correlation also improves efficiency of the estimated regression parameters (Wang and Carey 2004). Some alternative approaches have been proposed like [a] Second-order GEE in which the marginal mean parameters are simultaneously estimated with the marginal correlation parameters (Zhao and Prentice 1990; Liang, Zeger and Qaqish 1992) and, [b] the careful estimation of the correlation parameters in GEE using Quasi Least

Squares, developed in a suite of papers by Chaganty (1997), Shults and Chaganty (1998), and Chaganty and Shults (1999). See also Wang and Carey (2004) and Sun, Shults, and Leonard (2009) for more on estimating the correlation in the framework of GEE.

### 2.2 General Form of Pseudo-likelihood

In likelihood-based modeling frameworks, the marginal (log)likelihood is usually maximized to estimate the unknown parameters. For continuous longitudinal data, the marginal distribution and therefore the marginal (log)likelihood involves a product of the normal distributions for the data and the random effects resulting in a normal distribution as the marginal distribution. This presents no computational challenges and has been widely implemented in statistical software packages like SAS. For non-normal data, on the other hand, specification of the full likelihood can be very prohibitive computationally when measurement sequences are of moderate to large length (Molenberghs and Verbeke 2005). Rather than specifying the full likelihood, the idea of pseudo-likelihood, or composite likelihood (Arnold and Strauss 1991, Le Cessie and van Houwelingen 1991, Geys, Molenberghs, and Ryan 1999, Aerts et al. 2002, Zhao and Joe 2005, Molenberghs and Verbeke 2005, Yi, Zeng, and Cook 2011) is to specify, for example, all univariate densities, or all pairwise densities over the set of all possible pairs within a sequence of repeated measures in place of the full likelihood. In the case of pairwise densities, the likelihood contribution $f(y_{i1}, \ldots, y_{in_i})$ of subject $i$ to the full likelihood is substituted with a product of $f(y_{is}, y_{it})$. For example, when $n_i = 3$, $f(y_{i1}, y_{i2}, y_{i3})$ is replaced by $f(y_{i1}, y_{i2}) \times f(y_{i1}, y_{i3}) \times f(y_{i2}, y_{i3})$ and the corresponding log-likelihood $\log f(y_{i1}, y_{i2}, y_{i3})$ is replaced by $\log f(y_{i1}, y_{i2}) + \log f(y_{i1}, y_{i3}) + \log f(y_{i2}, y_{i3})$. In the general case of $n_i$ measurements per subject $i$, the contribution of subject $i$ to the log pseudo-likelihood is $p\ell_i = \sum_{1 \le s < t \le n_i} \log f(y_{is}, y_{it})$ and the marginal log-pseudo-likelihood is given by

$$p\ell(\boldsymbol{\lambda}|\mathbf{Y}) = \sum_{i=1}^{K} \sum_{s<t} \log f(y_{is}, y_{it}), \tag{4}$$

where $\boldsymbol{\lambda}$ contains the unknown parameters estimated by setting the first derivative of (4) equal to zero. With correct model specification, consistent and nor-

mally distributed estimators are obtained (Molenberghs and Verbeke 2005), the variance-covariance matrix calculated using a sandwich estimator similar to that of GEE in (2).

Regularity conditions have to be invoked to ensure that (4) can be maximized by solving the pseudo-likelihood (score) equations. These can be found, for example, in Molenberghs *et al* (2011). Importantly, because the components in (4) are derived from marginalizing the original distribution, a valid pseudo-likelihood function results. Details can be found in Joe and Lee (2008), who use weighting for reasons of efficiency in pairwise likelihood. **Let $\boldsymbol{\lambda}_0$ be** the true parameter. Under the aforementioned regularity conditions, maximizing (4) produces a consistent and asymptotically normal estimator $\widetilde{\boldsymbol{\lambda}}_0$ so that $\sqrt{N}(\tilde{\boldsymbol{\lambda}}_N - \boldsymbol{\lambda}_0)$ converges in distribution to $N_p[\mathbf{0}, I_0(\boldsymbol{\lambda}_0)^{-1} I_1(\boldsymbol{\lambda}_0) I_0(\boldsymbol{\lambda}_0)^{-1}]$. **The regularity conditions, as well as explicit forms for $I_0(\boldsymbol{\lambda}_0)$ and $I_1(\boldsymbol{\lambda}_0)$, are provided in Appendix A.**

Troxel et al. (1998) used the product of all univariate distributions as an approximation to the full-likelihood. This significantly reduces the computational burden encountered in the full-likelihood approach yet still results in asymptotically unbiased estimators of the regression parameters. However, specifying univariate distributions for longitudinal data is based upon the unrealistic working assumption of no dependence between the several responses within a subject and may lead to highly inefficiently estimated regression parameters (Parzen et al. 2007). Specifying the bivariate distribution for all the pairs of the responses from each subject may be a better approach. This has been used by Parzen et al. (2007) for longitudinal binary data with non-ignorable non-monotone missingness. We apply the approach to hierarchical count data in the context of marginal models.

### 2.3 A Model for Hierarchical Count Data

Assuming that $W_k$ are independent Poisson random variables with means $\theta_k$, $k = s, t$ or $st$. The random variables $Y_{is} = (W_{is} + W_{ist})$ and $Y_{it} = (W_{it} + W_{ist})$ then follow a bivariate Poisson distribution, i.e., $(Y_{is}, Y_{it}) \sim BP(\theta_{is}, \theta_{it}, \theta_{ist})$

given by

$$f(y_{is}, y_{it}) = e^{-(\theta_{is}+\theta_{it}+\theta_{ist})} \frac{\theta_{is}^{y_{is}}}{y_{is}!} \frac{\theta_{it}^{y_{it}}}{y_{it}!} \sum_{k=0}^{\min(y_{is},y_{it})} \binom{y_{is}}{k} \binom{y_{it}}{k} k! \left(\frac{\theta_{ist}}{\theta_{is}\theta_{it}}\right)^k. \quad (5)$$

Let $\theta_{is}^{\star} = \theta_{is} + \theta_{ist}$ and $\theta_{it}^{\star} = \theta_{it} + \theta_{ist}$ where $\log(\theta_{is}) = X_{is}\boldsymbol{\beta}$ and $\log(\theta_{it}) = X_{it}\boldsymbol{\beta}$. Marginally, $Y_{is} \sim \text{Poisson}(\theta_{is}^{\star})$, $Y_{it} \sim \text{Poisson}(\theta_{it}^{\star})$ and $\theta_{ist}$ is the covariance between subsequent pairs of the random variables $Y_{is}$ and $Y_{it}$. The marginal log pseudo-likelihood takes the form (4). Estimation of the parameters in $\boldsymbol{\lambda} = (\boldsymbol{\beta}, \theta_{ist})^T$ is done in SAS/IML® using the Newton-Raphson (NR) algorithm; a macro has been written to this effect. See Appendix B for the gradient and Hessian functions of the log PL function in equation (4), with respect to the unknown parameters in $\boldsymbol{\lambda}$, which are supplied to the NR optimization step.

Note that we have formulated a bivariate model only, even though our aim is to analyze hierarchical data with more than two repetitions. Fortunately, the use of GEE and PL methodology obviates the need to explicitly specify the higher-order joint distributions. We assume the covariance ($\theta_{ist}$) to be the same for all subjects and pairs ($=\theta_{st}$) in this paper. This bears resemblance **to** an exchangeable correlation structure, but one must remember that, because the variance depends on the mean, the corresponding correlations will fluctuate with the mean, even though the covariances may be constant. The exception is when the mean is constant as well; in that case a classical exchangeable correlation matrix will follow. This assumption of equal covariance term can however be relaxed.

## 3 Simulation Study

Simulations have been done to compare the performance of GEE and our proposed pseudo-likelihood approach in the cases of both correlated and independent outcomes. We study the effect of varying sample sizes and number of measurements per subject for GEE with an exchangeable working correlation structure in comparison to pseudo-likelihood, based on 1000 simulations. The absolute bias, MSE, and the percent samples for which convergence has been reached, quantify the behavior of the two methods.

## 3.1 Design of Simulation Study

### 3.1.1 Simulation of Independent Data

We generated data for $K = 10, 100, 1000, 10000$ subjects, assuming the following model:

$$\mu_{ij} = \exp(\beta_0 + \beta_1 * \text{trt}_i + \beta_2 * \text{time}_{ij} + \beta_3 * \text{trt}_i * \text{time}_{ij}), \tag{6a}$$

$$Y_{ij} \sim \text{Poisson}(\mu_{ij}), \tag{6b}$$

for subjects $i = 1, \ldots, n$ and measurements $j = 1, \ldots, n_i$. The subjects are equally distributed across the two treatment groups ($\text{trt}_i = 0$ or 1) and $\text{time}_{ij}$ is the ordering of the $j^{th}$ measurement within subject $i$. Further, $n_i$ was fixed to values of 2, 4, 8, and 16 for all subjects within a given run for simulation purposes, even though our methods allows for varying cluster sizes. The regression parameters were specified as $\beta_0 = 1.4531$, $\beta_1 = -0.1869$, $\beta_2 = -0.0328$, and $\beta_3 = 0.0195$.

### 3.1.2 Simulation of Dependent Data

To generate dependent data, a subject-specific intercept $b_i$ is introduced to equation (6a), changing it to

$$\mu_{ij} = \exp\left(\beta_0 + b_i + \beta_1 * \text{trt}_i + \beta_2 * \text{time}_i + \beta_3 * \text{trt} * \text{time}_{ij}\right). \tag{7}$$

First, the fixed-effect parameters were specified as in the case of no association in Section 3.1.1, while $b_i$ is a subject-specific parameter that was assumed to follow a normal distribution with zero mean and a variance of $0.25^2$, thus $b_i \sim N(0, 0.25^2)$. Datasets of varying sample sizes and cluster sizes were then generated from model (7) and the "true" marginal parameters obtained by fitting a univariate Poisson model ignoring the correlation. The parameters obtained are consistent though the efficiency with which they are estimated is compromised. Note that data are generated from a hierarchical model to which marginal models are then fitted. This implies that the true values for the $\beta$ parameters in (7) do not correspond to the true values for the marginal model. To deal with this issue, very large sample sizes were generated (starting from 1000 and going up all the way to 250,000) using the hierarchical model and then the subsequent marginal model was fitted.

For the largest sample sizes, very stable estimates were obtained. These values, $\beta_0^{(m)} = 1.5807$, $\beta_1^{(m)} = -0.1881$, $\beta_2^{(m)} = -0.0340$, and $\beta_3^{(m)} = 0.0192$ were used to calculate the bias in the case of dependent data. The superscript $(m)$ refers to 'marginal.'

### 3.2 Results

A comparison between GEE and PL is done in the context of hierarchical count data. GEE has been widely implemented in statistical software like SAS and R. Our proposed PL approach is implemented in SAS and a macro is available from the authors' web pages.

Not surprisingly, for independent data, GEE and PL parameter estimates are very similar (Table 1), with differences especially occurring in Table 1 for $K = 10$, $n_i = 2$. PL, however, has the covariance parameter $\theta_{st}$ estimated, which indicates a relative tendency to zero with increase in sample size and number of measurements per subject, as expected for independent data. For very small sample sizes, however, PL's performance is compromised as can also be seen from Table 2. Though $\theta_{st}$ hails from a Poisson distribution and is expected to be strictly positive, we argue that this interpretation takes effect in a hierarchical modeling framework. In the context of marginal modeling, this parameter can also take on negative values as is seen in Table 1. This phenomenon is often a source of confusion, and it is less well understood in non-Gaussian cases than for continuously distributed hierarchical data. Pryseley *et al* (2011) describe how such negative correlations can be estimated and interpreted for both Gaussian and non-Gaussian settings. One important situation where negative association is natural is where cluster members are in a competitive relation with one another. Molenberghs and Verbeke (2011) further discuss how a negative correlation can be reconciled with a hierarhical model interpretation.

Simulations with $\theta_{st}$ strictly positive in the case of data with association, see Table 6, in a marginal model perspective slightly improved the convergence rate while the bias and the MSE were more or less the same. In the case of data without association, the bias and MSE were also similar whether or not $\theta_{st}$ was constrained to be positive but the rate of convergence was reduced in the case of strictly positive $\theta_{st}$.

In the presence of correlation, $\theta_{st}$ is estimated above 1 as can be seen from Table 3 for the various combinations of $K$ and $n_i$. Convergence (Table 4) issues still persist for $K \approx 10$.

# 4 Data Analysis

Data on epileptic seizures were obtained from a randomized double-blind, parallel group multi-center study to compare a placebo(treatment=0) and a new anti-epileptic drug (AED) in combination with one or two other AED's (treatment=1). The randomization of the epileptic patients took place after a 12-week stabilization period. The number of seizures were counted during this baseline period after which 45 patients were assigned to the placebo group and 44 to the AED group. Patients were then followed weekly for 16 weeks and then enrolled into a long-term open-extension study. Patient characteristics including race, age (years), sex, height, and weight were also recorded. Some of the patients were followed for up to 27 weeks. The outcome of interest is the number of epileptic seizures experienced during the last week. Molenberghs and Verbeke (2005) and related references therein give more details and a report of earlier analyses of this set of data. Here, we analyze this dataset using three different approaches: (a) independence; (b) GEE; and (c) PL. Table 5 shows results of fitting a model for the evolution of the two treatment arms over time and, the same model but with a correction for baseline characteristics of the patients. Similar results are observed for GEE and PL, especially as far as the standard errors are concerned.

# 5 Concluding Remarks

We have put forward a particular form of pseudo-likelihood, also termed pairwise likelihood, to estimate parameters for a model fitted to repeated count data. Beneficially, the specification of a bivariate count-data model only is required. Unlike conventional generalized estimating equations, our method allows for the assessment of the association between pairs of measurements, in addition to the usual marginal mean parameters. Of course, one could consider a very general correlation structure with GEE, but this cannot be subjected to standard statisti-

cal assessment, e.g., based on hypothesis-testing based assessment. Alternatively, one could switch to second-order GEE (Zhao and Prentice 1990), but this come with considerable computational complexity.

Pseudo-likelihood, like generalized estimating equations, yields consistent and asymptotically normally distributed parameter estimates with a sandwich estimator used to calculate the variance. On the one hand, GEE remains computationally faster than PL because it only evaluates the first moment and plugs in working assumptions for the second. But because it allows for the misspecification of the working correlation structure, one cannot rely on the correlation estimates from GEE for formulating answers to scientific questions, should interest be in the association as well. The computational burden encountered in PL grows with the number of measurements per subject or cluster size, as evaluation of the marginal PL is done for all $[n_i(n_i - 1)]/2$ possible pairs of a subject.

It is important to realize that the method used for simulation does not match the assumed model. This can be seen as a drawback, but underscores that more and more flexible methods for simulating correlated Poisson data are needed. It is a topic of ongoing research.

The constant covariance terms, considered in this paper, can and will be relaxed in future developments.

In conclusion, pseudo-likelihood is a viable alternative when pairwise association between repeated counts is of interest. Of course, while these pairwise association parameters are fully part of the model, in spite of the fact that full likelihood is not specified, there may be a price in terms of efficiency loss. At the same time, with pairwise pseudo-likelihood, no three-way of higher-order parameters can be estimated.

**Further, and importantly, GEE2 and pairwise likelihood are less robust to misspecification of the association structure than conventional GEE. Of course, we have to place this against the background of functional restrictions on the correlation structure in marginal models. There are situations, especially with binary data, where a pairwise correlation structure is incompatible with the specified univariate mean functions. In such a case, it is better to have non-converging GEE and**

**PL, than a converged GEE which nevertheless cannot correspond to a valid joint distribution.**

**Generally, the less parametric the model, the higher the robustness towards misspecification. This simply means that whatever is not specified, cannot be misspecified. In this spirit, PL is robust against the entire higher-order association structure, given that it is not specified.**

**Robustness should also be seen against the existence of so-called parent distributions, i.e., full joint distributions that are compatible with the moments specified, e.g., the first and second moments in pairwise likelihood. Work has been done in this respect, e.g., by Molenberghs and Kenward (2010). These authors show that the parent provides a natural description of the framework into which the semi-parametrically specified parameters fit. The implication is that such semi-parametric methods as GEE1, GEE2, ALR, etc. can always be applied because there is always a valid parent, and hence a probabilistic basis. The sole condition is that the parametrically specified portion of the model be valid, but this is no different to any other statistical modeling exercise. It follows from the above that, when the pairwise correlation structure is grossly misspecified, the pairwise probabilities may be jeopardized and more so the parent distribution. This implies that robustness can come with important drawbacks. In pairwise likelihood, the modeler's obligation to reflect carefully on all that is specified is straightforwardly built in.**

## Acknowledgments

The authors gratefully acknowledge 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*. Boca Raton, FL: Chapman & Hall/CRC.

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

Breslow, N. and Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 813–820.

Chaganty, N.R. (1997). An alternative approach to the analysis of longitudinal data via Generalized Estimating Equations. *Journal of Statistical planning and Inference* **63**, 39–54.

Chaganty, N.R. and Shults, J. (1999). On eliminating the asymptotic bias in the quasi-least squares estimate of the correlation parameter. *Journal of Statistical Planning and Inference* **76**, 145-161.

Geys, H., Molenberghs, G., and Ryan, L.M. (1999). Pseudolikelihood modeling of multivariate outcomes in developmental toxicology. *Journal of the American Statistical Association* **94**, 734–745.

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

Karlis, D. (2003). An EM algorithm for multivariate Poisson distribution and related models. *Journal of Applied Statistics* **30**, 63–77.

Karlis, D. and Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *The Statistician* **52**, 381–393.

Kocherlakota, S. and Kocherlakota, K., (2001). Regression in the bivariate Poisson distribution. *Communications in Statistics, Theory & Methods* **30**, 815–825.

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

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

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

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

Molenberghs, G., Kenward, M.G., Verbeke, G., and Teshome Ayele, B. (2011). Pseudo-likelihood estimation for incomplete data. *Statistica Sinica*, **21**, 187–206.

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

Molenberghs, G. and Verbeke, G. (2011). A note on the hierarchical interpretation for negative variance components. *Statistical Modeling*, **11**, 389–408.

Molenberghs, G., Verbeke, G., and Demétrio, C.G.B. (2007). An extended random effects approach to modeling repeated, overdispersed count data. *Lifetime Data Analysis* **13**, 513–531.

Molenberghs, G., Verbeke, G., Demétrio, C.G.B., and Vieira, A. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, **25**, 325–347.

Parzen, M., Lipsitz, S.R., Fitzmaurice, G.M., Ibrahim, J.G., Troxel, A., and Molenberghs, G. (2007). Pseudo-likelihood methods for the analysis of longitudinal binary data subject to nonignorable non-monotone missingness. *Journal of Data Science* **5**, 1–21.

Pryseley, A., Tchonlafi, C., Verbeke, G., and Molenberghs, G. (2011). Estimating negative varinace components from Gaussian and non-Gaussian data: a mixed models approach. *Computational Statistics and Data Analysis*, **55**, 1071–1085.

Shults, J. and Chaganty, N.R. (1998). Analysis of serially correlated data using quasi-least squares. *Biometrics* **54**, 1622–1630.

Sun, W., Shults, J., and Leonard, M. (2009). A note on the use of unbiased estimating equations to estimate correlation in analysis of longitudinal trials. *Biometrical Journal* **51**, 5–18.

Troxel, A. B., Lipsitz, S.R., and Harrington, D.P. (1998). Marginal models for the analysis of longitudinal measurements with nonignorable non-monotone missing data. *Biometrika* **85**, 661–672.

Wang, Y. G. and Carey, V. J. (2004). Unbiased estimating equations from working correlation models for irregularly timed repeated measures. *Journal of the American Statistical Association* **99**, 845–852.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computing and Simulation* **48**, 233–243.

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

Zhao, L. and Prentice, R. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika* **77**, 642–648.

Zhao, Y. and Joe, H. (2005). Composite likelihood estimation in multivariate data analysis. *Canadian Journal of Statistics*, **33**, 335–356.

Zhao, L.P. and Prentice, R.L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika*, **77**, 642–648.

Table 1: *Simulation study, no association: Parameter estimates for GEE (exch. correlation) and pseudo-likelihood for varying number of measurements per subject ($n_i$) and sample size ($K$)*

| | | GEE | | | | Pseudo likelihood | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $K$ | $n_i$ | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\theta_{st}$ |
| *True value* | | 1.4531 | -0.1869 | -0.0328 | 0.0195 | 1.4531 | -0.1869 | -0.0328 | 0.0195 | . |
| 10 | 2 | 1.4147 | -0.1691 | -0.0306 | 0.0156 | 0.7424 | -0.3586 | -0.1007 | 0.1927 | 0.8545 |
| 10 | 4 | 1.4424 | -0.1842 | -0.0353 | 0.0203 | 1.4650 | -0.1754 | -0.0333 | 0.0178 | -0.1502 |
| 10 | 8 | 1.4452 | -0.1886 | -0.0329 | 0.0203 | 1.4653 | -0.1851 | -0.0323 | 0.0199 | -0.0906 |
| 10 | 16 | 1.4474 | -0.1863 | -0.0327 | 0.0196 | 1.4578 | -0.1836 | -0.0323 | 0.0193 | -0.0465 |
| 100 | 2 | 1.4512 | -0.1777 | -0.0336 | 0.0122 | 1.4488 | -0.1803 | -0.0339 | 0.0122 | -0.0066 |
| 100 | 4 | 1.4527 | -0.1882 | -0.0337 | 0.0201 | 1.4561 | -0.1875 | -0.0335 | 0.0200 | -0.0179 |
| 100 | 8 | 1.4525 | -0.1868 | -0.0331 | 0.0197 | 1.4547 | -0.1864 | -0.0330 | 0.0197 | -0.0101 |
| 100 | 16 | 1.4525 | -0.1868 | -0.0328 | 0.0197 | 1.4536 | -0.1866 | -0.0328 | 0.0197 | -0.0051 |
| 1,000 | 2 | 1.4546 | -0.1874 | -0.0342 | 0.0200 | 1.4546 | -0.1874 | -0.0342 | 0.0199 | -0.0016 |
| 1,000 | 4 | 1.4538 | -0.1863 | -0.0332 | 0.0193 | 1.4541 | -0.1863 | -0.0332 | 0.0193 | -0.0015 |
| 1,000 | 8 | 1.4527 | -0.1857 | -0.0328 | 0.0194 | 1.4532 | -0.1856 | -0.0328 | 0.0194 | -0.0021 |
| 1,000 | 16 | 1.4527 | -0.1858 | -0.0328 | 0.0194 | 1.4529 | -0.1858 | -0.0328 | 0.0194 | -0.0005 |
| 10,000 | 2 | 1.4536 | -0.1883 | -0.0330 | 0.0202 | 1.4536 | -0.1883 | -0.0330 | 0.0202 | 0.0000 |
| 10,000 | 4 | 1.4536 | -0.1865 | -0.0329 | 0.0194 | 1.4537 | -0.1865 | -0.0329 | 0.0194 | -0.0006 |
| 10,000 | 8 | 1.4536 | -0.1875 | -0.0329 | 0.0196 | 1.4536 | -0.1875 | -0.0329 | 0.0196 | -0.0001 |
| 10,000 | 16 | 1.4531 | -0.1871 | -0.0328 | 0.0195 | 1.4531 | -0.1871 | -0.0328 | 0.0195 | -0.0000 |

Table 2: *Simulation study, no association: Absolute bias in the parameter estimates and percent rate of convergence ($RATE_c$) for GEE and pseudo-likelihood for varying number of measurements per subject ($n_i$) and sample size ($K$)*

| | | GEE | | | | | Pseudo likelihood | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $K$ | $n_i$ | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $RATE_c$ | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $RATE_c$ |
| 10 | 2 | 0.0384 | 0.0178 | 0.0022 | 0.0039 | 99 | 0.7107 | 0.1717 | 0.0679 | 0.1732 | 68 |
| 10 | 4 | 0.0107 | 0.0027 | 0.0025 | 0.0008 | 100 | 0.0119 | 0.0115 | 0.0005 | 0.0017 | 95 |
| 10 | 8 | 0.0079 | 0.0017 | 0.0001 | 0.0008 | 100 | 0.0122 | 0.0018 | 0.0005 | 0.0004 | 100 |
| 10 | 16 | 0.0057 | 0.0006 | 0.0001 | 0.0001 | 100 | 0.0047 | 0.0033 | 0.0005 | 0.0002 | 100 |
| 100 | 2 | 0.0019 | 0.0092 | 0.0008 | 0.0073 | 100 | 0.0043 | 0.0066 | 0.0011 | 0.0073 | 100 |
| 100 | 4 | 0.0004 | 0.0013 | 0.0009 | 0.0006 | 100 | 0.0030 | 0.0006 | 0.0007 | 0.0005 | 100 |
| 100 | 8 | 0.0006 | 0.0001 | 0.0003 | 0.0002 | 100 | 0.0016 | 0.0005 | 0.0002 | 0.0002 | 100 |
| 100 | 16 | 0.0006 | 0.0001 | 0.0000 | 0.0002 | 100 | 0.0005 | 0.0003 | 0.0000 | 0.0002 | 100 |
| 1,000 | 2 | 0.0015 | 0.0005 | 0.0014 | 0.0005 | 100 | 0.0015 | 0.0005 | 0.0014 | 0.0004 | 100 |
| 1,000 | 4 | 0.0007 | 0.0006 | 0.0004 | 0.0002 | 100 | 0.0010 | 0.0006 | 0.0004 | 0.0002 | 100 |
| 1,000 | 8 | 0.0004 | 0.0012 | 0.0000 | 0.0001 | 100 | 0.0001 | 0.0013 | 0.0000 | 0.0001 | 100 |
| 1,000 | 16 | 0.0004 | 0.0011 | 0.0000 | 0.0001 | 100 | 0.0002 | 0.0011 | 0.0000 | 0.0001 | 100 |
| 10,000 | 2 | 0.0005 | 0.0014 | 0.0002 | 0.0007 | 100 | 0.0005 | 0.0014 | 0.0002 | 0.0007 | 100 |
| 10,000 | 4 | 0.0005 | 0.0004 | 0.0001 | 0.0001 | 100 | 0.0006 | 0.0004 | 0.0001 | 0.0001 | 100 |
| 10,000 | 8 | 0.0005 | 0.0006 | 0.0001 | 0.0001 | 100 | 0.0005 | 0.0006 | 0.0001 | 0.0001 | 100 |
| 10,000 | 16 | 0.0000 | 0.0002 | 0.0000 | 0.0000 | 100 | 0.0000 | 0.0002 | 0.0000 | 0.0000 | 100 |

Table 3: *Simulation study, association: Parameter estimates of GEE (exch. correlation) and pseudo-likelihood for varying number of measurements per subject ($n_i$) and sample size ($K$)*

| | | GEE | | | | Pseudo likelihood | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $K$ | $n_i$ | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\theta_{st}$ |
| *True value* | | 1.5807 | -0.1881 | -0.0340 | 0.0192 | 1.5807 | -0.1881 | -0.0340 | 0.0192 | . |
| 10 | 2 | 1.5183 | -0.1862 | -0.0470 | 0.0362 | 0.8550 | -0.4958 | -0.2284 | 0.3179 | 1.9756 |
| 10 | 4 | 1.5219 | -0.1749 | -0.0318 | 0.0222 | 1.2986 | -0.2531 | -0.0529 | 0.0343 | 1.5239 |
| 10 | 8 | 1.5328 | -0.1765 | -0.0329 | 0.0199 | 1.3454 | -0.2226 | -0.0473 | 0.0240 | 1.3882 |
| 10 | 16 | 1.5442 | -0.1716 | -0.0333 | 0.0199 | 1.4176 | -0.2365 | -0.0473 | 0.0256 | 1.2150 |
| 100 | 2 | 1.5651 | -0.1834 | -0.0311 | 0.0173 | 1.2493 | -0.2745 | -0.0561 | 0.0318 | 1.9167 |
| 100 | 4 | 1.5670 | -0.1806 | -0.0309 | 0.0191 | 1.2811 | -0.2670 | -0.0528 | 0.0282 | 1.8183 |
| 100 | 8 | 1.5703 | -0.1882 | -0.0326 | 0.0193 | 1.3179 | -0.2612 | -0.0520 | 0.0283 | 1.7083 |
| 100 | 16 | 1.5725 | -0.1798 | -0.0328 | 0.0195 | 1.3931 | -0.2718 | -0.0509 | 0.0284 | 1.4705 |
| 1,000 | 2 | 1.5774 | -0.1853 | -0.0328 | 0.0187 | 1.2591 | -0.2709 | -0.0535 | 0.0293 | 1.9175 |
| 1,000 | 4 | 1.5788 | -0.1870 | -0.0329 | 0.0192 | 1.2802 | -0.2687 | -0.0532 | 0.0283 | 1.8566 |
| 1,000 | 8 | 1.5776 | -0.1867 | -0.0329 | 0.0196 | 1.3179 | -0.2713 | -0.0524 | 0.0290 | 1.7337 |
| 1,000 | 16 | 1.5783 | -0.1865 | -0.0327 | 0.0195 | 1.3909 | -0.2738 | -0.0512 | 0.0287 | 1.4998 |
| 10,000 | 2 | 1.5779 | -0.1872 | -0.0326 | 0.0196 | 1.2628 | -0.2712 | -0.0549 | 0.0298 | 1.9194 |
| 10,000 | 4 | 1.5787 | -0.1880 | -0.0329 | 0.0198 | 1.2810 | -0.2706 | -0.0533 | 0.0289 | 1.8568 |
| 10,000 | 8 | 1.5778 | -0.1863 | -0.0328 | 0.0195 | 1.3179 | -0.2715 | -0.0525 | 0.0290 | 1.7367 |
| 10,000 | 16 | 1.5780 | -0.1871 | -0.0328 | 0.0195 | 1.3897 | -0.2742 | -0.0512 | 0.0287 | 1.5018 |

Table 4: *Simulation study, association: Absolute bias in the parameter estimates and percent rate of convergence ($RATE_c$) for GEE and pseudo-likelihood for varying number of measurements per subject ($n_i$) and sample size ($K$)*

| | | GEE | | | | | Pseudo likelihood | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $K$ | $n_i$ | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $RATE_c$ | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $RATE_c$ |
| 10 | 2 | 0.0624 | 0.0019 | 0.0130 | 0.0170 | 85 | 0.7257 | 0.3077 | 0.1944 | 0.2987 | 97 |
| 10 | 4 | 0.0588 | 0.0132 | 0.0022 | 0.0030 | 100 | 0.2821 | 0.0650 | 0.0189 | 0.0151 | 100 |
| 10 | 8 | 0.0479 | 0.0116 | 0.0011 | 0.0007 | 100 | 0.2353 | 0.0345 | 0.0133 | 0.0048 | 100 |
| 10 | 16 | 0.0365 | 0.0165 | 0.0007 | 0.0007 | 100 | 0.1631 | 0.0484 | 0.0133 | 0.0064 | 100 |
| 100 | 2 | 0.0156 | 0.0047 | 0.0029 | 0.0019 | 100 | 0.3314 | 0.0864 | 0.0221 | 0.0126 | 100 |
| 100 | 4 | 0.0137 | 0.0075 | 0.0031 | 0.0001 | 100 | 0.2996 | 0.0789 | 0.0188 | 0.0090 | 100 |
| 100 | 8 | 0.0104 | 0.0001 | 0.0014 | 0.0001 | 100 | 0.2628 | 0.0731 | 0.0180 | 0.0091 | 99 |
| 100 | 16 | 0.0082 | 0.0083 | 0.0012 | 0.0003 | 100 | 0.1876 | 0.0837 | 0.0169 | 0.0092 | 99 |
| 1,000 | 2 | 0.0033 | 0.0028 | 0.0012 | 0.0005 | 100 | 0.3216 | 0.0828 | 0.0195 | 0.0101 | 100 |
| 1,000 | 4 | 0.0019 | 0.0011 | 0.0011 | 0.0000 | 100 | 0.3005 | 0.0806 | 0.0192 | 0.0091 | 99 |
| 1,000 | 8 | 0.0031 | 0.0014 | 0.0011 | 0.0004 | 100 | 0.2628 | 0.0832 | 0.0184 | 0.0098 | 99 |
| 1,000 | 16 | 0.0024 | 0.0016 | 0.0013 | 0.0003 | 100 | 0.1898 | 0.0857 | 0.0172 | 0.0095 | 97 |
| 10,000 | 2 | 0.0028 | 0.0009 | 0.0014 | 0.0004 | 100 | 0.3179 | 0.0831 | 0.0209 | 0.0106 | 98 |
| 10,000 | 4 | 0.0020 | 0.0001 | 0.0011 | 0.0006 | 100 | 0.2997 | 0.0825 | 0.0193 | 0.0097 | 97 |
| 10,000 | 8 | 0.0029 | 0.0018 | 0.0012 | 0.0003 | 100 | 0.2628 | 0.0834 | 0.0185 | 0.0098 | 98 |
| 10,000 | 16 | 0.0027 | 0.0010 | 0.0012 | 0.0003 | 100 | 0.1910 | 0.0861 | 0.0172 | 0.0095 | 98 |

Table 5: *Epilepsy data: Parameter estimates (standard errors) for a univariate Poisson model, GEE (exchangeable correlation) and pseudo-likelihood (4).* The first block refers to a model testing for a difference in number of epileptic seizures between the two treatment arms over time. The second block corrects for patient characteristics including race, age, sex, height and weight.

| Parameter | Univariate | GEE | Pseudo-likelihood |
|---|---|---|---|
| Intercept | 1.4531 (0.0383) | 1.3165 (0.1799) | 0.91439 (0.29449) |
| treatment (0) | -0.1869 (0.0571) | 0.0156 (0.2931) | -0.06423 (0.41424) |
| study week | -0.0328 (0.0038) | -0.0147 (0.0168) | -0.03891 (0.01875) |
| study week×treatment (0) | 0.0195 (0.0058) | 0.0035 (0.0201) | 0.02845 (0.03558) |
| $\theta_{st}$ | | | 1.10170 (0.26994) |
| Intercept | 2.3963 (0.3576) | 4.0954 (3.9610) | 3.91804 (5.06465) |
| treatment (0) | -0.0992 (0.0578) | -0.0925 (0.2619) | -0.08047 (0.40335) |
| study week | -0.0299 (0.0039) | -0.0146 (0.0168) | -0.03403 (0.01824) |
| study week×treatment (0) | 0.0168 (0.0058) | 0.0033 (0.0206) | 0.02247 (0.03298) |
| race (1) | -0.0811 (0.0506) | -0.3298 (0.2904) | -0.07743 (0.54786) |
| age (years) | -0.0188 (0.0017) | -0.0200 (0.0115) | -0.02025 (0.01936) |
| sex (1) | 0.5747 (0.0575) | 0.8959 (0.3936) | 0.77549 (0.44018) |
| height | -0.0133 (0.0055) | -0.0429 (0.0576) | -0.03617 (0.07108) |
| weight | 0.0008 (0.0005) | 0.0023 (0.0040) | -0.00235 (0.00830) |
| $\theta_{st}$ | | | 1.07935 (0.25153) |

# Appendix

## A Consistency and Asymptotic Normality of the Pseudo-likelihood Estimator

We first list the required regularity conditions on the density functions $f_s(\boldsymbol{y}^{(s)}; \boldsymbol{\lambda})$.

**A0** The densities $f_s(\boldsymbol{y}^{(s)}; \boldsymbol{\lambda})$ are distinct for different values of the parameter $\boldsymbol{\lambda}$.

**A1** The densities $f_s(\boldsymbol{y}^{(s)}; \boldsymbol{\lambda})$ have common support, which does not depend on $\boldsymbol{\lambda}$.

**A2** The parameter space $\Omega$ contains an open region $\omega$ of which the true parameter value $\boldsymbol{\lambda}_0$ is an interior point.

**A3** $\omega$ is such that for all $s$, and almost all $\boldsymbol{y}^{(s)}$ in the support of $\boldsymbol{Y}^{(s)}$, the densities admit all third derivatives

$$\frac{\partial^3 f_s(\boldsymbol{y}^{(s)}; \boldsymbol{\lambda})}{\partial\theta_j \partial\theta_k \partial\theta_\ell}.$$

**A4** The first and second logarithmic derivatives of $f_s$ satisfy

$$E_{\boldsymbol{\lambda}}\left(\frac{\partial \ln f_s(\boldsymbol{y}^{(s)}; \boldsymbol{\lambda})}{\partial\theta_k}\right) = 0, \qquad k = 1, \ldots, q,$$

and

$$0 < E_{\boldsymbol{\lambda}}\left(\frac{-\partial^2 \ln f_s(\boldsymbol{y}^{(s)}; \boldsymbol{\lambda})}{\partial\theta_k \partial\theta_\ell}\right) < \infty, \qquad k, \ell = 1, \ldots, q.$$

**A5** The matrix $I_0$, defined in (8), is positive definite.

**A6** There exist functions $M_{klr}$ such that

$$\sum_{s\in S} \delta_s E_{\boldsymbol{\lambda}}\left|\frac{\partial^3 \ln f_s(\boldsymbol{y}^{(s)}; \boldsymbol{\lambda})}{\partial\theta_k \partial\theta_\ell \partial\theta_r}\right| < M_{k\ell r}(\boldsymbol{y})$$

for all $\boldsymbol{y}$ in the support of $f$ and for all $\boldsymbol{\theta} \in \omega$ and $m_{k\ell r} = E_{\boldsymbol{\lambda}_0}(M_{k\ell r}(Y)) < \infty$.

Theorem 1, proven by Arnold and Strauss (1991), guarantees the existence of at least one solution to the pseudo-likelihood equations, which is a consistent and asymptotically normal estimator. Without loss of generality, we can assume $\boldsymbol{\lambda}$ is constant. Replacing it by $\boldsymbol{\lambda}_i$, and modeling it as a function of covariates is straightforward.

**Theorem 1 (Consistency and Asymptotic Normality)** *Assume that $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N)$ are i.i.d. with common density that depends on $\boldsymbol{\lambda}_0$. Then under regularity conditions (A1)–(A6):*

1. *the pseudo-likelihood estimator $\tilde{\boldsymbol{\lambda}}_N$, defined as the maximizer of the pseudo-score function, converges in probability to $\boldsymbol{\lambda}_0$.*
2. *$\sqrt{N}(\tilde{\boldsymbol{\lambda}}_N - \boldsymbol{\lambda}_0)$ converges in distribution to $N_p(\mathbf{0}, I_0(\boldsymbol{\lambda}_0)^{-1} I_1(\boldsymbol{\lambda}_0) I_0(\boldsymbol{\lambda}_0)^{-1})$ with $I_0(\boldsymbol{\lambda})$ defined by*

$$I_{0,k\ell}(\boldsymbol{\lambda}) = -\sum_{s \in S} \delta_s E_{\boldsymbol{\lambda}} \left( \frac{\partial^2 \ln f_s(\boldsymbol{y}^{(s)}; \boldsymbol{\lambda})}{\partial \theta_k \partial \theta_\ell} \right) \tag{8}$$

*and $I_1(\boldsymbol{\lambda})$ by*

$$I_{1,k\ell}(\boldsymbol{\lambda}) = \sum_{s,t \in S} \delta_s \delta_t E_{\boldsymbol{\lambda}} \left( \frac{\partial \ln f_s(\boldsymbol{y}^{(s)}; \boldsymbol{\lambda})}{\partial \theta_k} \frac{\partial \ln f_t(\boldsymbol{y}^{(t)}; \boldsymbol{\lambda})}{\partial \theta_\ell} \right). \tag{9}$$

# B The First and Second Derivatives of the Log Pseudo-likelihood Function

Let

$$\mathbf{B} = \sum_{k=0}^{min(y_{is}, y_{it})} \frac{e^{\mathbf{X}_{is}\boldsymbol{\beta}(y_{is}-k) + \mathbf{X}_{it}\boldsymbol{\beta}(y_{it}-k)} \theta_{ist}^k}{(y_{is}-k)!(y_{it}-k)!k!}.$$

Then, the bivariate Poisson distribution for the two measurements $y_{is}$ and $y_{it}$ expressed in terms of the covariates at the two time points $s$ and $t$ is

$$f(y_{is}, y_{it}) = \exp\left[ -\left( e^{\mathbf{X}_{is}\boldsymbol{\beta}} + e^{\mathbf{X}_{it}\boldsymbol{\beta}} + \theta_{ist} \right) \right] \times \mathbf{B}. \tag{10}$$

This leads to the log PL function given as

$$p\ell(\boldsymbol{\lambda}|\mathbf{Y}) = \sum_{i=1}^{K} \sum_{s<t} \log f(y_{is}, y_{it})$$

from which the gradient and Hessian functions are derived with respect to $\boldsymbol{\beta}$ and $\theta_{ist}$ ($\theta_{st}$ here) as

$$\begin{aligned}
\frac{\partial p\ell}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{K}\sum_{s<t}\Big\{ \ -\Big(\mathbf{X}_{is}^T e^{\mathbf{X}_{is}\boldsymbol{\beta}} + \mathbf{X}_{it}^T e^{\mathbf{X}_{it}\boldsymbol{\beta}}\Big) + \mathbf{B}^{-1}A \ \Big\} \\
\frac{\partial p\ell}{\partial \theta_{st}} &= \sum_{i=1}^{K}\sum_{s<t}\big\{-1 + \mathbf{B}^{-1}\mathbf{C}_2\big\}
\end{aligned} \tag{11}$$

and

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}}\Big(\frac{\partial p\ell}{\partial \boldsymbol{\beta}}\Big) &= \sum_{i=1}^{K}\sum_{s<t}\Big\{ \ -(\mathbf{X}_{is}^T\mathbf{X}_{is}e^{\mathbf{X}_{is}\boldsymbol{\beta}} + \mathbf{X}_{it}^T\mathbf{X}_{it}e^{\mathbf{X}_{it}\boldsymbol{\beta}}) + \mathbf{B}^{-2}(A_d\mathbf{B} - \mathbf{A}\mathbf{A}^T) \ \Big\} \\
\frac{\partial}{\partial \theta_{st}}\Big(\frac{\partial p\ell}{\partial \theta_{st}}\Big) &= \sum_{i=1}^{K}\sum_{s<t}\mathbf{B}^{-2}(\mathbf{B}\mathbf{C}_3 - \mathbf{C}_2^2) \\
\frac{\partial}{\partial \theta_{st}}\Big(\frac{\partial p\ell}{\partial \boldsymbol{\beta}}\Big) &= \sum_{i=1}^{K}\sum_{s<t}\mathbf{B}^{-2}(\mathbf{B}\mathbf{C} - \mathbf{C}_2\mathbf{A})
\end{aligned} \tag{12}$$

where

$$\begin{aligned}
\mathbf{A}_1 =& e^{\mathbf{X}_{is}\boldsymbol{\beta}(y_{is}-k)+\mathbf{X}_{it}\boldsymbol{\beta}(y_{it}-k)} \\
\mathbf{A}_2 =& (y_{is}-k)\mathbf{X}_{is}^T + (y_{it}-k)\mathbf{X}_{it}^T \\
\mathbf{A} =& \sum_{k=0}^{min(y_{is},y_{it})} \frac{\theta_{st}^k}{(y_{is}-k)!(y_{it}-k)!k!}\mathbf{A}_1\mathbf{A}_2 \\
\mathbf{A}_d =& \sum_{k=0}^{min(y_{is},y_{it})} \frac{\theta_{st}^k}{(y_{is}-k)!(y_{it}-k)!k!}\mathbf{A}_2\mathbf{A}_2^T\mathbf{A_1} \\
\mathbf{C} =& \sum_{k=0}^{min(y_{is},y_{it})} \frac{k\theta_{st}^{k-1}\mathbf{A}_1\mathbf{A}_2}{(y_{is}-k)!(y_{it}-k)!k!} \\
\mathbf{C}_2 =& \sum_{k=0}^{min(y_{is},y_{it})} \frac{k\theta_{st}^{k-1}\mathbf{A}_1}{(y_{is}-k)!(y_{it}-k)!k!} \\
\mathbf{C}_3 =& \sum_{k=0}^{min(y_{is},y_{it})} \frac{\mathbf{A}_1}{(y_{is}-k)!(y_{it}-k)!k!}k(k-1)\theta_{st}^{k-2}
\end{aligned}$$

# C Additional Results

We present an additional Table 6 related to the hierarchical interpretation, where $\theta_{st}$ is constrained to be strictly positive.

Table 6: *Simulation study, association: Parameter estimates, MSE and convergence rate of pseudo-likelihood for varying number of measurements per subject ($n_i$) and sample size ($K$), when the covariance($\theta_{st}$) is constrained to be positive*

| | | Parameter Estimates | | | | | MSE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $K$ | $n_i$ | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\theta_{st}$ | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $RATE_c$ |
| *True value* | | 1.5807 | -0.1881 | -0.0340 | 0.0192 | . | | | | | |
| 10 | 2 | 1.0275 | -0.3366 | -0.1742 | 0.1689 | 1.9424 | 12.7503 | 18.6687 | 5.6976 | 8.0308 | 93 |
| 10 | 4 | 1.2937 | -0.2635 | -0.0534 | 0.0362 | 1.5647 | 0.3822 | 0.5165 | 0.0271 | 0.0497 | 98 |
| 10 | 8 | 1.3449 | -0.2225 | -0.0473 | 0.0241 | 1.3904 | 0.2210 | 0.3112 | 0.0033 | 0.0052 | 100 |
| 10 | 16 | 1.4176 | -0.2365 | -0.0473 | 0.0256 | 1.2149 | 0.1379 | 0.2314 | 0.0006 | 0.0007 | 100 |
| 100 | 2 | 1.2493 | -0.2745 | -0.0561 | 0.0318 | 1.9167 | 0.1760 | 0.1524 | 0.0236 | 0.0506 | 100 |
| 100 | 4 | 1.2811 | -0.2670 | -0.0528 | 0.0282 | 1.8182 | 0.1126 | 0.0592 | 0.0023 | 0.0046 | 100 |
| 100 | 8 | 1.3175 | -0.2615 | -0.0519 | 0.0283 | 1.7079 | 0.0821 | 0.0364 | 0.0006 | 0.0006 | 100 |
| 100 | 16 | 1.3923 | -0.2709 | -0.0509 | 0.0284 | 1.4696 | 0.0434 | 0.0281 | 0.0003 | 0.0002 | 100 |
| 1,000 | 2 | 1.2591 | -0.2709 | -0.0535 | 0.0293 | 1.9175 | 0.1094 | 0.0204 | 0.0022 | 0.0046 | 100 |
| 1,000 | 4 | 1.2807 | -0.2691 | -0.0533 | 0.0284 | 1.8555 | 0.0924 | 0.0119 | 0.0006 | 0.0005 | 100 |
| 1,000 | 8 | 1.3179 | -0.2713 | -0.0524 | 0.0290 | 1.7340 | 0.0703 | 0.0099 | 0.0004 | 0.0002 | 100 |
| 1,000 | 16 | 1.3907 | -0.2737 | -0.0512 | 0.0287 | 1.4994 | 0.0369 | 0.0095 | 0.0003 | 0.0001 | 100 |
| 10,000 | 2 | 1.2626 | -0.2709 | -0.0547 | 0.0296 | 1.9194 | 0.1018 | 0.0083 | 0.0006 | 0.0006 | 100 |
| 10,000 | 4 | 1.2811 | -0.2709 | -0.0534 | 0.0290 | 1.8570 | 0.0900 | 0.0074 | 0.0004 | 0.0001 | 100 |