

A finite mixture model for multivariate counts under endogenous selectivity

Marco Alfò · Antonello Maruotti · Giovanni Trovato

Received: 9 January 2009 / Accepted: 4 November 2009
© Springer Science+Business Media, LLC 2009

Abstract We describe a selection model for multivariate counts, where association between the *primary* outcomes and the endogenous selection source is modeled through outcome-specific latent effects which are assumed to be dependent across equations. Parametric specifications of this model already exist in the literature; in this paper, we show how model parameters can be estimated in a finite mixture context. This approach helps us to consider overdispersed counts, while allowing for multivariate association and endogeneity of the selection variable. In this context, attention is focused both on bias in estimated effects when exogeneity of selection (treatment) variable is assumed, as well as on consistent estimation of the association between the random effects in the primary and in the treatment effect models, when the latter is assumed endogeneous. The model behavior is investigated through a large scale simulation experiment. An empirical example on health care utilization data is provided.

Keywords Multivariate counts · Selection models · Random effects · Nonparametric ML · Finite mixtures

M. Alfò (✉)
Dipartimento di Statistica, Probabilità e Statistiche Applicate,
“Sapienza” Università di Roma, Piazzale A. Moro, 5,
00185 Rome, Italy
e-mail: marco.alfò@uniroma1.it

A. Maruotti
Dipartimento di Istituzioni Pubbliche, Economia e Società,
Università di Roma Tre, Via G. Chiabrera, 199, 00145 Rome,
Italy

G. Trovato
Dipartimento di Economia e Istituzioni, Università di Roma
“Tor Vergata”, Via Columbia, 2, 00133 Rome, Italy

1 Introduction

In this paper we discuss a mixed-effects model for multivariate counts when a binary selection variable is included in the linear predictor for the primary responses. If the selection variable depends on unobservable, individual characteristics which affect the primary outcomes also, it can not be considered *exogenous*, and dependence on unobserved individual characteristics should be taken into account to prevent *endogeneity bias*, see e.g. Davidson and Mackinnon (1993).

Relevant examples can be derived from empirical studies on the effect of private insurance coverage on health services utilization; one may get health insurance to protect financially himself/herself from an unexpected, substantial, medical bill due to a serious accident or a major illness. If health insurance through *group coverage* is not available, one could turn to private insurance plans. Individual choices for health insurance plans can influence patients' behavior and demand for health care services; in this view, private insurance choice is often considered as a covariate in regression models for utilization data, see Winkelmann (2000b) for a thorough review of the literature on the subject. But, the demand for health care services depends on observed socio-economic and health status and on additional, unobserved, factors that may also affect the insurance plan choice. Therefore, the choice for private insurance can not be considered as strictly exogenous.

In this paper, we propose a random effects model to account for extra-Poisson variation and dependence between the selection mechanism and the random effects in the outcomes equations, see Terza (1998) or Winkelmann (2000a). If the selection mechanism depends on unobservable heterogeneity sources influencing the (primary) counted outcomes, and estimation is based on the model for the primary outcomes only, parameter estimates for the selection variable

and for all variables depending on the selection, could be severely biased; Crouchley and Davies (1999) discuss this topic in the biostatistical context. In this case, parameter distinctiveness does not hold and the likelihood can not be factorized; therefore, we need to account for potential dependence of the selection variable on the random effects in the primary outcomes equations. An additional purpose of this work is to ascertain if correlation between random effects can be consistently estimated within this framework; analyzing bivariate binary responses, Smith and Moffatt (1999) found that the correlation parameter could not be estimated with enough precision, while Alfò and Trovato (2004) show different results in the bivariate count data context.

We start from the parametric model introduced by Munkin (2003) and Munkin and Trivedi (2003), where association among the outcomes and the selection variable is modeled through dependent outcome-specific random effects, with a known parametric joint distribution. Instead, we leave the random effects distribution $\mathcal{G}(\cdot)$ unspecified, and show how parameter estimation can be performed using a nonparametric ML (NPML) approach in a finite mixture context. This model can be also considered as an extension of the model described by Mroz (1999) to assess the impact of a dummy endogenous variable on a continuous outcome. It is worth noticing that an alternative method of estimation can be based on the proposal of Zimmer and Trivedi (2006), who discuss estimation of a trivariate model (two count equations plus one binary choice) using copulas. However, while computationally and theoretically appealing, their approach allows only for positive association between counts and the selection variable and the extension to cases where $J > 2$ counts are observed appears to be problematic.

The paper is structured as follows: in Sect. 2 we discuss model specification, while in Sect. 3 computational details of the adopted EM algorithm are given. Finite sample performance of the proposed model is investigated through a Monte Carlo study in Sect. 4, where attention is focused on empirical cases where $J \geq 2$ counts are considered. In Sect. 5 the effect of private health insurance on the demand for health care services is investigated through an empirical analysis of a well-known health care dataset. The last section contains concluding remarks and future research agenda.

2 Model specification

Regression models for multivariate counts have raised great interest in the last few years; reviews of this subject can be found in Cameron and Trivedi (2005) and Winkelmann (2000b). Loosely speaking, three main approaches have been proposed in this context: convolution models, see e.g.

Karlis and Meligkotsidou (2006, 2007), latent effect models, see e.g. Chib and Winkelmann (2001), Alfò and Trovato (2004), and copula-based models, see e.g. Harry (1997), Nikoloulopoulos and Karlis (2006), Zimmer and Trivedi (2006). We adopt the latent effects approach and define a set of conditional univariate models, linked by a common latent structure which accounts for both heterogeneity (in the univariate profiles) and dependence between counts. The latent effects approach may be motivated as follows: it is probably the only available with standard statistical packages, e.g. Stata or SAS, and EM-based maximization algorithms may be easily programmed using R or Matlab. Thus, while other approaches may be adopted as well, the latent effects approach does not impose any constraint on (explicit or implicit) model parameters and is probably the most widely used.

Let us suppose we have recorded counts Y_{ij} , $i = 1, \dots, n$ and $j = 1, \dots, J$, together with a set of m_j covariates $\mathbf{x}_{ij}^T = (x_{ij1}, \dots, x_{ijm_j})$ and a binary selection variable, say Y_{i0} . To describe association among counts, it is reasonable to assume that they share some common unobservable feature: let u_{ij} , $i = 1, \dots, n$, $j = 1, \dots, J$ denote a set of individual and outcome-specific random effects, accounting for heterogeneity and dependence between outcomes. Conditional on the covariates and the random effects, the observed counts Y_{ij} are independent Poisson random variables:

$$Y_{ij} | \lambda_{ij} \sim \text{Poisson}(\lambda_{ij}) \quad (1)$$

with canonical parameters $\lambda_{ij} = E(Y_{ij} | x_{ij}, y_{i0}, u_{ij})$ modeled as follows:

$$\begin{aligned} \log[E(Y_{ij} | x_{ij}, y_{i0}, u_{ij})] \\ = \beta_{0j} + \sum_{l=1}^{m_j} x_{ijl} \beta_{jl} + \alpha_j y_{i0} + u_{ij}. \end{aligned} \quad (2)$$

In this context, $\beta_j = (\beta_{j0}, \beta_{j1}, \dots, \beta_{jm_j})$ is an outcome-specific vector of fixed effects and α_j represents the effect of the selection variable on the j -th outcome, $j = 1, \dots, J$. The random effects $\mathbf{u}_i = (u_{i1}, \dots, u_{iJ})$ are drawn from a known multivariate parametric distribution, say $\mathcal{G}(\cdot)$, and account for extra-Poisson variation and marginal dependence among counts. Various alternatives have been proposed, including Poisson-Log-Normal, see e.g. Munkin and Trivedi (1999), and latent Poisson-Normal, see e.g. van Ophem (2000), two examples of multi-factor models. Uni-factor models have been proposed as well, for example the bivariate negative binomial of Winkelmann (2000c). Semiparametric multivariate alternatives with unspecified $\mathcal{G}(\cdot)$ are discussed in Alfò and Trovato (2004).

Given the model assumptions, the likelihood function conditional on the selection variable can be written as follows:

$$L(\cdot | y_{i0}) = \prod_{i=1}^n \left\{ \int_{\mathcal{U}} f(y_i | \mathbf{x}_i, y_{i0}, \mathbf{u}_i) d\mathcal{G}(\mathbf{u}_i | y_{i0}) \right\} \times \prod_{i=1}^n \left\{ \int_{\mathcal{U}} \prod_{j=1}^J f(y_{ij} | \mathbf{x}_{ij}, y_{i0}, u_{ij}) d\mathcal{G}(\mathbf{u}_i | y_{i0}) \right\}, \quad (3)$$

where \mathcal{U} denotes the support for $\mathcal{G}(\cdot)$. As can be easily noticed, the likelihood is defined integrating with respect to $\mathcal{G}(\mathbf{u}_i | y_{i0})$; if the selection variable is predetermined or, at least, independent of the random effects in the primary outcomes equations, we have that $\mathcal{G}(\mathbf{u}_i | y_{i0}) = \mathcal{G}(\mathbf{u}_i)$. Otherwise, the dependence between the random effects and the selection variable has to be taken explicitly into account; integrating with respect to $\mathcal{G}(\mathbf{u}_i)$ rather than with respect to $\mathcal{G}(\mathbf{u}_i | y_{i0})$ may produce biased and inconsistent estimates, as remarked by Crouchley and Davies (1999), Fotuhi (2005), Alfò and Aitkin (2006). This is usually known in the literature as endogeneity bias, see Davidson and Mackinnon (1993) for a thorough discussion of the topic.

To solve this problem, we may proceed by defining a *secondary* model for the selection variable, linked to the *primary* response model through a common latent structure. We assume that the selection variable Y_{i0} represents, conditional on a selection-specific latent effect, say u_{i0} , a Bernoulli random variable with canonical parameter described through the following mixed effect logistic model:

$$Y_{i0} | \mathbf{x}_{i0}, u_{i0} \sim \text{Bin}(1, p_i),$$

$$\text{logit}(p_i) = \beta_{00} + \sum_{l=1}^{m_0} x_{i0l} \beta_{0l} + u_{i0}, \quad i = 1, \dots, n.$$

Obviously, the restriction for Y_{i0} to be binary can be easily relaxed by considering a multinomial selection variable; this would simply increase the number of selection equations, given that a nominal variable with G categories may be represented by $(G - 1)$ binary (indicator) variables, with a log-ratio link representation. We will describe only the cross sectional case, since handling panel data would represent a straightforward extension; in fact, when longitudinal measurements are handled, random effects account for overdispersion in the univariate profiles, dependence among outcomes and among repeated measures on the same outcome; the proposed approach can be extended to primary outcomes of mixed type as well, if identifiability conditions are fulfilled. Given the model assumptions, the likelihood function can be written as:

$$L(\cdot) = \prod_{i=1}^n \left\{ \int_{\mathcal{U}} \left[f(y_{i0} | \mathbf{x}_{i0}, u_{i0}) \times \prod_{j=1}^J f(y_{ij} | \mathbf{x}_{ij}, y_{i0}, u_{ij}) \right] d\mathcal{G}(\mathbf{u}_i) \right\}. \quad (4)$$

For Gaussian assumptions on \mathbf{u}_i , the marginal likelihood can not be written in closed form; to obtain ML estimates, we may choose among several alternatives. We may adopt numerical integration techniques based on standard or adaptive Gaussian Quadrature (GQ, AGQ), for a comparison see Rabe-Heskett and Skrondal (2002); however, the corresponding estimation algorithms can be cumbersome when the number of outcomes increases and thus can be slow to achieve convergence (this point will be further discussed). A further alternative is to rely on Monte Carlo or simulation-based techniques, see e.g. McCulloch (1994), Chib and Winkelmann (2001), Geyer and Thompson (1992), Munkin and Trivedi (1999). The latter approach is known to be inefficient for non optimal importance samplers; furthermore, the conditional distribution of the random effects given the observed data and the current parameter estimates can be quite difficult to sample from. In addition, simulation, see e.g. McCulloch (1997), as well as analytical investigations, see e.g. Jank and Booth (2003), found SML estimation to perform poorly relative to Monte Carlo (EM) ML estimation. Marginal maximization procedures using Gaussian quadrature or Monte Carlo approximations can be computationally very intensive, as noted in different contexts by Crouch and Spiegelman (1990) and Gueorguieva (2001), since the number of quadrature points/simulation draws need to be significantly increased. As a general matter, the assumption of multivariate normality for the latent effects can be too restrictive and is actually unverifiable; parametric alternatives could result in oversmoothing, masking individual discontinuities when the mixing distribution is defined over a finite number of well-separated groups. From this perspective, a more appealing approach is to leave $\mathcal{G}(\cdot)$ unspecified, and rely on the theory of NPML, see e.g. Kiefer and Wolfowitz (1956), Laird (1978), Heckman and Singer (1984). Using a simple geometric result, Lindsay (1983a, 1983b) showed that finding the MLE involves a standard optimization problem, the maximization of a concave function over a convex set. As long as the likelihood is bounded, it is maximized with respect to $\mathcal{G}(\cdot)$ by at least a discrete distribution $\mathcal{G}_K(\cdot)$ with at most $K \leq n$ support points. Thus, moreover, fitting a finite mixture model can be considered as an additional tool to perform a sensitivity analysis for a standard, parametric, random effect model. Let us suppose that $\mathcal{G}_K(\cdot)$ puts masses π_k on locations $\mathbf{u}_k = (u_{k0}, u_{k1}, \dots, u_{kJ})$, $k = 1, \dots, K$. The resulting likelihood function is:

$$L(\cdot) = \prod_{i=1}^n \left\{ \sum_{k=1}^K \pi_k \left[f(y_{i0} | \mathbf{x}_{i0}, u_{k0}) \times \prod_{j=1}^J f(y_{ij} | \mathbf{x}_{ij}, y_{i0}, u_{kj}) \right] \right\}, \quad (5)$$

where $\pi_k = \Pr(\mathbf{u}_i = \mathbf{u}_k) = \Pr(u_{k0}, u_{k1}, \dots, u_{kj})$, $k = 1, \dots, K$, represent the prior probabilities of locations \mathbf{u}_k . Similar finite mixture models have been successfully used to model overdispersed counts and may outperform mixed models based on parametric specifications, particularly when extreme (very high or very low) intensity users are present, see e.g. Deb and Trivedi (1997), Deb and Holmes (2000), Alfö and Trovato (2004). Finite mixture models have several significant advantages over parametric mixture models; first, the discrete nature of the estimate helps to classify subjects in clusters characterized by homogeneous values of random parameters. This is particularly appealing in health sciences, where components can be interpreted as groups with similar behavior with respect to the event of interest. Second, the locations and the corresponding probabilities are completely free to vary over the corresponding support, and therefore, extreme departures from the basic (homogeneous) model can be accommodated. Estimation may be accomplished by a standard EM algorithm; the elements of \mathbf{u}_k can be estimated by introducing in the linear predictor the interaction between a K -level factor and a set of outcome-specific dummies d_{ij} , $i = 1, \dots, n$, $j = 0, 1, \dots, J$, indicating which outcome is fitted. A standard EM algorithm for finite mixtures applies, with the substantial change that the posterior weights w_{ik} are computed by considering the joint distribution of the counted responses and the selection variable(s).

3 Computational details

Given the finite mixture representation above, each unit can be conceptualized as drawn from one of K distinct groups: we denote with π_k the prior probability that a unit belongs to the k -th component and with z_{ik} the indicator variable which is equal to 1 if the i -th unit belongs to the k -th component of the finite mixture. The data vector is composed by an observable part, \mathbf{y}_i , and by an unobservable one, the membership vector, which is denoted by $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})$. If the vector \mathbf{z}_i is known, the proposed model reduces to a simple multivariate nonlinear regression model with group-specific intercepts; since the \mathbf{z}_i are unknown, we may use an EM-type algorithm. Various authors have discussed algorithms for joint estimation of K and model parameters, such as the *Vertex Exchange* or the *Vertex Direction* methods: for an insightful discussion see Böhning (2000). A possible solution is to update estimates for a fixed K improving step by step, as in the EM algorithm with gradient function update, see Böhning (2003). We follow the latter approach: the algorithm is run for fixed K and reaches a solution which is used to estimate model parameters when $K + 1$ components

are considered. Using this approach, we reduce to a *standard* ML estimation problem where the solution obtained for K is compared with that obtained for $(K + 1)$ components. Karlis and Meligkotsidou (2007) have employed a large-scale simulation study to investigate relative performance of penalized likelihood criteria, while Feng and McCulloch (1996) discuss how hypothesis testing can be performed by bootstrapping the likelihood ratio test.

For fixed K the following routine for parameter estimation is followed: we assume that the unobservable vector \mathbf{z}_i is a multinomial random variable with weights $\pi_k = \Pr(\mathbf{u}_i = \mathbf{u}_k)$, $i = 1, \dots, n$, $k = 1, \dots, K$. Given the model assumptions, the likelihood for the complete data $(\mathbf{y}_i, \mathbf{z}_i)$, $i = 1, \dots, n$, is given by:

$$L_c(\cdot) = \prod_{i=1}^n \prod_{k=1}^K (\pi_k f_{ik})^{z_{ik}} = \prod_{i=1}^n \prod_{k=1}^K \left[\pi_k \left(f_{i0k} \prod_{j=1}^J f_{ijk} \right) \right]^{z_{ik}}. \quad (6)$$

The corresponding complete data log-likelihood function is therefore defined by:

$$\ell_c(\cdot) = \sum_{i=1}^n \sum_{k=1}^K z_{ik} \{ \log(\pi_k) + \log f_{ik} \} = \sum_{i=1}^n \sum_{k=1}^K z_{ik} \left\{ \log(\pi_k) + \log(f_{i0k}) + \sum_{j=1}^J \log(f_{ijk}) \right\}, \quad (7)$$

where $\pi_k = \Pr(\mathbf{u}_k)$. Within the E-step, the presence of missing data is handled by taking the conditional expectation of the log-likelihood for the complete data given the observed data \mathbf{y}_i and the current ML parameter estimates, say $\boldsymbol{\theta}^{(r)} = \{\boldsymbol{\delta}^{(r)}, \boldsymbol{\pi}^{(r)}\} = \{\boldsymbol{\beta}^{(r)}, \boldsymbol{\alpha}^{(r)}, \mathbf{u}^{(r)}, \boldsymbol{\pi}^{(r)}\}$. In other words, we replace z_{ik} with its conditional expectation:

$$E(z_{ik} | \mathbf{y}_i, \boldsymbol{\theta}^{(r)}) = w_{ik}^{(r+1)} = \frac{\pi_k f_{ik}}{\sum_{l=1}^K \pi_l f_{il}}, \quad (8)$$

which is the posterior probability that the i -th unit belongs to the k -th component of the finite mixture, $i = 1, \dots, n$, $k = 1, \dots, K$. The conditional expectation of the complete log-likelihood function given the data vector \mathbf{y}_i and the current parameter estimates, $\boldsymbol{\theta}^{(r)}$, is

$$Q(\cdot) = E_{\boldsymbol{\theta}^{(r)}} \{ \ell_c(\cdot) | \mathbf{y}_i \} = \sum_{i=1}^n \sum_{k=1}^K w_{ik}^{(r+1)} \{ \log(\pi_k) + \log f_{ik} \}$$

$$= \sum_{i=1}^n \sum_{k=1}^K w_{ik}^{(r+1)} \left\{ \log(\pi_k) + \log(f_{i0k}) + \sum_{j=1}^J \log(f_{ijk}) \right\}, \tag{9}$$

which corresponds to a finite mixture of K multivariate ($J + 1$ -dimensional) distributions with common weights $w_{ik}^{(r+1)}$. Conditional on updated weights $w_{ik}^{(r)}$, we maximize $Q(\cdot)$ with respect to θ to obtain updated ML parameter estimates $\hat{\theta}^{(r+1)}$. The estimated parameters are the solution of the following M-step equations:

$$\frac{\partial Q}{\partial \pi_l} = \sum_{i=1}^n \left\{ \frac{w_{il}}{\pi_l} - \frac{w_{iK}}{\pi_K} \right\} = 0, \tag{10}$$

$$\frac{\partial Q}{\partial \delta} = \sum_{i=1}^n \sum_{k=1}^K w_{ik} \frac{\partial}{\partial \delta} \left[\log(f_{i0k}) + \sum_{j=1}^J \log(f_{ijk}) \right]. \tag{11}$$

Solving the first equation we obtain:

$$\hat{\pi}_k^{(r+1)} = \sum_{i=1}^n \frac{w_{ik}^{(r+1)}}{n}, \tag{12}$$

which represents a well known result from ML in finite mixtures. Given $w_{ik}^{(r+1)}$, score equations in (11) are weighted sums of likelihood equations for $J + 1$ independent GLMs; the standard EM algorithm for finite mixtures of univariate distributions applies, where w_{ik} are computed considering the joint distribution of the counted responses and the selection variable(s). Since closed form solutions of equations (11) are unavailable, we use a standard Newton-Raphson algorithm. The E- and M-steps are repeatedly alternated until the log-likelihood relative difference changes by an arbitrarily small amount. Given that the log likelihood function increases, at the $r + 1$ -th step of the algorithm we have $\ell^{(r+1)} \geq \ell^{(r)}$, $r = 0, 1, 2, \dots$; convergence is obtained with a sequence of likelihood values which are bounded from above.

As a prerequisite for ML estimation, we assume that the mixture is identifiable, i.e. that two sets of parameters which do not agree after permutation cannot yield the same mixture distribution. A sufficient condition for identifiability in the context of Poisson regression models is that the covariate matrix \mathbf{X} is full rank, see Wang et al. (1996). Finite mixtures of multivariate counts can be considered as mixtures of univariate counts with outcome-specific random effects, and are, therefore, identifiable provided the conditions above are satisfied. On the other side, Teicher (1963) shows that if $Y_i | p_i \sim \text{Binomial}(m, p_i)$, $p_i \sim \mathcal{G}(\cdot)$, then $\mathcal{G}(\cdot)$ can be uniquely identified from the marginal density provided it has no more than $(m + 1)/2$ components; thus, finite mixtures

of Bernoulli distributions are not identifiable. However, the result entails finite mixtures of Bernoulli distributions, but does not apply to the regression setting. Follmann and Lambert (1991) discuss identifiability of $\mathcal{G}(\cdot)$ in random effects models for binary responses and give simple bounds for the number of components.

To explain this point, let us choose a generic individual vector, say the h -th, \mathbf{x}_h , $h = 1, \dots, n$ and define S_{hs} as the set of all vectors, \mathbf{x}_i , $i = 1, \dots, n$, $i \neq h$, which agree with the h -th for all but the s -th coordinate, $s = 1, \dots, p$. Let us denote by N_{hs} the number of distinct elements in this set, and the corresponding maximum value by $N^* = \max_{h,s} N_{hs}$. Theorem 2 of Follmann and Lambert (1991) guarantees identifiability of a mixing distribution \mathcal{G} with K components bounding K by a function of N^* , $K \leq \sqrt{N^* + 2} - 1$. To get further insight, let us define the (conditional) probabilities:

$$\begin{aligned} \pi_{k,1} &= \Pr(\mathbf{u}_{i \setminus 0} = \mathbf{u}_{k \setminus 0} \mid u_{i0} = u_{k0}) \\ &= \Pr[(u_{i1}, \dots, u_{iJ}) = (u_{k1}, \dots, u_{kJ}) \mid u_{i0} = u_{k0}], \end{aligned}$$

$$\pi_{k,2} = \Pr(u_{i0} = u_{k0}),$$

where $\mathbf{u}_{i \setminus 0}$ and u_{i0} are the random effects associated to the counted outcomes and the selection process, respectively. Using such quantities, we may factorize the mixing weights in (7) as follows:

$$\begin{aligned} \pi_k &= \Pr(\mathbf{u}_i = \mathbf{u}_k) \\ &= \Pr(\mathbf{u}_{i \setminus 0} = \mathbf{u}_{k \setminus 0} \mid u_{i0} = u_{k0}) \Pr(u_{i0} = u_{k0}). \end{aligned}$$

The log-likelihood function for *complete* data in (7) can therefore be factorized accordingly:

$$\begin{aligned} \ell_c(\cdot) &= \sum_{i=1}^n \sum_{k=1}^K z_{ik} \left\{ \log(\pi_k) + \log(f_{i0k}) + \sum_{j=1}^J \log(f_{ijk}) \right\} \\ &= \sum_{i=1}^n \sum_{k=1}^K z_{ik} \left\{ \left[\log(\pi_{k,1}) + \sum_{j=1}^J \log(f_{ijk}) \right] \right. \\ &\quad \left. + [\log(\pi_{k,2}) + \log(f_{i0k})] \right\}. \end{aligned}$$

We may notice that the first term represents the contribution to the log-likelihood of a mixture of Poisson distributions (conditional on u_{i0} , priors given by $\pi_{k,1}$), while the second represents the contribution to the loglikelihood of a mixture of logistic models (priors $\pi_{k,2}$). Therefore, provided that conditions given above are satisfied, the proposed mixture is identifiable.

Since results obtained using the EM algorithm tend to be very sensitive to initial parameter values, we started the algorithm from the estimates obtained through a short-length

CEM (Classification EM) algorithm as suggested in Biernacki et al. (2003). Empirical experience confirms that the use of CEM-based starting values substantially enhances the speed of the corresponding EM algorithm; consequently, a higher number of EM solutions can be attained and a more thorough search over the likelihood surface can be performed, likely leading to (nearly) optimal results. In the simulation section, we retained, according to Keribin (2000), the model corresponding to the best BIC value; this often represents a conservative rule, which has been adopted considering that, in the present context, random effects represent nuisance parameters. We aim at retaining the more parsimonious model rather than the best estimate for $\mathcal{G}(\cdot)$. Other alternatives are however possible, if a more refined estimate of the random effects distribution is needed; for example, Karlis and Meligkotsidou (2007) found that AIC performs best among various criteria in terms of selecting the *right* number of components in finite mixtures of multivariate Poisson models.

4 An empirical application to health care utilization

In this section the proposed model is used to study the effect of private insurance plans on health care utilization data when strong controls for bias are imposed. In particular, utilization is measured through counts of event such as the number of admissions to medical facilities (hospitals, emergency rooms, etc.).

We use data from the National Medical Expenditure Survey (NMES), previously analyzed by Deb and Trivedi (1997) and Munkin and Trivedi (1999). The NMES has been conducted in 1987 and 1988 to provide a comprehensive picture of how Americans use and pay for health services. In addition to health care data, the NMES provides information on (mainly self-perceived) health status, employment and other social, demographic and economic characteristics.

We consider individuals aged 66 and over (4406 observations) all of whom are covered by Medicare, a public insurance programme that bothers substantial protection against health care costs. We focus on a joint model for two responses, the number of visits to an emergency room (*EMR*) and the number of hospital stays (*HOSP*). As one might expect, most elderly see a physician at least once a year and a considerably small fraction enters an emergency room or is admitted to a hospital. Hence, an important feature of these data is that they include a high proportion of zero counts, corresponding to a null demand of medical services, and a high degree of marginal overdispersion. Thus, some kind of mixture model is needed to provide a good fit to the data. The adopted covariates and responses are defined in Table 1.

A further modeling challenge come from the potential endogeneity of health insurance. Several studies in the health economics literature allow for endogeneity, see e.g. Mello et al. (2002), Deb et al. (2006), Munkin and Trivedi (2008). A widely held perception in the health economic literature is

Table 1 Variable definitions and summary statistics

<i>Variable</i>	<i>Definition</i>	<i>Mean</i>	<i>St. Dev.</i>
<i>EMR</i>	Number of emergency rooms visits	0.26	0.70
<i>HOSP</i>	Number of hospitalizations	0.30	0.75
<i>EXCLHLTH</i>	1 if self-perceived health is excellent	0.08	0.27
<i>POORHLTH</i>	1 if self-perceived health is poor	0.13	0.33
<i>NUMCHRON</i>	Number of chronic conditions	1.54	1.35
<i>ADLDIFF</i>	1 if the personal condition limits daily life	0.20	0.40
<i>NOREAST</i>	1 if the person lives in Northeastern USA	0.19	0.39
<i>MIDWEST</i>	1 if the person lives in Midwestern USA	0.26	0.44
<i>WEST</i>	1 if the person lives in Western USA	0.18	0.39
<i>AGE</i>	Age in years (divided by 10)	7.44	0.63
<i>BLACK</i>	1 if African American	0.12	0.32
<i>MALE</i>	1 if male	0.40	0.49
<i>MARRIED</i>	1 if married	0.55	0.50
<i>SCHOOL</i>	Number of years of education	10.3	3.74
<i>FAMINC</i>	Family income in \$10,000	2.53	2.92
<i>EMPLOYED</i>	1 if employed	0.10	0.30
<i>PRIVINS</i>	1 if covered by private health insurance	0.78	0.42
<i>MEDICAID</i>	1 if covered by MEDICAID	0.09	0.29

that healthier individuals tend to select themselves into managed care plans with a gatekeeper and smaller premiums; while less health but more risk-adverse individuals tend to select indemnity plans with higher premiums and more extensive coverage.

According to these studies, we fitted a multivariate mixed model dealing with both heterogeneity sources and endogeneity.

Ignoring selection effects means that we cannot separate out the pure treatment effect from the one due to self-selection. Individuals and households are more likely to choose insurance based on personal characteristics such as overall health status, severity of chronic health conditions and physical limitations, preferences for risk, preferences over intensity of treatment. If all such variables are introduced into the outcome equation, then one could control for the effects of selection, but this is difficult because some of these factors are intrinsically unobservable. Hence the observable variables included in the outcome equation may not help to adequately control for the influence of these factors, and, likely, some additional statistical controls for selection on unobservables will be required.

We start our analysis by fitting a random effects regression model for the bivariate outcomes; overdispersion and dependence between responses come through a bivariate la-

tent variable which is additively inserted in the linear predictor. We assume that the responses are conditionally independent Poisson variates given the covariates (including the selection variable PRIVINS, indicating if a subject is covered by a private health insurance plan) and the latent effects, (u_{i1}, u_{i2}) . The inclusion of individual- and outcome-specific random effects implies correlation across equations. In this study, unobserved heterogeneity, modeled via a random effects model, may represent chronic or permanent health conditions that make individuals less likely to join insurance plan and more likely to have some hospital and physician use. These unobserved factors are constant across equation for an individual and also over time since the health condition is chronic in nature and affect both the outcome and the insurance choice variables.

This model is a *constrained* version of the *full* model, since the insurance choice is treated as exogeneous. Estimates obtained by using a bivariate Poisson log-Normal model, see Munkin and Trivedi (1999), are shown in Table 2, while parameter estimates obtained using the semi-parametric approach developed by Alfd and Trovato (2004) are displayed in Table 3; in both cases parameter estimates for the mixed logit modeling private insurance are shown. Since the effect of private insurance, assuming endogeneity can not be identified without introducing legitimate ex-

Table 2 Estimates of the Bivariate Poisson—lognormal model. Exogenous selection. *Variables significant at 5% level

VARIABLE	EMR		HOSP		PRIVINS	
	Coef.	s.e.	Coef.	s.e.		
CONSTANT	-3.734*	0.475	-4.237*	0.462		
EXCLHLTH	-0.596*	0.205	-0.696*	0.203		
POORHLTH	0.526*	0.105	0.531*	0.103		
NUMCHRON	0.249*	0.028	0.282*	0.027		
ADLDIFF	0.434*	0.096	0.337*	0.094		
NOREAST					0.1994	0.1140
MIDWEST					0.5387*	0.1097
WEST					-0.2494*	0.1163
AGE	0.120*	0.061	0.180*	0.059		
BLACK					-1.3871*	0.1088
MALE					-0.0915	0.0925
MARRIED					0.6324*	0.0923
SCHOOL					0.1732*	0.0115
FAMINC					0.1138*	0.0245
EMPLOYED					0.2581	0.1572
PRIVINS	0.002	0.106	0.205*	0.097		
MEDICAID	0.199	0.139	0.172	0.140		
σ^2		1.6402		1.6452		0.0000
$\rho_{u_{i1}, u_{i2}}$						0.9999
log-likelihood						-5247.4090
						-1917.7417

Table 3 Nonparametric Estimates of Bivariate Poisson—Correlated Random Effects. Exogenous selection. *Variables significant at 5% level

VARIABLE	EMR		HOSP		PRIVINS	
	Coef.	s.e.	Coef.	s.e.	Coef.	s.e.
CONSTANT	−3.876*	0.522	−4.206*	0.472	−1.426*	0.280
EXCLHLTH	−0.598*	0.197	−0.696*	0.195		
POORHLTH	0.458*	0.102	0.465*	0.101		
NUMCHRON	0.211*	0.026	0.246*	0.025		
ADLDIFF	0.397*	0.102	0.300*	0.103		
NOREAST					0.272	0.147
MIDWEST					0.710*	0.159
WEST					−0.287*	0.148
AGE	0.101	0.058	0.169*	0.057		
BLACK					−1.566*	0.145
MALE					−0.071	0.119
MARRIED					0.760*	0.134
SCHOOL					0.212*	0.021
FAMINC					0.214*	0.065
EMPLOYED					0.463	0.246
PRIVINS	−0.044	0.102	0.162*	0.081		
MEDICAID	0.100	0.131	0.080	0.133		
σ^2		2.803		2.03		1.772
$\rho_{u_{i1}, u_{i2}}$		0.9976				
log-likelihood		−5232.962				−1914.616

clusion restrictions, we separate variables affecting the insurance choice but not the outcomes. In fact, if one is interested in the effect of an endogenous *treatment variable* on an outcome variable, one may perform causal inference by exploiting the presence of variables that causally affect the treatment status but do not have direct causal effect on the primary outcome. This means that empirical results critically depend on the validity of the exclusion restrictions; being identifying restrictions, they can not be tested.

In the exogenous case, the $\hat{G}(\cdot)$ estimate is a three-point distribution which puts masses (0.5654, 0.0577, 0.3778) on locations (−1.4503, −1.2233), (2.9583, 2.7814) and (1.7155, 1.4033), where the coordinates correspond to the EMR and HOSP equations, respectively.

It is worth noticing that the first location corresponds to low frequencies which represent people who either do not use health care services or use them very rarely, while the second and third locations correspond, respectively, to very high and high frequencies of health care utilization. The correlation estimate is equal to 0.9976; this findings confirm that unobservable heterogeneity sources influencing the counted outcomes are closely related and, further, that a shared latent structure for both health care utilization outcomes may be taken into account. Looking at parameter estimates, we may observe that parametric and semiparamet-

ric models strongly agree. Individual characteristics have a direct effect on the overall probability of choosing private insurance as well as on health care utilization. Education levels are significant (in both parametric and nonparametric latent effects models), with higher levels of attainment increasing the probability of being insured. While gender is not a significant determinant of the demand for insurance, older respondents have a higher propensity to use health care services. Being single lowers the probability of choosing a private insurance scheme, while a higher income is associated to increased probability of choosing private insurance. Self-reported health status variables show that poorer self-reported health levels are significant and increase the probability of using a health care service. The insurance has a significant effect only on the number of hospital stays, while its effect is negligible when the entrance to emergency room is considered. However, these findings could be due to the implicit assumption of strict exogeneity for the selection variable.

For this reason, we define a model for PRIVINS and proceed considering the choice for a private insurance plan as endogenous. We assume that PRIVINS is, conditionally on the covariates and on an outcome-specific random effect, u_{i0} , a Bernoulli random variable; furthermore, u_{i0} is supposed to be dependent on the random effects in the primary

Table 4 Parametric Parameter Estimates for the Bivariate Poisson Model with Endogenous Selectivity

VARIABLE	EMR		HOSP		PRIVINS	
	Coef.	s.e.	Coef.	s.e.	Coef.	s.e.
<i>COSTANT</i>	-3.7522*	0.5088	-4.2525*	0.4967	-0.8408*	0.1303
<i>EXCLHLTH</i>	-0.5967*	0.2055	-0.6961*	0.2033		
<i>POORHLTH</i>	0.5265*	0.1060	0.5316*	0.1033		
<i>NUMCHRON</i>	0.2491*	0.0280	0.2821*	0.0271		
<i>ADLDIFF</i>	0.4344*	0.0967	0.3375*	0.0944		
<i>NOREAST</i>					0.1997*	0.1140
<i>MIDWEST</i>					0.5390*	0.1098
<i>WEST</i>					-0.2487*	0.1167
<i>AGE</i>	0.1209*	0.0615	0.1805*	0.0596		
<i>BLACK</i>					-1.3871*	0.1088
<i>MALE</i>					-0.0911	0.0926
<i>MARRIED</i>					0.6322*	0.0923
<i>SCHOOL</i>					0.1732*	0.0115
<i>FAMINC</i>					0.1138*	0.0245
<i>EMPLOYED</i>					0.2588	0.1574
<i>PRIVINS</i>	0.0173	0.2051	0.2208	0.2053		
<i>MEDICAID</i>	0.2003	0.1402	0.1785	0.1412		
$\rho(u_{i1}, u_{i2}), u_{i3}$			-1.0000			
σ^2	1.6412		1.6412		1.6735	
log-likelihood			-7165.1801			

outcome equations. However, as remarked above, the bivariate model for EMR and HOSP suggests to assume that the random effects are shared by the primary outcomes. In the following, we will refer to this model specification, also known as one-factor model, Winkelmann (2000c), where correlation among outcomes arises since all the outcomes measured on the i -th individual share the same random term, which account for individual extra-Poisson variation. In this setting, PRIVINS is considered endogenous: Tables 4 and 5 show results of the parametric and semiparametric approaches, respectively. As can be noticed, the corresponding point estimates in the primary outcomes equation show changes in both magnitude and significance. In particular, the choice for private insurance is no longer significant in the utilization equations, either for EMR or for HOSP.

Adopting the semiparametric approach described above, the estimate \hat{G} is, still, a three-point distribution with masses (0.5556, 0.0611, 0.3833) on locations (-1.3666, 1.3851), (2.8542, -3.1722), (1.5258, -1.5019) for the random effects in the outcomes (shared random effects) and PRIVINS equations, respectively.

The first location corresponds to people who have a high propensity to subscribe private insurance plans but use less frequently health services. This suggests that people with higher health and social status subscribe more often individ-

ual health insurance plans, but use less frequently health care resources. The second component groups individuals with a very low propensity to subscribe health insurance plans, but with higher health needs; lower levels of health status are, in fact, considerably more diffuse amongst people without insurance than among those with a private insurance. The third component is characterized by a reduced, but not extreme, propensity to health insurance and a positive propensity to utilization of hospital resources. The random effect in the private insurance equation is negatively correlated with those in the utilization equations. While the random terms u_{i1} and u_{i2} in the health care equations show the impact of *latent* health needs on the demand for health services with increased utilization in component 2 and 3, the random effect u_{i0} seems to represent the influence of this *latent* health status on the choice for private health insurance with increased probability in component 1.

Usually, if high risk individuals purchase more insurance (adverse selection) the effect of insurance should actually go down, once this type of endogeneity is accounted for. The strong and negative correlation between the different sources of heterogeneity suggests something else. Individuals with higher health status show a higher propensity to subscribe health insurances than those with lower status, regardless of their health needs.

Table 5 Nonparametric Parameter Estimates for the Bivariate Poisson Model with Endogenous Selectivity

VARIABLE	EMR		HOSP		PRIVINS	
	Coef.	s.e.	Coef.	s.e.	Coef.	s.e.
COSTANT	-3.9310*	0.5660	-4.4465*	0.5572	-0.8421*	0.1305
EXCLHLTH	-0.6047*	0.1973	-0.7014*	0.1946		
POORHLTH	0.4618*	0.1027	0.4662*	0.0997		
NUMCHRON	0.2108*	0.0263	0.2442*	0.0253		
ADLDIFF	0.4070*	0.0947	0.3068*	0.0929		
NOREAST					0.2013	0.1142
MIDWEST					0.5418*	0.1101
WEST					-0.2441*	0.1172
AGE	0.1130	0.0594	0.1748*	0.0576		
BLACK					-1.3870*	0.1090
MALE					-0.0880	0.0929
MARRIED					0.6328*	0.0924
SCHOOL					0.1731*	0.0115
FAMINC					0.1140*	0.0246
EMPLOYED					0.2629	0.1578
PRIVINS	0.0576	0.2424	0.2613	0.2412		
MEDICAID	0.1034	0.1318	0.0781	0.1318		
$\rho(u_{i1}, u_{i2}), u_{i3}$			-0.9983			
σ^2	2.4278		2.4278		2.5455	
log-likelihood			-7151.4870			

We computed standard errors of parameter estimates using parametric bootstrap, as standard errors based on the observed information matrix are often unstable. That is, we re-fitted the model to bootstrap samples drawn from the estimated model. This process was repeated 250 times.

The choice of the number of components in the finite mixture is based on the BIC criterion; as well as from a formal LRT (performed conditional on \hat{K}), the null hypothesis of exogeneity of the selection variable ($\rho_{13} = \rho_{23} = 0$) would be rejected, suggesting the choice of the unconstrained model where the selection variable is formally treated as endogenous. We can conclude that ignoring the selection variable is endogenous would have led to biased parameter estimates for all those effects which depend on it.

As can be easily observed, the estimated correlation between the random effects in the outcome and in the insurance equations is close to unit in magnitude, which is the boundary of the parameter space. This is true for both parametric and semiparametric case, even if in the latter case the problem is more serious. In fact, when a discrete mixing distribution is adopted, the correlation coefficient is estimated only on a small (K) number of points and when ρ is high this can lead to a set of K points that are almost aligned. Thus,

when correlation is high, parameter estimates tend towards the bounds of the interval $(-1, 1)$.

5 Simulation study

To investigate the empirical behavior of the proposed model, we have conducted the following simulation study. A dataset is generated according to an endogenous selection model, with multiple counted outcomes; to model dependence between the outcomes and the selection process, the random effects in the corresponding equations have been drawn from a multivariate normal density with non-diagonal covariance matrix. The aim of this study is to investigate if ignoring endogeneity of the selection variable could lead to biased estimates for the corresponding effect as well as for the effects of those covariates which are correlated, to some extent, with the selection variable.

We have generated $B = 250$ samples of size $n = 250$, 1000, 2000 each, according to the following scheme:

$$Y_{ij} | \lambda_{ij} \sim \text{Poisson}(\lambda_{ij}), \quad j = 1, \dots, J$$

for $J = 2, 3, 4$ counts and

$$Y_{i0} | p_i \sim \text{Bin}(1, p_i)$$

for the selection variable. The following regression models hold:

$$\log(\lambda_{ij}) = \beta_{j0} + \beta_{j1}x_{i1} + \beta_{j2}y_{i0} + u_{ij},$$

$$j = 1, \dots, J, \quad (13)$$

$$\text{logit}(p_i) = \beta_{00} + \beta_{01}x_{i2} + u_{i0}.$$

The covariates have been drawn from a standard Gaussian density, while the correlated latent effects \mathbf{u}_i have been drawn from a multivariate normal density, $\text{MVN}(\mathbf{0}, \mathbf{\Sigma})$. For ease of discussion, the elements in $\mathbf{\Sigma}$ are imposed to be equal to:

$$\sigma_j^2 = 1, \quad \sigma_{jj'} = \rho, \quad j \neq j' = 0, \dots, J; \quad \rho = \{0.5; 0.75\}$$

therefore, the random effects are dependent and the selection variable can not be considered exogeneous. The choice of a common covariance term, denoted by ρ , is motivated by our interest in analyzing the behavior of parameter estimates when correlation approaches 1; in this context, the reliable estimation of the correlation between random effects in the primary and selection equations may be of interest *per se*. We assume the following *true* values for the parameter vectors:

$$\beta_0 = \begin{bmatrix} \beta_{00} \\ \beta_{01} \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}, \quad \beta_1 = \begin{bmatrix} \beta_{10} \\ \beta_{11} \\ \beta_{12} \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.75 \\ 1 \end{bmatrix},$$

$$\beta_2 = \begin{bmatrix} \beta_{20} \\ \beta_{21} \\ \beta_{22} \end{bmatrix} = \begin{bmatrix} 0.75 \\ 0.5 \\ 1 \end{bmatrix}, \quad \beta_3 = \begin{bmatrix} \beta_{30} \\ \beta_{31} \\ \beta_{32} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0.6 \end{bmatrix},$$

$$\beta_4 = \begin{bmatrix} \beta_{40} \\ \beta_{41} \\ \beta_{42} \end{bmatrix} = \begin{bmatrix} 1.5 \\ 0.6 \\ 0.5 \end{bmatrix}.$$

Thus, β_0 is the fixed effect vector for the selection model, while the effect of the selection variable on the j -th count is denoted by β_{j2} , $j = 1, \dots, J$. We fitted the proposed model using the algorithm described in Sect. 3 for varying number of components, $K = 2, \dots, 10$. In each case, the model with the best BIC value has been retained according to Keribin (2000). Similar estimates for regression parameters have been obtained using other penalized likelihood criteria, such as the AIC; the substantial difference is in the mean number of components used to estimate the unknown mixing distribution, which tend to be higher if the AIC is employed. In this context, AIC has proved to perform slightly better when the covariance matrix estimate is entailed. However, observed differences are not substantial, and thus we proceed on employing the BIC. The EM algorithm has been implemented in MatLab; just to give an idea of the computational effort involved, it required about 5 hours CPU time (on a P4 based machine) to obtain parameter estimates for 250 samples with $n = 2000$ and $k = 2, \dots, 10$.

Complete simulation results are given in the Appendix; here we focus on correlation estimates and on estimates for the selection variable effect.

We notice the reliable behavior of the proposed model even for small to moderate sample sizes, i.e. for $n = 250, 1000$; the modest additional amount of computational resources which is needed to estimate the endogenous selection model suggests that it can be used also when the association between outcomes is weak or we cast some doubts about the endogeneity of the selection variable. If the endogeneity of the selection mechanism is not accounted for, a severe bias of parameter estimates is recorded, regardless of the sample size. This bias seems to increase with absolute value of ρ (see Table 10 in the Appendix). As can be seen, the estimated effects of the selection variable are permanently biased, and the bias does not show any regular behavior for increasing n or ρ . Other model parameters are consistently estimated, as the adopted covariates are independent on the mechanism generating the selection variable; this setting is likely unrealistic, and the behavior of the exogenous model should be better investigated to understand the effect of ignoring endogeneity of the selection variable on parameter estimates when some of the adopted covariates are dependent on the selection mechanism. On the other hand, taking into account the endogeneity of the selection variable, estimates of selection variable effects, $\hat{\beta}_{j2}$, $j = 1, \dots, 4$ are nearly unbiased (see Fig. 1).

From a computational perspective, it might be interesting to show the behavior of the proposed algorithm by starting the EM algorithm from the real parameter values; this would help us to distinguish between variability due to the performance of the proposed estimator and that intrinsically due to the EM algorithm. Just to give an idea, we fitted the bivariate model for $n = 250$ by using true parameter values as starting values (results are displayed in Table 11 in the Appendix). As expected, both bias and standard errors reduce in magnitude, showing the presence of a (non-significant) bias due to computational aspects related to algorithm implementation.

Until now, we have shown the behavior of the proposed model for bivariate counts, but the model can be easily implemented for $J \geq 2$ counted responses. Tables 12–13 in the Appendix entail the case with $J = 3, 4$ counts and a binary, endogenous, selection, corresponding estimated effects are summarized in Figs. 2–3.

Computational complexity does increase linearly with the number of responses, say J ; thus, the proposed approach is not computationally intensive or time consuming *per se*. The need for a formal strategy to choose starting values may be somewhat cumbersome for $J = 3, 4$, but, still, the computational effort is linear in the number of the starting value sets, the (fixed) number of components and the sample size. As an empirical evidence, we found out that using a short-length CEM algorithm to start the *core* EM algorithm tends

Fig. 1 Simulation results for the bivariate Poisson model with endogenous selectivity—Boxplot for the selection variable effect estimates, true values in brackets

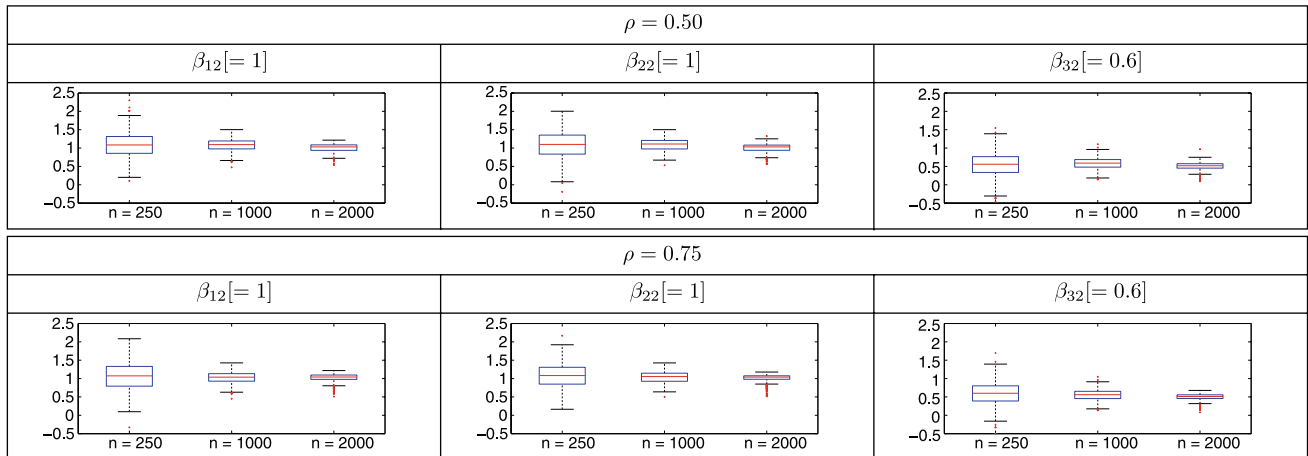
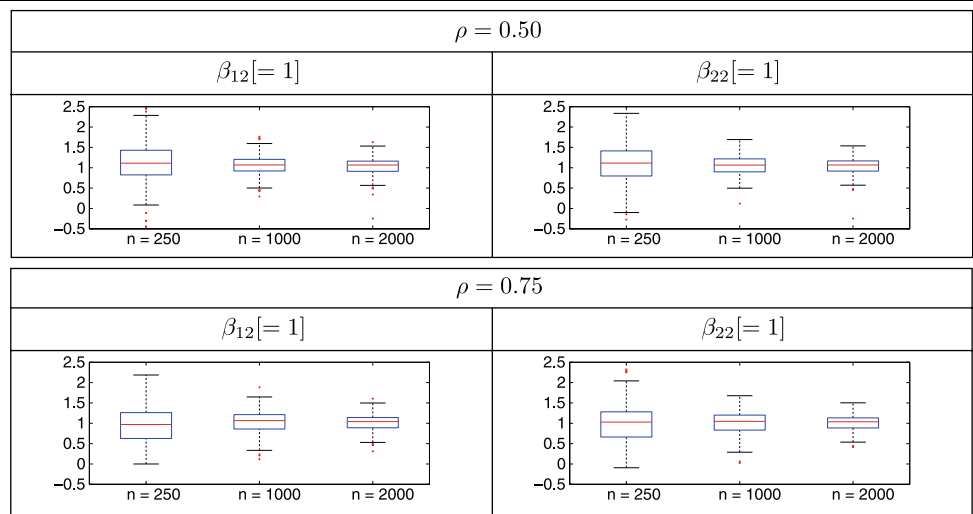


Fig. 2 Simulation results for the trivariate Poisson model with endogenous selectivity—Boxplot for the selection variable effect estimates, true values in brackets

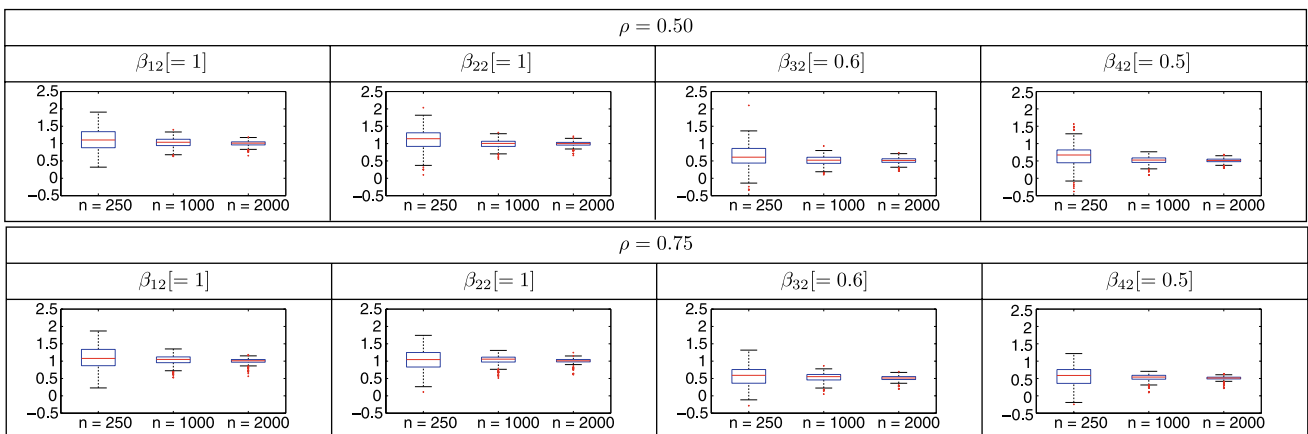


Fig. 3 Simulation results for the quadrivariate Poisson model with endogenous selectivity—Boxplot for the selection variable effect estimates, true values in brackets

Table 6 Simulation results for the bivariate Poisson model with endogeneous selectivity—Correlation coefficient estimates

	$n = 250$		$n = 1000$		$n = 2000$	
	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.5$	$\rho = 0.75$
$\hat{\rho}_{12}$	0.5482	0.7825	0.5142	0.7478	0.5085	0.7501
$\hat{\rho}_{10}$	0.5908	0.8520	0.5473	0.8098	0.5336	0.8063
$\hat{\rho}_{20}$	0.5848	0.8490	0.5423	0.8093	0.5297	0.8082

Table 7 Simulation results for the trivariate Poisson model with endogeneous selectivity—Correlation coefficient estimates

	$n = 250$		$n = 1000$		$n = 2000$	
	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.5$	$\rho = 0.75$
$\hat{\rho}_{12}$	0.5466	0.7971	0.5107	0.7629	0.5083	0.7536
$\hat{\rho}_{13}$	0.5439	0.7961	0.5281	0.7607	0.5067	0.7582
$\hat{\rho}_{23}$	0.5433	0.8035	0.5194	0.7676	0.5049	0.7488
$\hat{\rho}_{10}$	0.5651	0.8574	0.5541	0.8182	0.5475	0.8107
$\hat{\rho}_{20}$	0.5655	0.8503	0.5583	0.8152	0.5521	0.8115
$\hat{\rho}_{30}$	0.5634	0.8475	0.5612	0.8254	0.5514	0.8215

Table 8 Simulation results for the quadrivariate Poisson model with endogeneous selectivity—Correlation coefficient estimates

	$n = 250$		$n = 1000$		$n = 2000$	
	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.5$	$\rho = 0.75$
$\hat{\rho}_{12}$	0.5658	0.7896	0.5158	0.7436	0.5050	0.7509
$\hat{\rho}_{13}$	0.5542	0.7912	0.5122	0.7609	0.4988	0.7522
$\hat{\rho}_{14}$	0.5564	0.7868	0.5214	0.7676	0.5099	0.7516
$\hat{\rho}_{23}$	0.5683	0.7805	0.5172	0.7413	0.5019	0.7489
$\hat{\rho}_{24}$	0.5661	0.7764	0.5156	0.7614	0.5083	0.7491
$\hat{\rho}_{34}$	0.5611	0.7825	0.5093	0.7644	0.5080	0.7508
$\hat{\rho}_{10}$	0.5667	0.8485	0.5525	0.8201	0.5511	0.8174
$\hat{\rho}_{20}$	0.5772	0.8548	0.5518	0.8171	0.5516	0.8096
$\hat{\rho}_{30}$	0.5607	0.8494	0.5565	0.8325	0.5520	0.8274
$\hat{\rho}_{40}$	0.5684	0.8574	0.5546	0.8369	0.5538	0.8270

to substantially reduce the need to employ several different starting values.

When we look at parameter estimates, we observe a behavior which is close to the one observed for $J = 2$ counts; increasing the sample size leads to nearly unbiased parameter estimates, for fixed value of ρ (see Tables 12–13 in the Appendix). For sake of brevity, we do not report here parameter estimates for the model postulating exogeneity of the selection variable, since the behavior of the parameter estimates is qualitatively similar to the one observed before. The variance of the parameter estimates reduces with increasing sample size, but the bias does not, at least not substantially; in particular, the bias is increasing with ρ , regardless of the corresponding sample size.

While the proposed model is relatively easy to implement, there are a number of problems that may limit its usefulness in practice; a major problem is related to the estimation of covariance between the random effects in the primary and selection equations. While identifiability of the global model is ensured by conditions given above, and the covariance matrix is straightforwardly calculated, the corresponding estimates are not completely reliable. In fact, as the value of ρ increases, we registered a clear tendency to provide less stable estimates of the covariance/correlation coefficient between the random effects in the counted response and in the selection equations (see Tables 6–8). Estimates of $\hat{\rho}$ are shown in Table 6 for the bivariate case; a clear and consistent path can be noticed as the sample dimension,

n , increases, regardless of the values of the correlation coefficient (see Table 9 in the Appendix for estimates of $\hat{\beta}$). The only differences which can be observed when the correlation among the latent effects varies are in the accuracy of parameter estimates, since the results seem more stable for higher (absolute) values of ρ . This can be explained by the smaller number of components which is necessary to estimate $G(\cdot)$ when $\rho \simeq 1$; this produces a smaller variability in the resulting estimates due to reduced uncertainty with respect to the number of mixture components. The vice-versa is true when ρ decreases in absolute value towards zero. In the following tables terms $\rho_{jj'}$, $j \neq j' = 1, \dots, 4$ refer to the estimated correlation coefficients between the latent effects in the two outcome equations while values ρ_{j0} , $j = 1, \dots, 4$ refer to the correlation between the random effects in the j -th outcome equation and those in the selection equation.

It may be noticed that the correlation coefficient is not a formal parameter of the proposed model, but it is rather a by-product of the adopted estimation approach. Our proposal is roughly based on the use of a finite mixture to approximate the distribution of the primary responses and the selection variable; the mixing distribution estimates is, in this context, a discrete distribution on K support points and, unless K is fixed great enough, the resulting estimate is only a rough approximation of the true $\mathcal{G}(\cdot)$. Reasonably, this can be also due to the variance estimates of the random effect in the (binary) selection process which is usually higher than those in the counted responses equations; the estimated locations for the random effect in the binary regression are usually greater than those in the primary response equations, and may force the ρ estimates towards the corresponding bounds.

6 Concluding remarks

In this paper we discuss a mixed effect model for multivariate counts, when a binary, potentially endogenous, selection source is included in the linear predictor for the primary responses. The proposed model is a semiparametric ML alternative to the models discussed by Munkin (2003) and Munkin and Trivedi (2003). We propose to use a finite mixture representation for the regression model, estimating the unknown (multivariate) random effect distribution through a discrete (multivariate) distribution. Finite mixture models are particularly suited for health data modeling since they lead to identify potential clusters characterized by a similar propensity to the event(s) of interest. Obviously, mixture components do not necessarily correspond to clusters: the component specific densities could simply represent the departure from a misspecified homogeneous model (in this case the Poisson model).

Results obtained in a simulation study seem to suggest a reliable finite sample behavior of the proposed model in a wide variety of empirical applications; in particular, even for moderate sample sizes model parameter estimates seem to be nearly unbiased. Correlation estimates suffer, however, of some overestimation when the true ρ is near to the bounds; this can be due to the discrete nature of the estimated mixing distribution which is able to capture dependence between random effects and/or extra-Poisson variation, but is not able to provide accurate estimates for the correlation coefficients.

Appendix: Simulation results

Table 9 Simulation results for the bivariate Poisson model with endogenous selectivity

	$n = 250$		$n = 1000$		$n = 2000$	
	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.5$	$\rho = 0.75$
$\hat{\beta}_{10}$	0.4828	0.4723	0.4981	0.4903	0.5021	0.5047
(s.d.)	(0.4774)	(0.4726)	(0.2991)	(0.3035)	(0.1864)	(0.1833)
$\hat{\beta}_{11}$	0.7262	0.7276	0.7462	0.7483	0.7469	0.7502
(s.d.)	(0.2250)	(0.2223)	(0.1057)	(0.1089)	(0.0862)	(0.0738)
$\hat{\beta}_{12}$	1.0920	1.0576	1.0669	1.0299	1.0334	1.0138
(s.d.)	(0.4949)	(0.4302)	(0.2420)	(0.2731)	(0.2108)	(0.1948)
$\hat{\beta}_{20}$	0.7213	0.7258	0.7399	0.7445	0.7582	0.7532
(s.d.)	(0.4510)	(0.4595)	(0.2990)	(0.2970)	(0.1831)	(0.1769)
$\hat{\beta}_{21}$	0.4503	0.4557	0.4892	0.5017	0.4952	0.5006
(s.d.)	(0.2176)	(0.2244)	(0.1003)	(0.1042)	(0.0792)	(0.0735)
$\hat{\beta}_{22}$	1.0942	1.0751	1.0651	1.0204	1.0301	1.0002
(s.d.)	(0.4499)	(0.4221)	(0.2807)	(0.2649)	(0.2110)	(0.1948)
$\hat{\beta}_{00}$	0.5427	0.5449	0.4923	0.5031	0.4943	0.5003
(s.d.)	(0.5023)	(0.4866)	(0.3012)	(0.3192)	(0.1579)	(0.1509)
$\hat{\beta}_{01}$	1.0921	1.0794	0.9713	0.9978	0.9936	1.0012
(s.d.)	(0.2292)	(0.2364)	(0.1353)	(0.1234)	(0.0806)	(0.0959)

Table 10 Simulation results for the bivariate Poisson model with exogenous selectivity

	$n = 250$		$n = 1000$		$n = 2000$	
	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.5$	$\rho = 0.75$
$\hat{\beta}_{10}$	0.5253	0.4458	0.4757	0.6611	0.5167	0.5193
(s.d.)	(0.3325)	(0.3438)	(0.1742)	(0.1717)	(0.0961)	(0.0832)
$\hat{\beta}_{11}$	0.6253	0.6177	0.7005	0.6997	0.7409	0.7441
(s.d.)	(0.1957)	(0.2018)	(0.1026)	(0.1075)	(0.0860)	(0.0738)
$\hat{\beta}_{12}$	1.3604	1.4304	1.3271	1.5102	1.3282	1.3342
(s.d.)	(0.2526)	(0.2522)	(0.1707)	(0.1646)	(0.1026)	(0.0954)
$\hat{\beta}_{20}$	0.6486	0.6525	0.7101	0.6925	0.7502	0.7343
(s.d.)	(0.3340)	(0.3480)	(0.1737)	(0.1689)	(0.1052)	(0.0995)
$\hat{\beta}_{21}$	0.3863	0.3875	0.4637	0.4531	0.4932	0.4951
(s.d.)	(0.1912)	(0.1986)	(0.1010)	(0.1023)	(0.0759)	(0.0700)
$\hat{\beta}_{22}$	1.3873	1.4242	1.3322	1.4171	1.3308	1.3961
(s.d.)	(0.2294)	(0.2470)	(0.1494)	(0.1515)	(0.1088)	(0.1001)
$\hat{\beta}_{00}$	0.6560	0.6407	0.5431	0.5429	0.4899	0.4914
(s.d.)	(0.3217)	(0.3432)	(0.1695)	(0.1741)	(0.1034)	(0.0964)
$\hat{\beta}_{01}$	0.8769	0.8691	0.9430	0.9608	0.9806	0.9846
(s.d.)	(0.1908)	(0.2002)	(0.1766)	(0.1779)	(0.0867)	(0.0859)
$\hat{\rho}_{12}$	0.5925	0.8024	0.5842	0.7478	0.5085	0.7501

Table 11 Simulation results for the bivariate Poisson model with endogenous selectivity. Starting values set to real parameter values

	$n = 250$	
	$\rho = 0.5$	$\rho = 0.75$
$\hat{\beta}_{10}$	0.4890	0.4970
(s.d.)	(0.1840)	(0.1876)
$\hat{\beta}_{11}$	0.7404	0.7422
(s.d.)	(0.0387)	(0.0355)
$\hat{\beta}_{12}$	1.0242	0.9922
(s.d.)	(0.1595)	(0.1787)
$\hat{\beta}_{20}$	0.7346	0.7467
(s.d.)	(0.1577)	(0.1051)
$\hat{\beta}_{21}$	0.5091	0.5019
(s.d.)	(0.0343)	(0.0355)
$\hat{\beta}_{22}$	1.0406	1.0186
(s.d.)	(0.1548)	(0.1736)
$\hat{\beta}_{00}$	0.4877	0.4912
(s.d.)	(0.3484)	(0.3311)
$\hat{\beta}_{01}$	0.9544	0.9873
(s.d.)	(0.2278)	(0.2304)

Table 12 Simulation results for the trivariate Poisson model with endogenous selectivity

	$n = 250$		$n = 1000$		$n = 2000$	
	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.5$	$\rho = 0.75$
$\hat{\beta}_{10}$	0.5417	0.4823	0.4995	0.4994	0.5024	0.4974
(s.d.)	(0.3867)	(0.3565)	(0.1585)	(0.1478)	(0.0924)	(0.0802)
$\hat{\beta}_{11}$	0.7002	0.6976	0.7395	0.7345	0.7474	0.7469
(s.d.)	(0.2237)	(0.2222)	(0.1690)	(0.1540)	(0.0945)	(0.0957)
$\hat{\beta}_{12}$	1.0871	1.0663	1.0508	1.0267	1.0036	1.0064
(s.d.)	(0.3924)	(0.3782)	(0.1700)	(0.1615)	(0.1171)	(0.1244)
$\hat{\beta}_{20}$	0.6828	0.7326	0.7389	0.7346	0.7503	0.7441
(s.d.)	(0.3312)	(0.3118)	(0.1518)	(0.1679)	(0.1011)	(0.0904)
$\hat{\beta}_{21}$	0.4408	0.4516	0.4586	0.4678	0.5016	0.5027
(s.d.)	(0.2247)	(0.2362)	(0.1414)	(0.1332)	(0.0810)	(0.0921)
$\hat{\beta}_{22}$	1.0924	1.0785	1.0471	1.0328	1.0018	0.9993
(s.d.)	(0.3444)	(0.3344)	(0.1642)	(0.1760)	(0.1073)	(0.1086)
$\hat{\beta}_{30}$	1.0652	0.9541	1.0466	1.0453	1.0042	0.9920
(s.d.)	(0.3378)	(0.3285)	(0.1499)	(0.1469)	(0.0940)	(0.0925)
$\hat{\beta}_{31}$	0.9553	0.9807	0.9842	0.9931	1.0033	0.9980
(s.d.)	(0.2241)	(0.2212)	(0.1081)	(0.1103)	(0.1196)	(0.1051)
$\hat{\beta}_{32}$	0.5480	0.5337	0.5742	0.5675	0.5982	0.5898
(s.d.)	(0.2613)	(0.2355)	(0.1709)	(0.1623)	(0.1186)	(0.1090)
$\hat{\beta}_{00}$	0.5539	0.5810	0.4867	0.4819	0.5044	0.5029
(s.d.)	(0.3538)	(0.3346)	(0.1923)	(0.1940)	(0.1134)	(0.1002)
$\hat{\beta}_{01}$	1.0532	1.0700	0.9562	0.9753	1.0093	1.0175
(s.d.)	(0.2222)	(0.2409)	(0.1692)	(0.1706)	(0.1039)	(0.1172)

Table 13 Simulation results for the quadrivariate Poisson model with endogenous selectivity

	$n = 250$		$n = 1000$		$n = 2000$	
	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.5$	$\rho = 0.75$
$\hat{\beta}_{10}$	0.5745	0.4811	0.5456	0.5076	0.5128	0.4975
(s.d.)	(0.2500)	(0.2774)	(0.1294)	(0.1141)	(0.1023)	(0.0980)
$\hat{\beta}_{11}$	0.7032	0.7254	0.7273	0.7318	0.7592	0.7529
(s.d.)	(0.1843)	(0.2113)	(0.1171)	(0.1230)	(0.0752)	(0.0861)
$\hat{\beta}_{12}$	1.1137	1.0971	1.0271	1.0278	1.0040	0.9994
(s.d.)	(0.3054)	(0.3172)	(0.1865)	(0.1967)	(0.1105)	(0.1145)
$\hat{\beta}_{20}$	0.7875	0.7284	0.7466	0.7335	0.7482	0.7508
(s.d.)	(0.2508)	(0.2771)	(0.1438)	(0.1424)	(0.0960)	(0.1071)
$\hat{\beta}_{21}$	0.4638	0.4973	0.4840	0.4907	0.4980	0.5004
(s.d.)	(0.2076)	(0.2336)	(0.1322)	(0.1293)	(0.0874)	(0.0914)
$\hat{\beta}_{22}$	1.1036	1.0362	0.9832	1.0248	0.9904	1.0035
(s.d.)	(0.3055)	(0.2949)	(0.1947)	(0.2096)	(0.1052)	(0.1120)
$\hat{\beta}_{30}$	1.0367	0.9689	0.9837	0.9915	1.0011	1.0037
(s.d.)	(0.2721)	(0.2806)	(0.1599)	(0.1637)	(0.0930)	(0.0823)
$\hat{\beta}_{31}$	1.0170	0.9808	0.9785	0.9809	1.0024	0.9993
(s.d.)	(0.2178)	(0.2226)	(0.1215)	(0.1190)	(0.0855)	(0.0820)
$\hat{\beta}_{32}$	0.6321	0.5659	0.6137	0.6266	0.6071	0.6087
(s.d.)	(0.3287)	(0.2730)	(0.1329)	(0.1263)	(0.0798)	(0.0724)

Table 13 (Continued)

	$n = 250$		$n = 1000$		$n = 2000$	
	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.5$	$\rho = 0.75$
$\hat{\beta}_{40}$	1.5541	1.4313	1.4810	1.5239	1.5084	1.5063
(s.d.)	(0.2436)	(0.2683)	(0.1350)	(0.1473)	(0.1021)	(0.1031)
$\hat{\beta}_{41}$	0.5359	0.5557	0.5770	0.5901	0.5938	0.6011
(s.d.)	(0.1843)	(0.2004)	(0.1307)	(0.1368)	(0.0972)	(0.0901)
$\hat{\beta}_{42}$	0.6398	0.5590	0.5111	0.5167	0.5076	0.5061
(s.d.)	(0.3133)	(0.2856)	(0.1861)	(0.2007)	(0.1159)	(0.1213)
$\hat{\beta}_{00}$	0.5995	0.6073	0.4983	0.4903	0.4994	0.5016
(s.d.)	(0.3560)	(0.3644)	(0.1918)	(0.2066)	(0.1190)	(0.1236)
$\hat{\beta}_{01}$	0.9988	1.0234	0.9993	1.0213	0.9975	1.0111
(s.d.)	(0.2181)	(0.2113)	(0.1458)	(0.1528)	(0.0807)	(0.0825)

References

- Alfö, M., Aitkin, M.: Variance component models for longitudinal count data with baseline information: epilepsy data revisited. *Stat. Comput.* **16**, 231–238 (2006)
- Alfö, M., Trovato, G.: Semiparametric mixture models for multivariate count data, with application. *Econom. J.* **7**, 1–29 (2004)
- Biernacki, C., Celeux, G., Govaert, G.: Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Comput. Stat. Data Anal.* **41**, 561–575 (2003)
- Böhning, D.: *Computer-assisted Analysis of Mixtures and Applications: Meta-analysis, Disease Mapping and Others*. Chapman & Hall/CRC, London (2000)
- Böhning, D.: The EM algorithm with gradient function update for discrete mixtures with known (fixed) number of components. *Stat. Comput.* **13**, 257–265 (2003)
- Cameron, A., Trivedi, P.: *Microeconometrics: Methods and Applications*. Cambridge University Press, Cambridge (2005)
- Chib, S., Winkelmann, R.: Markov chain Monte Carlo analysis of correlated count data. *J. Bus. Econ. Stat.* **19**, 428–435 (2001)
- Crouch, E., Spiegelman, D.: The evaluation of integrals of the form $\int_{-\infty}^{\infty} f(t) \exp(-t^2) dt$: application to logistic-normal models. *J. Am. Stat. Assoc.* **85**, 464–469 (1990)
- Crouchley, R., Davies, R.: A comparison of population average and random effect models for the analysis of longitudinal count data with baseline information. *J. R. Stat. Soc. A* **162**, 331–347 (1999)
- Davidson, R., Mackinnon, J.: *Estimation and Inference in Econometrics*. Oxford University Press, Oxford (1993)
- Deb, P., Holmes, A.: Estimates of use and costs of behavioural health care: a comparison of standard and finite mixture models. *Health Econ.* **9**, 475–489 (2000)
- Deb, P., Trivedi, P.: Demand for medical care by the elderly: a finite mixture approach. *J. Appl. Econom.* **12**, 313–336 (1997)
- Deb, P., Munkin, M., Trivedi, P.: Bayesian analysis of the two-part model with endogeneity: application to health care expenditure. *J. Appl. Econom.* **21**, 1081–1099 (2006)
- Feng, Z., McCulloch, C.: Using bootstrap likelihood ratios in finite mixture models. *J. R. Stat. Soc. B* **58**, 609–617 (1996)
- Follmann, D., Lambert, D.: Identifiability of finite mixtures of logistic regression models. *J. Stat. Plan. Inference* **27**, 375–381 (1991)
- Fotouhi, A.: The initial conditions problem in longitudinal binary process: a simulation study. *Simul. Model. Pract. Theory* **13**, 566–583 (2005)
- Geyer, C., Thompson, E.: Constrained Monte Carlo maximum likelihood for dependent data. *J. R. Stat. Soc. B* **54**, 657–699 (1992)
- Gueorguieva, R.: A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Stat. Model.* **1**, 177–193 (2001)
- Harry, J.: *Multivariate Models and Multivariate Dependence Concepts*. Chapman & Hall/CRC, London (1997)
- Heckman, J., Singer, B.: A method for minimizing the impact of distributional assumptions in econometric models of duration. *Econometrica* **52**, 271–320 (1984)
- Jank, W., Booth, J.: Efficiency of Monte Carlo EM and simulated maximum likelihood in two-stage hierarchical models. *J. Comput. Graph. Stat.* **12**, 214–229 (2003)
- Karlis, D., Meligkotsidou, L.: Multivariate Poisson regression with covariance structure. *Stat. Comput.* **15**, 255–265 (2006)
- Karlis, D., Meligkotsidou, L.: Finite mixtures of multivariate Poisson distributions with application. *J. Stat. Plan. Inference* **137**, 1942–1960 (2007)
- Keribin, C.: Consistent estimation of the order of mixture models. *Sankhyā: Indian J. Stat.* **62**, 49–66 (2000)
- Kiefer, J., Wolfowitz, J.: Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Stat.* **27**, 887–906 (1956)
- Laird, N.: Nonparametric maximum likelihood estimation of a mixing distribution. *J. Am. Stat. Assoc.* **73**, 805–811 (1978)
- Lindsay, B.: The geometry of mixture likelihoods: a general theory. *Ann. Stat.* **11**, 86–94 (1983a)
- Lindsay, B.: The geometry of mixture likelihoods, part II: the exponential family. *Ann. Stat.* **11**, 783–792 (1983b)
- McCulloch, C.: Maximum likelihood estimation of variance components for binary data. *J. Am. Stat. Assoc.* **89**, 330–335 (1994)
- McCulloch, C.: Maximum likelihood algorithms for generalized linear mixed models. *J. Am. Stat. Assoc.* **92**, 162–170 (1997)
- Mello, M., Stearns, S., Norton, E.: Do medicare hmos still reduce health services use after controlling for selection bias? *Health Econ.* **11**, 323–340 (2002)
- Mroz, T.: Discrete factor approximations in simultaneous equation models: Estimating the impact of a dummy endogenous variable on a continuous outcome. *J. Econom.* **92**, 233–274 (1999)
- Munkin, M.: The MCMC and SML estimation of a self-selection model with two outcomes. *Comput. Stat. Data Anal.* **42**, 403–424 (2003)
- Munkin, M., Trivedi, P.: Simulated maximum likelihood estimation of multivariate mixed-Poisson regression models, with application. *Econom. J.* **2**, 29–48 (1999)
- Munkin, M., Trivedi, P.: Bayesian analysis of a self-selection model with multiple outcomes using simulation-based estimation: an application to the demand for healthcare. *J. Econom.* **114**, 197–220 (2003)

- Munkin, M., Trivedi, P.: Bayesian analysis of the ordered probit model with endogenous selection. *J. Econom.* **143**, 334–348 (2008)
- Nikoloulopoulos, A., Karlis, D.: Modelling multivariate count data using copula functions. In: Rizzi, A., Vichi, M. (eds.) *COMPSTAT 2006—Electronic Proceedings in Computational Statistics*. Physica-Verlag, Heidelberg (2006)
- Rabe-Heskett, S., Skrondal, A.: Reliable estimation of generalized linear mixed models using adaptive quadrature. *Stata J.* **2**, 1–21 (2002)
- Smith, M., Moffatt, P.: Fisher's information on the correlation coefficient in bivariate logistic models. *Aust. N. Z. J. Stat.* **41**, 315–330 (1999)
- Teicher, H.: Identifiability of finite mixtures. *Ann. Math. Stat.* **34**, 244–248 (1963)
- Terza, J.: Estimating count data models with endogenous switching: sample selection and endogenous treatment effects. *J. Econom.* **84**, 129–154 (1998)
- van Ophem, H.: Modeling selectivity in count data models. *J. Bus. Econ. Stat.* **18**, 503–510 (2000)
- Wang, P., Puterman, M., Cockburn, I., Le, N.: Mixed Poisson regression models with covariate dependent rates. *Biometrics* **52**, 381–400 (1996)
- Winkelmann, R.: Count data models with selectivity. *Econom. Rev.* **17**, 339–359 (2000a)
- Winkelmann, R.: *Econometric Analysis of Count Data*. Springer, Berlin (2000b)
- Winkelmann, R.: Seemingly unrelated negative binomial regression. *Oxf. Bull. Econ. Stat.* **62**, 553–560 (2000c)
- Zimmer, D., Trivedi, P.: Using trivariate copulas to model sample selection and treatment effects: Application to family health care demand. *J. Bus. Econ. Stat.* **24**, 63–76 (2006)