A Mixed Autoregressive Probit Model for Ordinal Longitudinal Data

Cristiano Varin
Department of Statistics, University Ca' Foscari
San Giobbe, Cannaregio 873, I-30121 Venice, Italy
e-mail: sammy@unive.it
and
Claudia Czado
Center of Mathematical Sciences, Munich University of Technology
Boltzmann str. 3, D-85747 Garching, Munich, Germany
e-mail: cczado@ma.tum.de

September 11, 2009

Summary

Longitudinal data with binary and ordinal outcomes routinely appear in medical applications. Existing methods are typically designed to deal with short measurement series. In contrast, modern longitudinal data can result in large numbers of subject-specific serial observations. In this framework, we consider multivariate probit models with random effects to capture heterogeneity and autoregressive terms for describing the serial dependence. Since likelihood inference for the proposed class of models is computationally burdensome because of high dimensional intractable integrals, a pseudolikelihood approach is followed. The methodology is motivated by the analysis of a large longitudinal study on the determinants of migraine severity.

Key words: Autoregressive residuals; Composite likelihood; Longitudinal data; Migraine severity; Ordinal probit; Mixed models; Pairwise likelihood.

1 Introduction

Pain severity is often measured on rating scales which involve four to eleven categories ranging from the absence of symptoms to the most severe pain (Von Korff *et al.*, 2000, e.g.). For chronic and recurrent pain conditions, such as migraine and back pain, studying the symptom severity over a time period is crucial to detect common- and person-specific pain trigger conditions. To this aim, patients record the pain severity in a diary over some time period. See Bolger *et al* (2003) for general design, technology and analysis questions. With the availability of electronic data collection methods such as palmtop computers, the frequency of such assessments can be very high. Thus, it is important to develop statistical methods that are able to deal adequately with large longitudinal ordinal response data in cross sectional setups.

There exist several methods to deal with short longitudinal setups involving ordinal responses measured typically over four to seven time points. Many of them require the inclusion of random effects to deal with the dependence between subject-specific measurements, *e.g.* Hedeker and Gibbons (1994), Gibbons and Hedeker (1997), Liu and Hedeker (2005) and Todem *et al.* (2007). Estimation in such models are often based on Gauss-Hermite quadrature for the integration of random effects. Another proposal involves the global odds ratio suggested by Dale (1986); see Molenberghs and Lesaffre (1994) and Williamson *et al.* (1995). Still another approach is based on Markov transition models. Lee and Daniels (2007) extend this method from binary (Heagerty, 2002) to ordinal longitudinal data involving six time points. Böckenholt (1999) uses a first-order Markov process on the category indicators to capture the time dependence. His model is able to fit longer ordinal time series, but requires that all time points are equidistant and common to all units.

For studying binary time series, Piorecky et al. (1996) use generalized estimating equations (Liang and Zeger, 1986) to adjust for the dependency between measurements. Generalized estimating equations could also be used for ordinal valued time series if one is only interested in inference for regression parameters, see for example Liang *et al.* (1992), Lipsitz and Kim (1994), Heagerty and Zeger (1996), Fahrmeir and Pritscher (1996) and Delfino *et al* (2001).

All the above approaches are limited by the number of person-specific measurements or by other restrictions such as common equidistant time points. Motivated by a longitudinal study on migraine severity determinants, we propose a class of mixed ordered probit models with an autocorrelated component to capture subject-specific time-series variability. In §2, we describe the model class. In §3, we develop a computationally convenient composite likelihood approach for inference and model selection. §4 illustrates the application to the migraine pain severity data. The paper closes with some final remarks.

2 Mixed autoregressive ordinal probit models

Let Y_{ij} represent a categorical response with h possible ordered categories and let $\boldsymbol{x}_{ij}$ be a vector of p exploratory variables observed at time t_{ij} for observation $j = 1, \dots, n_i$ on subject $i = 1, \dots, m$. As usual in longitudinal studies, the m subjects are assumed to be independent.

The ordinal response Y_{ij} may be viewed as a censored observation from a hidden continuous variable Y_{ij}^*,

$$Y_{ij} = y_{ij} \quad \leftrightarrow \quad \alpha_{y_{ij}-1} < Y_{ij}^* \leq \alpha_{y_{ij}}, \quad y_{ij} \in \{1, \dots, h\},$$

where $-\infty \equiv \alpha_0 < \alpha_1 < \ldots < \alpha_{h-1} < \alpha_h \equiv \infty$ are suitable threshold parameters. The important case of binary response corresponds to $h = 2$ and a single threshold parameter α_1. Among several possible specifications for the relationship between the unobserved Y_{ij}^* and the vector of regressors $\boldsymbol{x}_{ij}$, a common choice is a linear mixed model of type

$$Y_{ij}^* = \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + U_i + \epsilon_{ij}, \tag{1}$$

where $\boldsymbol{\beta}$ is a vector of p unknown coefficients, also termed fixed effects, while the U_i are m mutually independent random effects describing the heterogeneity among different subjects and the ϵ_{ij} are underlying errors. Popular assumptions for the marginal distribution of ϵ_{ij} are logistic and normal distributions, leading to the cumulative logit and cumulative probit models for the observed Y_{ij}, respectively. Additionally we assume independence between ϵ_{ij} and U_i. For more details see Agresti (2002, §7). Here, we choose a probit model and assume that the random effects are normally distributed, $U_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. We consider these distributional assumptions for ease of interpretation and mathematical manageability, although the methodology discussed in this paper holds more generally.

Model identifiability for the resulting multivariate probit model requires restrictions on both the location and the scale of the unobserved process Y_{ij}^*. These requirements are met when the residuals ϵ_{ij} have unit variance, and the first cutpoint α_1 or, alternatively, the intercept β_1 is fixed to zero, see for example Chib and Greenberg (1998).

Probit models with random effects have a particularly convenient interpretation. In fact, it is straightforward to move from a subject-specific interpretation to a population level interpretation. For example, consider the probability that subject i experiences a certain level y_{ij} at time t_{ij}

$$\mathrm{pr}\left(Y_{ij} = y_{ij}; \boldsymbol{\theta}\right) = \mathrm{pr}\left(Y_{ij}^* \in (\alpha_{y_{ij}-1}, \alpha_{y_{ij}}]; \boldsymbol{\theta}\right) = \Phi\left(\frac{\alpha_{y_{ij}} - \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}}{\sqrt{\sigma^2+1}}\right) - \Phi\left(\frac{\alpha_{y_{ij}-1} - \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}}{\sqrt{\sigma^2+1}}\right), \tag{2}$$

where $\Phi(z)$ denotes the cumulative probability function of a standard Normal variable and $\boldsymbol{\theta}$ is the parameter vector, including the cutpoints $\boldsymbol{\alpha} = (\alpha_2, \cdots, \alpha_h)^{\mathrm{T}}$, the regressor coefficients $\boldsymbol{\beta}$ and the variance component σ^2. While the subject-specific effect of the covariates on the response is described by $\boldsymbol{\beta}$, from expression (2) it follows that the average population effect is governed by the rescaled coefficient $\boldsymbol{\beta}^{\mathrm{pop}} = \boldsymbol{\beta}/\sqrt{\sigma^2+1}$.

Commonly, probit models with random effects are constructed by assuming that the underlying errors ϵ_{ij} are mutually independent, $\epsilon_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. It follows that the joint distribution of the hidden variables for the ith subject $(Y_{i1}^*, \ldots, Y_{in_i}^*)^{\mathrm{T}}$ is multivariate Normal with standardized mean vector

$$\left(\frac{\boldsymbol{x}_{i1}^{\mathrm{T}}\boldsymbol{\beta}}{\sqrt{\sigma^2+1}}, \ldots, \frac{\boldsymbol{x}_{in_i}^{\mathrm{T}}\boldsymbol{\beta}}{\sqrt{\sigma^2+1}}\right)^{\mathrm{T}} \tag{3}$$

and correlation matrix with constant non-diagonal entries given by $\sigma^2/(\sigma^2+1)$.

Model fitting is typically performed by maximum likelihood. Denote by $\boldsymbol{y} = (\boldsymbol{y}_1^{\mathrm{T}}, \ldots, \boldsymbol{y}_m^{\mathrm{T}})^{\mathrm{T}}$ the vector of all observations, with $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^{\mathrm{T}}$ being the subvector of observations pertaining to the ith patient. Similarly denote the vectors of the corresponding hidden variables $\boldsymbol{Y}^*$ and $\boldsymbol{Y}_i^*$, respectively. The likelihood function for the usual probit model with underlying independent residuals involves m intractable integrals

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta};\boldsymbol{y}) =& \prod_{i=1}^{m}\int_{-\infty}^{\infty}\prod_{j=1}^{n_i}\mathrm{pr}\left(Y_{ij}=y_{ij}|\boldsymbol{x}_{ij},u_i;\boldsymbol{\theta}\right)f(u_i;\boldsymbol{\theta})\mathrm{d}u_i \\
=& \prod_{i=1}^{m}\int_{-\infty}^{\infty}\prod_{j=1}^{n_i}\left(\int_{\alpha_{y_{ij}-1}}^{\alpha_{y_{ij}}} f(y_{ij}^*|\boldsymbol{x}_{ij},u_i;\boldsymbol{\theta})\mathrm{d}y_{ij}^*\right)f(u_i;\theta)\mathrm{d}u_i \\
=& \prod_{i=1}^{m}\int_{-\infty}^{\infty}\prod_{j=1}^{n_i}\left(\Phi(\alpha_{y_{ij}}-\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}-u_i)-\Phi(\alpha_{y_{ij}-1}-\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}-u_i)\right)\phi\left(\frac{u_i}{\sigma}\right)\mathrm{d}u_i, \qquad (4)
\end{aligned}$$

where $\phi(z)$ denotes the probability density function of a standard Normal variable. This likelihood may be approximated by Gauss-Hermite quadrature or, more accurately, by adapative Gauss-Hermite quadrature (see for example, Pinheiro and Bates (2000)).

Although the above described probit mixed model is widely used, its underlying equal correlation assumption seems unsatisfactory for many longitudinal studies, especially for those characterized by moderate to long subject-specific series. Better models should take into account the serial correlation within each subject-specific time-series. In this paper, we propose to model the within-subject serial correlation by a smooth temporal decaying correlation function as, for example, the exponential correlation model (Diggle *et al.*, 2002), $\mathrm{corr}(\epsilon_{ij},\epsilon_{ik})=\exp(-\delta|t_{ij}-t_{ik}|)$, where t_{ij} denotes the measurement time of observation y_{ij}. This correlation function reduces to the autoregressive model of order one, $\gamma^{|t_{ij}-t_{ik}|}$ with $\gamma=e^{-\delta}$, for equi-spaced observations times. Correspondingly, the correlation between two hidden continuous variables is formed by a constant subject-specific level plus a smooth serial component

$$\mathrm{corr}(Y_{ij}^*,Y_{ik}^*)=\frac{\sigma^2}{\sigma^2+1}+\frac{e^{-\delta|t_{ij}-t_{ik}|}}{\sigma^2+1}. \qquad (5)$$

Thus, by assuming serial correlation among the residuals, we obtain a multivariate probit model with the same marginal interpretation as in (2) but with a more realistic longitudinal structure. Thereafter, the proposed class of models will be termed mixed autoregressive ordinal probit (MAOP) models.

Further model flexibility may be obtained by allowing the parameter δ to depend on a subject-specific factor S_i with q different levels. Thus, the model may describe different memory effects in different groups of subjects. For example, in the migraine data discussed in §4 different pain memory effects can be postulated in patients taking medications or not, or in patients with different headaches types.

The cost for the versatility of the MAOP model is paid in terms of computational difficulties. The likelihood function still involves m intractable integrals but with dimensions corresponding to the cluster sizes $n_1,\ldots,n_m$. Denote always by $\boldsymbol{\theta}$ the parameter vector that now contains also the autocorrelation parameters $\boldsymbol{\delta}$. The likelihood function for the model with serially correlated residuals is

$$\mathcal{L}(\boldsymbol{\theta};\boldsymbol{y})=\prod_{i=1}^{m}\int_{-\infty}^{\infty}\left(\int_{\alpha_{y_{i1}-1}}^{\alpha_{y_{i1}}}\cdots\int_{\alpha_{y_{in_i}-1}}^{\alpha_{y_{in_i}}} f(y_{i1}^*,\ldots,y_{in_i}^*|\boldsymbol{x}_{i1},\ldots,\boldsymbol{x}_{in_i},u_i;\boldsymbol{\theta})\mathrm{d}y_{i1}^*\ldots\mathrm{d}y_{in_i}^*\right)f(u_i;\boldsymbol{\theta})\mathrm{d}u_i.$$

By using the assumptions of normality for both the random effects U_i and the hidden errors ϵ_{ij}, the likelihood may be rewritten as the product of m integrals of multivariate Normal densities of dimensions $n_1, \ldots, n_m$

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{m} \int_{\tilde{\alpha}_{y_{i1}-1}}^{\tilde{\alpha}_{y_{i1}}} \ldots \int_{\tilde{\alpha}_{y_{in_i}-1}}^{\tilde{\alpha}_{y_{in_i}}} \phi_{n_i}\left(z_{i1}, \ldots, z_{in_i}; \mathrm{R}_i\right) \mathrm{d}z_{i1} \ldots \mathrm{d}z_{in_i}, \tag{6}$$

where $\tilde{\alpha}_{y_{ij}}$ indicates the standardized cutpoint, $\tilde{\alpha}_{y_{ij}} = \left(\alpha_{y_{ij}} - \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}\right)/\sqrt{\sigma^2+1}$. The integrands $\phi_{n_i}\left(z_{i1}, \ldots, z_{in_i}; \mathrm{R}_i\right)$ are n_i-dimensional multivariate Normal densities with zero means and correlation matrix R_i whose entries are given by expression (5).

Except for longitudinal data with small numbers of observations per subject, the direct computation of likelihood (6) is time-consuming and possibly numerically unstable.

MAOP models for discrete-time observations are categorized Gaussian linear state space models. The celebrated Kalman filter (Kalman, 1960) allows efficient iterative computation of the exact likelihood function in Gaussian linear state space models, but cannot be applied to censored observations. Reliable approaches use several kinds of Monte Carlo approximations based typically on Kalman filter-type iterations, see for example Durbin and Koopman (2001). A Bayesian analysis of binary time series allowing for covariates using Markov Chain Monte Carlo methods and the simulation smoother of De Jong and Shephard (1995) for block updates of the hidden process variables were developed in Czado and Song (2008). It would be feasible to extend their approach to ordinal-valued time series models using ideas of Müller and Czado (2005) and Müller and Czado (2008) to update the threshold parameters.

Unfortunately, these computer-intensive approaches may be difficult to apply in large longitudinal data sets, such as the migraine data analysed in §4. Moreover, even if the computational cost would be tolerable, a full likelihood approach might be impractical due to the difficulty of assessing the adequacy of the multivariate assumptions underlying the model. These considerations lead us to consider a composite likelihood approach (Lindsay, 1988).

3 Composite likelihood inference

The term composite likelihood denotes a rich class of pseudolikelihoods constructed by compounding valid likelihoods based on data subsets. Recent applications include genetics, spatial statistics, time series and longitudinal data analysis; see Varin (2008) for a recent review.

Here, we focus on the composite likelihood constructed combining likelihoods for pairs of observations, also called pairwise likelihood (Le Cessie and Van Houwelingen, 1994). Since pairs formed from closest observations are likely to be more informative, it is convenient to restrict to the pseudolikelihood constructed from the marginal probabilities of observed pairs of outcomes less distant than q units,

$$p\ell^{(q)}(\boldsymbol{\theta};\boldsymbol{y}) = \sum_{i=1}^{m}\sum_{j<k}^{n_i} \log \mathrm{pr}(Y_{ij}=y_{ij}, Y_{ik}=y_{ik};\boldsymbol{\theta})\mathbb{1}_{[-q,q]}(t_{ij}-t_{ik}),$$

where $\mathbb{1}_{\mathcal{A}}(x)$ is the indicator of the event $\{x : x \in \mathcal{A}\}$. Note that $p\ell^{(q)}(\cdot;\boldsymbol{y})$ is a weighted pairwise likelihood with dummy weights used to exclude pairs too far apart. A recent detailed discussion of weighted versions of pairwise likelihood can be found in Joe and Lee (2009).

In contrast to a full likelihood approach, the pairwise likelihood for MAOP models involves only two-dimensional Gaussian integrals,

$$\mathrm{pr}(Y_{ij}=y_{ij}, Y_{ik}=y_{ik};\boldsymbol{\theta}) = \int_{\tilde{\alpha}_{y_{ij}-1}}^{\tilde{\alpha}_{y_{ij}}}\int_{\tilde{\alpha}_{y_{ik}-1}}^{\tilde{\alpha}_{y_{ik}}} \phi_2\left(z_{ij}, z_{ik}; \frac{\sigma^2}{\sigma^2+1} + \frac{e^{-\delta_{w_i}|t_{ij}-t_{ik}|}}{\sigma^2+1}\right)\mathrm{d}z_{ij}\mathrm{d}z_{ik}.$$

The maximum composite likelihood estimator for $\boldsymbol{\theta}$ solves the composite likelihood score equation,

$$u^{(q)}(\boldsymbol{\theta};\boldsymbol{y}) = \sum_{i=1}^{m} u_i^{(q)}(\boldsymbol{\theta};\boldsymbol{y}_i) = \sum_{i=1}^{m}\sum_{j<k}^{n_i} u_{i\cdot jk}(\boldsymbol{\theta};\boldsymbol{y}_i)\mathbb{1}_{[-q,q]}(t_{ij}-t_{ik}),$$

where $u_{i\cdot jk}(\boldsymbol{\theta};\boldsymbol{y}_i) = \nabla \log \mathrm{pr}(Y_{ij}=y_{ij}, Y_{ik}=y_{ik};\boldsymbol{\theta})$. Since $u^{(q)}(\boldsymbol{\theta};\boldsymbol{y})$ is a linear combination of proper score functions associated with each pairwise term forming the pseudolikelihood, it follows that, under standard assumptions (Molenberghs and Verbeke, 2005, §9), the maximum pairwise likelihood estimator $\hat{\boldsymbol{\theta}}^{(q)}$ is consistent and asymptotically Normally distributed. See also Cox and Reid (2004) for a discussion on situations in which consistency of maximum pairwise likelihood estimators may not hold, such as in long-memory temporal processes.

The asymptotic variance of $\hat{\boldsymbol{\theta}}^{(q)}$ assumes the typical "sandwich" form,

$$\Sigma^{(m)}(\boldsymbol{\theta}) = \mathrm{H}^{(q)}(\boldsymbol{\theta})^{-1}\mathrm{J}^{(q)}(\boldsymbol{\theta})\mathrm{H}^{(q)}(\boldsymbol{\theta})^{-1},$$

where $\mathrm{H}^{(q)}(\boldsymbol{\theta}) = -\mathrm{E}\{\nabla u^{(q)}(\boldsymbol{\theta};\boldsymbol{Y})\}$ and $\mathrm{J}^{(q)}(\boldsymbol{\theta}) = \mathrm{cov}\{u^{(q)}(\boldsymbol{\theta};\boldsymbol{Y})\}$. The inverse of $\Sigma^{(q)}(\boldsymbol{\theta})$ is also termed Godambe Information (Song, 2007, §3). An empirical estimate of $\mathrm{H}^{(q)}(\boldsymbol{\theta})$ is $-\nabla u^{(q)}(\hat{\boldsymbol{\theta}}^{(q)};\boldsymbol{y})$. Alternatively, exploiting the information identity for each pairwise term forming the pseudolikelihood, $\mathrm{H}^{(q)}(\boldsymbol{\theta})$ may be conveniently estimated by

$$\hat{\mathrm{H}}^{(q)}(\boldsymbol{y}) = \sum_{i=1}^{m}\sum_{j<k}^{n_i} u_{i\cdot jk}(\hat{\boldsymbol{\theta}}^{(q)};\boldsymbol{y}_i)u_{i\cdot jk}(\hat{\boldsymbol{\theta}}^{(q)};\boldsymbol{y}_i)^{\mathrm{T}}\mathbb{1}_{[-q,q]}(t_{ij}-t_{ik}), \tag{7}$$

thus avoiding the need to derive Hessian matrices. The natural empirical estimate of $\mathrm{J}^{(q)}(\boldsymbol{\theta})$ is

$$\hat{\mathrm{J}}^{(q)}(\boldsymbol{y}) = \sum_{i=1}^{m} u_i^{(q)}(\hat{\boldsymbol{\theta}}^{(q)};\boldsymbol{y}_i)u_i^{(q)}(\hat{\boldsymbol{\theta}}^{(q)};\boldsymbol{y}_i)^{\mathrm{T}}. \tag{8}$$

Matrices $\hat{\mathrm{H}}^{(q)}(\boldsymbol{y})$ and $\hat{\mathrm{J}}^{(q)}(\boldsymbol{y})$ are key ingredients for high-level inferential tasks such as hypothesis testing and model selection. The composite likelihood information criterion (CLIC) by Varin and Vidoni (2005) is a direct generalization of the Akaike (1973) criterion for model selection with composite likelihoods. The CLIC suggests to prefer models with smaller values of the quantity

$$\mathrm{CLIC}^{(q)} = -2\left(p\ell^{(q)}(\hat{\boldsymbol{\theta}};\boldsymbol{y}) - d^{(q)}(\boldsymbol{y})\right)$$

where $d^{(q)}(\boldsymbol{y})$ is an estimate of the effective number of parameters of the model. A consistent estimate of $d^{(q)}(\boldsymbol{y})$ is given by the trace of the matrix $\hat{\Sigma}^{(q)}(\boldsymbol{y})\hat{\mathrm{H}}^{(q)}(\boldsymbol{y})$. This information criterion may be seen as a form of the Takeuchi (1976) information criterion for model selection with misspecified likelihoods, being the pairwise likelihood a misspecified likelihood under the working assumption of independent pairs.

Regarding the choice of the maximal admissible distance q between pairs used in the pairwise likelihood, previous work on pairwise likelihood for temporal and spatial processes suggests that the inclusion of too-distant pairs is not only computationally inefficient, but may also not improve statistical efficiency, see Varin (2008). Here, we propose to choose the tuning parameter q as the value minimizing a global fitting criterion, for example the generalized variance defined as the determinant of $\hat{\Sigma}^{(q)}(\boldsymbol{y})$.

The web supplementary appendix (http://www.biostatistics.oxfordjournals.org) contains details on a simulation study carried out to evaluate the finite-sample performance of the proposed inferential methods. The results suggest that maximum pairwise likelihood estimators behave well for all the parameters even in case of strong serial correlation among the hidden variables. Computer code written in the R language (R Development Core Team, 2008) is also included in the supplementary material.

4 Migraine severity data

Prince *et al.* (2004) report that forty-five million Americans seek medical attention for headaches yearly, at an estimated labor cost of $13 billion. They show that only half of the migraine patients are affected by weather conditions. In contrast some studies show little or no effect of weather conditions on migraine severity, see Cooke *et al.* (2000) and Prince *et al.* (2004) for specific references. However in these studies only the frequency of headache occurrences, the daily maximum or total score of an ordinal severity levels have been studied.

Current strategies for the analysis of pain severity data measured on an ordinal scale require aggregating over periods to achieve continuous average or total pain scores, *e.g.* Cooke *et al.* (2000), Prince *et al.* (2004), Goldstein *et al* (2005) and Raskin *et al.* (2005). Such an approach ignores effects occurring during the aggregation periods.

Here, we directly model the observed severity categories collected using a headache diary. In particular, we investigate the four daily ratings – recorded at morning, noon, afternoon and bedtime – of the headache intensity of 133 Canadian (Toronto) patients in a study conducted by psychologist T. Kostecki-Dillon during the years 1993-1996. Records of the migraine severity were made on an ordinal scale with six categories described in Table 1.

Table 1 about here.

In addition to a subject-specific questionnaire with personal and clinical information, weather conditions were recorded. They were collected from the meteorological station closest to the place where patients spends most of their time. The weather covariates include measurements related to sunshine, humidity, wind direction and speed, windchill, pressure, air quality and many others.

Patients with a very large number of missing observations in subsequent measurements, or with less than one day of measurements, were omitted. The final data set comprises 119 patients with a total of 16,366 measurements, 1,157 of which are missing. We assume an ignorable missing mechanism and thus we base inference on the pairwise likelihood formed by pairs of observed outcomes. The numbers of observations per patient vary from 16 (4 days) to 1,352 (338 days). Observations are not necessarily consecutive. Often the subject-specific observations are organized in separated measurement periods, each of them formed by consecutive observations. The minimal measurement period is one day (four measurements), while the maximal one is 213 days (815 measurements).

Table 2 reports the observed proportions of the transitions between the ordinal categories in two consecutive measurements. Serial correlation in the data is suggested by the patterns of symptom persistence and of transitions between adjacent categories.

Table 2 about here.

For illustration, we study the relationship between headache severity using university degree status and the usage of analgesics as base variables. Three weather covariates are additionally included. The first is the change in atmospheric pressure from the previous day, categorized in three levels, namely from high ($> 1013\,hPa$) to low pressure ($\leq 1013\,hPa$), from low to high pressure and unchanged level of pressure (from low to low or from high to high). The second weather covariate is the relative humidity index with three levels, that is less than 60% of humidity, between 60% and 80% of humidity and more than 80% of humidity. The last weather covariate is windchill categorized into four classes, between $-50°$ and $-10°$ Celsius, between $-10°$ and $0°$ Celsius, between $0°$ and $10°$ Celsius and between $10°$ and $30°$ Celsius.

We consider the two binary covariates university degree and usage of analgesics as covariates of primary interest, thus they are included in all considered models. The base model is

$$\text{headache} \sim \text{university+analgesics}$$

Furthermore, we consider two different autocorrelation parameters γ for subjects with analgesics intake and those without.

For model comparison, it is necessary to fit all models of interest with a pairwise likelihood constructed from the same pairs of observations, that is with the same distance q. We choose q as the value minimizing the generalized variance for the larger model

$$\text{headache} \sim \text{university+analgesics+change+humidity+windchill}$$

According to this criterion, the overall best performance is obtained with $q = 12$. Thus, we fit all the other (nested) models with this value for q.

Table 3 shows the $2^3 = 8$ models obtained by adding all the possible combinations of the three weather covariates to the base model. The relative performance of the kth model with respect to the alternative models can be summarized by CLIC weights defined as $w_k = e^{-\Delta_k} / \sum_{k=1}^{8} e^{-\Delta_k}$, where $\Delta_k = (\mathrm{CLIC}_k - \min_i \mathrm{CLIC}_i)/2$.

Table 3 about here.

Qualitative conclusions should take into account the fitted models with their relative importance expressed through the CLIC weights. In Table 4, for illustration we show parameter estimates and standard errors only for the two best models, namely the model including the change in pressure and the base model.

Table 4 about here.

When considering also the other six fitted models, we obtain the following overall conclusions. Subjects with an university degree tend to suffer from lower levels of headache, while those taking analgesics have stronger symptoms. These variables have more predictive impact on the headache symptoms than the considered weather effects. Among the latter, only the change in the atmospheric pressure is significant, in that its decrease is associated with raised headache severity. The categorized relative humidity appears weakly significant and the windchill covariate even less.

Finally, in all fitted models there is no appreciable difference between the symptom persistence for patients who took analgesics and those who did not. Indeed, the difference between the autocorrelation parameters for the analgesic users (γ_T) and the non-analgesic users (γ_F) for all models is estimated between 0.133 and 0.148 with standard errors ranging between 0.096 and 0.099. This is confirmed by re-fitting the models with a common autocorrelation parameter γ for all patients: Table 4 reports the best model with and without diversified autocorrelation parameters. The model with common autocorrelation has a value of the CLIC statistic of 5915.8, thus somehow smaller than that of the corresponding model with two different autocorrelation parameters (CLIC equals to 5916.3).

5 Concluding remarks

We have developed a pseudolikelihood approach for analyzing a large longitudinal study on migraine severity symptoms. The proposed methodology is general and may be useful for other studies with ordinal, as well as binary, outcomes.

The main advantage from pairwise likelihood inference is its computational simplicity. Moreover since only the specification of bivariate margins is required, our approach relies on model assumptions to a lesser degree than any approach based on a full likelihood approximation. Some loss of efficiency may be experienced for the composite likelihood method compared to a full likelihood, but full likelihood is intractable for large numbers of observations per subject. The study of the efficiency of maximum pairwise likelihood estimators is possible only with a small number of observations per patients as in Joe and Lee (2009) whose results encourage the use of this pseudolikelihood.

The underlying Normal assumptions leading to the multivariate probit model were considered mainly for ease of interpretation. However, there are no theoretical restrictions against considering other distributional assumptions. An alternative of possible interest is a cumulative logit model (Agresti, 2002, §7) for the conditional distribution of the response given the random effects.

Other useful variants of the proposed class of models may involve robustification of the random effect distribution, for example by using a Student t-distribution instead of the traditional Gaussian distribution.

Often in longitudinal studies the missing data mechanism may not be assumed ignorable and thus likelihood-type analysis based on complete observations are not valid. Modification of the pairwise likelihood for non-ignorable missing data mechanism are described in Parzen et al. (2007).

As with standard likelihood inference, maximum pairwise likelihood estimators for variance components fail to correct for the degrees of freedom lost for estimating fixed effects and thus are prone to severe downward bias. When the number of covariates is not small compared to the number of subjects, bias in the estimate of σ^2 can be worthy of attention. Among several bias correction procedures, resampling methods such as jackknife and bootstrap are viable approaches given the low computational cost of pairwise likelihood evaluations. Further computational saving may be obtained by using first-order approximations instead of complete maximization of the pseudolikelihood for each resampled data.

Standard errors estimated from the empirical quantities (7) and (8) may be numerical imprecise for longitudinal studies with few subjects, typically leading to overoptimistic standard errors. More robust variance estimates for small numbers of subjects may obtained with resampling techniques such as jackknife and bootstrap.

Acknowledgements

This work was supported by *Ministero dell'Università e della Ricerca* [PRIN 2006] to C.V. and *Deutsche Forschungsgemeinschaft* [CZ 86/1-3] to C.C. We would like to thank T. Kostecki-Dillon for providing the data and Ruggero Bellio and Annamaria Guolo for helpful discussion. Finally we acknowledge that the comments of the referees, the Associate Editor and the Editor greatly improved the paper.

References

Agresti, A. (2002). *Categorical Data Analysis.* Wiley. Second edition.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In

B. Petrov and F. Caski (Eds.), Proc. Second International Symposium on Information Theory, Budapest: Akademiai Kiado, pp. 267–281.

Böckenholt, U. (1999). Measuring change: Mixed Markov models for ordinal panel data. *British Journal of Mathematical and Statistical Psychology* **52**, 125–136.

Bolger, N., Davis, A. and Rafaeli, E. (2003). Diary methods: capturing life as it is lived. *Annu. Rev. Psychol.* **54**, 579–616.

Chib S. and Greenberg E. (1998). Analysis of multivariate probit models. *Biometrika* **85**, 347–361.

Cooke, L., Rose, M. and Becker W. (2000). Chinook winds and migraine headache. *Neurology* **54**, 302–307.

Cox, D. and Reid N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **92**, 729–737.

Czado, C. and Song, P. X.-K. (2008). State space mixed models for longitudinal observations with binary and binomial responses. *Statistical Papers* **49**, 691–714.

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

Delfino, R.J., Jamner, L.D. and Whalen, C.K. (2001). Temporal analysis of the relationship of smoking behavior and urges to mood states in men versus women. *Nicotine & Tobacco Research* **3**, 235–248.

De Jong, P. and Shephard, N. (1995). The simulation smoother for time series models *Biometrika* **82**, 339–350.

Diggle, P.J., Heagerty, P.J., Liang K.Y. and Zeger, S.L. (2002). *The analysis of longitudinal data.* Oxford University Press. Second edition.

Durbin, J. and Koopman, S.J. (2001). *Time Series Analysis by State Space Methods.* Oxford University Press.

Fahrmeir, L. and Pritscher, L. (1996). Regression analysis of forest damage by marginal models for correlated ordinal responses. *Journal of Environmental and Ecological Statistics* **3**, 257–268.

Gibbons, R. and Hedeker, D. (1997). Random effects probit and logistic regression models for three-level data. *Biometrics* **53**, 1527–1537.

Goldstein, D.J., Lu Y., Detke, M.J., Lee, T.C. and Iyengar, S. (2005). Duloxine versus placebo in patients with painful diabetic neuropathy. *Pain* **116**, 109–118.

Heagerty, P. , (2002). Marginalized transition models and likelihood inference for longtudinal categorical data. *Biometrics* **58**, 342–351.

Heagerty, P. and Zeger, S. (1996). Marginal regression models for clustered ordinal measurements. *Journal of the American Statistical Association* **91**, 809–822.

Hedeker, D. and Gibbons, R. (1996). A random-effects ordinal regression model for multilevel analysis. *Biometrics* **50**, 933–944.

Joe, H. and Lee, Y. (2009). On weighted of bivariate margins in pairwise likelihood. *Journal of Multivariate Analysis* **100**, 670–685.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering* **82**, 35–45.

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

Lee, K. and Daniels, M.J. (2007). A class of Markov models for longitudnal ordinal data. *Biometrics* **63**, 1060-1067.

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

Liang, K.-Y., Zeger, S. and Qaqish, B. (1992). Multivariate regression analyses for categorical data (with discussion). *Journal of the Royal Statistical Society B* **54**, 2–24.

Lindsay (1988). Composite likelihood methods. In N. Prabhu (Ed.), Statistical Inference from Stochastic Processes, pp. 221–239. Providence, RI: American Mathematical Society.

Lipsitz, S.R. and Kim K. (1994). Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine* **13**, 1149–1163.

Liu, I. and Agresti, A. (2005). The analysis of ordered categorical data: an overview and a survey of recent developments (with discussion). *Test* **14**, 1–73.

Liu, L. C. and Hedeker, D. (2006). A mixed-effects regression model for longitudinal multivariate ordinal data. *Biometrics* **62**, 261–268.

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

Müller, G. and Czado, C. (2005). An autoregressive ordered probit model with application to high frequency financial data. *Journal of Computational and Graphical Statistics* **14**, 320–338.

Müller, G. and Czado, C. (2008). Stochastic volatility models for ordinal valued time series with application to Finance. *Statistical Modelling* to appear.

Molenberghs, G. and G. Verbeke (2005). *Models for Discrete Longitudinal Data*. Springer-Verlag.

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 Sub ject to Nonignorable Non-monotone Missingness. *Journal of Data Science* **5**, 1–21.

Pinheiro, J.C. and D.M. Bates (2000). *Mixed-Effects Models in S and S-PLUS*. 2nd Edition. Springer-Verlag.

Piorecky, J., Becker, W. and Rose, M. (1996). The effect of chinook winds on the probability of migraine headache occurence. *Headache* **37**, 153–158.

Prince, P., Rapoport, A., Sheftell, F., Tepper, S. and Bigal, M. (2004). The effect of weather on headache. *Headache* **44**, 153–158.

Raskin, J., Prichitt, Y.L., Wang F., D'Scouza, D.N., Waninger, A. L., Iyengar, S. and Wernicke, J.F. (2005). A double-blind, randomized multicenter trial comparing duloxetine with placebo in the management of diabetic perpheral neuropathic pain. *Pain Medicine* **6**, 346–356.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Song, P.X.-K. (2007). Correlated Data Analysis: Modeling, Analytics and Applications. Springer.

Takeuchi (1976). Distribution of information statistics and criteria for adequacy of models. *Mathematical Science* **153**, 12–18.

Todem, D., Kim, K. and Lesaffre, E. (2007). Latent-variable models for longitudinal data with bivariate ordinal outcomes. *Statistics in Medicine* **26**, 1034–1054.

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

Varin, C. and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika* **92**, 519–528.

Von Korff, M., Jensen, M.P. and Karoly, P. (2000). Assessing global pain severity by self-report in clinical and health services research. *Spine* **25**, 3140–3151.

Williamson, J.M., Kim, K. and Lipsitz, S.R. (1995) . Analyzing bivariate ordinal data using a global odds ratio. *Journal of the American Statistical Association* **90**, 1432–1437.

Table 1: Migraine data. Description of response categories with observed frequencies.

intensity	frequency	condition	description
0	9210	no headache	
1	2455	mild headache	aware of it only when attending to it
2	1685	moderate headache	could be ignored at times
3	1156	painful headache	continuously aware of it, but able to start or continue daily activities as usual
4	526	severe headache	continuously aware of it, difficult to concentrate and able to perform only undemanding tasks
5	177	intense headache	continuously aware of it, incapacitating unable to start or continue activity

Table 2: Migraine data. Observed two-step transition proportions.

	0	1	2	3	4	5
0	0.83	0.10	0.04	0.02	0.01	0.00
1	0.35	0.37	0.17	0.08	0.03	0.01
2	0.25	0.22	0.33	0.14	0.05	0.01
3	0.20	0.15	0.22	0.30	0.10	0.03
4	0.15	0.10	0.14	0.27	0.27	0.07
5	0.10	0.05	0.10	0.16	0.24	0.35

Table 3: Migraine data. Maximized log-pairwise likelihoods with $q = 12$, CLIC statistics and CLIC weights for various models fitted to the migraine data.

change	humidity	windchill	log-pair	CLIC	weights
-	-	-	-2935.16	5917.15	0.27
$*$	-	-	-2933.58	5916.30	0.41
-	$*$	-	-2934.36	5918.84	0.12
-	-	$*$	-2933.36	5922.53	0.02
$*$	$*$	-	-2932.90	5918.33	0.15
$*$	-	$*$	-2931.75	5921.93	0.02
-	$*$	$*$	-2932.62	5924.59	0.01
$*$	$*$	$*$	-2931.03	5924.20	0.01

Table 4: Migraine data. Estimates and standard errors from the pairwise likelihood with $q = 12$ for the base model (first two columns) and the best model accordingly to CLIC with different autocorrelation parameters for analgesic users and non-analgesic users (second two columns) and with a single common autocorrelation parameter (last two columns). The levels of the variable `change` are `1`: change from low to high atmospheric pressure, `2`: substantially unchanged atmospheric pressure, `3`: change from high to low atmospheric pressure. The baseline is "no university degree, no intake of analgesics, change from low to high pressure".

	est.	s.e.	est.	s.e.	est.	s. e.
α_2	0.588	0.046	0.588	0.046	0.589	0.046
α_3	1.136	0.069	1.136	0.069	1.137	0.069
α_4	1.786	0.079	1.787	0.080	1.788	0.080
α_5	2.505	0.109	2.506	0.111	2.508	0.112
intercept	−0.474	0.226	−0.522	0.223	-0.517	0.223
university	−0.523	0.172	−0.523	0.174	-0.525	0.173
analgesics	0.558	0.202	0.561	0.205	0.557	0.205
change2	—	—	0.031	0.051	0.031	0.051
change3	—	—	0.164	0.053	0.164	0.053
γ_F	0.415	0.094	0.424	0.094	0.540	0.031
γ_T	0.556	0.030	0.557	0.030	—	—
$\gamma_\mathrm{T} - \gamma_\mathrm{F}$	0.142	0.098	0.133	0.098	—	—
σ^2	0.566	0.110	0.564	0.111	0.566	0.112