Computational Statistics and Data Analysis 54 (2010) 1206–1218

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Bayesian estimation of random effects models for multivariate responses of mixed data☆

Helga Wagner[a,∗], Regina Tüchler[b]

[a] Department of Applied Statistics and Econometrics, Johannes Kepler Universität Linz, A-4040, Austria

[b] Austrian Federal Economic Chamber, Department of Statistics, Wien, Austria

ARTICLE INFO

Article history:
Received 25 November 2008
Received in revised form 11 August 2009
Accepted 15 December 2009
Available online 22 December 2009

Keywords:
Auxiliary mixture sampling
Generalized linear models
MCMC
Random effects model
Variable selection

ABSTRACT

A random effects model is presented to estimate multivariate data of mixed data types. Such data typically appear in studies where different response variables are measured repeatedly for one subject. It is possible to relate normal, binary, multinomial and count data by our joint model. Further flexibility with respect to model specification is obtained by including modern variable selection techniques. Auxiliary mixture sampling leads to a Gibbs sampling type scheme which is easy to implement since no additional tuning is needed. The method is illustrated by transaction data of a costumer cohort acquired by an apparel retailer.

1. Introduction

In this paper we model multidimensional data which arise when different response variables are measured repeatedly for one subject. Usually these responses are not of the same type but are measured on different scales, yielding mixed data with continuous and discrete outcomes. Since measurements are taken repeatedly over time on each subject under study not only dependencies between the response components but also within subject dependencies have to be taken into account. For repeated measurements of a single data type the usual approach is to use linear random effects models for normal or general random effects models for discrete data. However, combination of different data types to a joint model is a challenging problem. In the present paper we specify a random effects model which combines normal, binary, multinomial and count outcomes. We account for within subject dependencies by defining a random effects specification for the linear predictors of the single data types. These single response types are then linked by adding covariances between random effects of the different data types.

Such a general model for mixed data was not estimated in the literature before. This is mainly due to computational difficulties which arise when combining different data types. Clustered data of mixed type received attention in particular for a binary and a normal response component in the context of toxicity studies (Fitzmaurice and Laird, 1995; Catalano and Ryan, 1992; Regan and Catalano, 1999b,a). One approach is to model the joint distribution of both outcomes as the product of a marginal and a conditional distribution, see Cox and Wermuth (1992) for a discussion of different factorizations. Correlation of repeated measurements for one subject is taken into account in the marginal model as well as in the conditional model, estimation is accomplished by generalized estimation equations. The same type of approach is taken by Yang et al. (2007) for

☆ The views expressed in the paper are those of the authors and do not necessarily reflect those of the Austrian Federal Economic Chamber.

∗ Corresponding author.
E-mail addresses: helga.wagner@jku.at (H. Wagner), regina.tuechler@wko.at (R. Tüchler).

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.12.007

bivariate longitudinal data where one component is continuous and the other is Poisson count. Within subject correlation is taken into account for each response type by assuming a compound symmetry covariance matrix for observations of one subject. A different modeling approach, taken in Regan and Catalano (1999b); Gueorguieva and Agresti (2001) and Faes et al. (2008), is based on the interpretation of binary response as a dichotomization of an underlying normal variable and assuming a bivariate normal distribution for the normal response and the underlying normal variable. Correlation between the two responses and intra-cluster or within subject correlation can be taken into account by either explicit modeling of the covariance structure as in Regan and Catalano (1999b) or by a random effects specification where random effects and/or errors are assumed to follow a general bivariate normal distribution as in Gueorguieva and Agresti (2001). In principle this approach allows a full random effects specification for multivariate responses, however due to computational aspects so far researchers focused their work on simplified models. Faes et al. (2008) consider this problem in a classical setting and use pseudo-likelihood for joint estimation of all pairwise bivariate generalized linear mixed models.

In our paper we estimate a full random effects model. By using data augmentation we combine not only the normal responses but also the discrete ones to a linear model. The novel method of auxiliary mixture sampling then leads to a Gibbs sampling type scheme. Until recently Bayesian estimation of generalized linear models for categorical or count data was only possible if Metropolis–Hastings steps were included. Auxiliary mixture sampling for single data types was developed in Frühwirth-Schnatter and Wagner (2006) and Frühwirth-Schnatter et al. (2009) for Poisson counts, and in Frühwirth-Schnatter and Frühwirth (2007) for binomial and multinomial responses.

With many covariates at hand specification of random and fixed effects is a complicated problem. Recently variable selection tools are used to solve such model selection problems, see e.g. George and McCulloch (1997) for a description of the stochastic search variable approach, Smith and Kohn (2002) for covariance selection for normal data, and Frühwirth-Schnatter and Tüchler (2008) and Tüchler (2008) for covariance selection in normal and logistic random effects models, respectively. In our paper variable and covariance selection enable us to start with a very general model specification. All predictor variables at hand may be included and all effects may be specified as random effects. During the course of MCMC sampling those effects with zero means are detected and those effects which are fixed rather than random are restricted to fixed effects. Since the different data types are related through the variance–covariance matrix covariance selection also reveals whether such a relationship is present or not. If all covariances between effects of certain data types were selected as zero no relation between these data types would be present and the joint model would split into separate models.

The paper is structured as follows. In Section 2 we define the model. It is transformed into a Gaussian random effects model by auxiliary mixture sampling in Section 3.1 and variable and covariance selection is incorporated in Section 3.2. The prior and the simulation steps are described in Sections 3.3 and 3.4, respectively. The method is applied to simulated data in Section 4 and Section 5 gives a real-data example. Section 6 summarizes the results.

2. Random effects model for mixed data

Let $\mathbf{Y} = (Y^1, \dots, Y^K)'$ denote a multivariate response variable which is observed for $i = 1, \dots, N$ subjects on $t = 1, \dots, T_i$ occasions. The components Y^k, $k = 1, \dots, K$ may be either normal, binary, multinomial or Poisson counts. Let y_{it}^k denote the observation of the kth component measured for subject i at time point t, let $\mathbf{y}_i^k$ denote the sequence of T_i observations for the kth component of subject i, and let $\mathbf{y}_i$ summarize all T_iK observations of subject i.

To relate the mean $\mu_{it}^k = E(y_{it}^k)$ to the linear predictor η_{it}^k we introduce a distinct link function $g_k(\mu_{it}^k) = \eta_{it}^k, k = 1, \dots, K$ for each component depending on the type of the kth response component. For Poisson components we use the log-link function

$$\mu_{it}^k = \exp(\eta_{it}^k),$$

for binary components we consider the logit link function

$$\mu_{it}^k = \frac{\exp(\eta_{it}^k)}{1 + \exp(\eta_{it}^k)},$$

whereas for normal components y_{it}^k we use the identical link

$$\mu_{it}^k = \eta_{it}^k$$

and assume a constant variance $y_{it}^k \sim \mathcal{N}\left(\mu_{it}^k, \sigma_k^2\right)$.

We consider the following random effects specification for the linear predictors $\boldsymbol{\eta}_i^k$ of the sequence $\mathbf{y}_i^k$:

$$\boldsymbol{\eta}_i^k = \mathbf{X}_i \boldsymbol{\beta}_i^k.$$

$\mathbf{X}_i$ is a design matrix of dimension $T_i \times d$, where d equals the number of covariates in the model. $\boldsymbol{\beta}_i^k$ are normally distributed random effects. We assume that the same covariates are used for each of the K response components, whereas the random effects are allowed to differ between components.