

# On the estimation of normal copula discrete regression models using the continuous extension and simulated likelihood

Aristidis K. Nikoloulopoulos\*

## Abstract

The continuous extension of a discrete random variable is amongst the computational methods used for estimation of multivariate normal copula-based models with discrete margins. Its advantage is that the likelihood can be derived conveniently under the theory for copula models with continuous margins, but there has not been a clear analysis of the adequacy of this method. We investigate the asymptotic and small-sample efficiency of two variants of the method for estimating the multivariate normal copula with univariate binary, Poisson, and negative binomial regressions, and show that they lead to biased estimates for the latent correlations, and the univariate marginal parameters that are not regression coefficients. We implement a maximum simulated likelihood method, which is based on evaluating the multidimensional integrals of the likelihood with randomized quasi Monte Carlo methods. Asymptotic and small-sample efficiency calculations show that our method is nearly as efficient as maximum likelihood for fully specified multivariate normal copula-based models. An illustrative example is given to show the use of our simulated likelihood method.

*Keywords:* Continuous extension; Jitters; Multivariate normal copula; Rectangle probabilities; Simulated likelihood.

## 1 Introduction

For multivariate discrete data  $\mathbf{Y} = (Y_1, \dots, Y_d)$  given a vector of (continuous or discrete) covariates  $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_d)$  with  $\mathbf{x}_j \in \mathbb{R}^p$ ,  $j = 1, \dots, d$ , the discretized multivariate normal (MVN) distribution, or the MVN copula with discrete margins, has been in use for a considerable length of time, e.g. Joe (1997), and much earlier in the biostatistics (Ashford and Sowden, 1970), psychometrics (Muthén, 1978), and econometrics (Hausman and Wise, 1978) literature. It is usually known as a multivariate, or multinomial, probit model. The multivariate probit model is a simple example of the MVN copula with univariate probit regressions as the marginals. In the general case, the discretized MVN model has the following cumulative distribution function (cdf):

$$\Pr(Y_1 \leq y_1, \dots, Y_d \leq y_d; \mathbf{x}) = \Phi_d(\Phi^{-1}[F_{Y_1}(y_1; \mathbf{x}_1)], \dots, \Phi^{-1}[F_{Y_d}(y_d; \mathbf{x}_d)]; \mathbf{R}), \quad (1)$$

where  $\Phi_d$  denotes the standard MVN distribution function with correlation matrix  $\mathbf{R} = (\rho_{jk} : 1 \leq j < k \leq d)$ ,  $\Phi$  is the cdf of the univariate standard normal, and  $F_{Y_1}(y_1; \mathbf{x}_1), \dots, F_{Y_d}(y_d; \mathbf{x}_d)$  are the univariate discrete cdfs.

Other multivariate copulas for discrete response data have been around a long time, e.g. in Joe (1997), and earlier for some simple copula models. Simple parametric families of

---

\*A.Nikoloulopoulos@uea.ac.uk, School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK

copulas have a closed form cdf; hence the joint likelihood is straightforward to derive from the probability mass function (pmf) as a finite difference of the cdf, but they provide limited dependence. For example, Meester and MacKay (1994) used a Frank copula (Frank, 1979) with a closed form cdf to model multivariate binary data. The Frank copula is a member of the Archimedean class of copulas, which are limited to exchangeable structures.

The MVN copula generated by the MVN distribution inherits the useful properties of the latter, thus allowing a wide range for dependence, and overcomes the drawback of limited dependence inherent in simple parametric families of copulas (Nikoloulopoulos et al., 2011). The use of the MVN with logistic regression (or Poisson or negative binomial regression) is just a special case of the general theory of dependence modelling with copulas. Implementation of the MVN copula for discrete data (discretized MVN) is possible, but not easy, because the MVN distribution as a latent model for discrete response requires rectangle probabilities based on high-dimensional integrations or their approximations (Nikoloulopoulos and Karlis, 2009). Many approaches have been considered for computing high-dimensional normal probabilities, see e.g. Schervish (1984), Genz (1992), and Joe (1995). These could be used to evaluate the normal copula-based likelihood for discrete data with general dependence.

Four recent papers (Heinen and Rengifo, 2007, 2008; Madsen, 2009; Madsen and Fang, 2011) attempt to “approximate” the likelihood, by using the continuous extension (CE) of a discrete random variable developed in Denuit and Lambert (2005). Heinen and Rengifo (2007, 2008) proposed a surrogate likelihood, assuming the latent uniform variables in the CE of a discrete random variable are observed, while Madsen (2009) and Madsen and Fang (2011) used also the CE but their method is actually an example of a simulated likelihood to compute the MVN rectangle probabilities. The methods based on the CE cannot be recommended until its properties have been studied and compared to existing methods. The CE method in Denuit and Lambert (2005) has been used to prove theoretical results for copula-based concordance measures for discrete data. Although its application to copula dependence modeling for discrete data is novel, its theoretical and small-sample efficiency has yet to be established in that context. The contribution in this paper is (a) to examine thoroughly the accuracy and the adequacy of the surrogate and simulated likelihood method based on the CE using asymptotics and simulations; (b) to improve the efficiency of simulated likelihood by transforming the rectangle integrals as in Genz (1992); and (c) to give precise guidelines for handling the MVN copula-based likelihood for regression models with dependent discrete response data.

For ease of exposition, we consider the case that the univariate marginal parameters are common to different univariate margins;  $\gamma$  denotes the  $r$ -dimensional vector of univariate marginal parameters that are not regression coefficients, and  $\beta$  the  $p$ -dimensional vector of the regression parameters. The marginal means,  $E(Y_j) = \mu_j$ ,  $j = 1, \dots, d$ , depend on the vector of the covariates  $\mathbf{x}_j$ , via the vector  $\beta$  and a link function,  $\eta(\mu_j) = \mathbf{x}_j^\top \beta$ ,  $j = 1, \dots, d$ . The remainder of the paper proceeds as follows. Section 2 provides an overview of the surrogate and simulated likelihood method using the CE. In Section 3 we describe appropriate methods for handling the MVN copula-based likelihood for regression models with dependent discrete response data, and propose a maximum simulated likelihood method based on evaluating the multidimensional integrals of the likelihood with randomized quasi Monte Carlo methods. Section 4 and Section 5 contain theoretical (e.g. asymptotic properties of the estimators) and small sample efficiency calculations, respectively, to assess the accuracy of the methods. Section 6 presents an application of the proposed simulated likelihood to the toenail infection data in Madsen and Fang (2011). We conclude this article with some discussion, followed by a technical Appendix.

## 2 Surrogate and simulated likelihood using the CE

Denuit and Lambert (2005) proposed a CE of integer-valued random variables to study concordance measures for dependent discrete data. They associate an integer-valued random variable  $Y$  with a jittered random variable  $Y^*$ , such that,

$$Y^* = Y + (V - 1),$$

where  $V$  is a uniform random variable in the unit interval and independent of  $Y$ .  $Y^*$  is a CE of  $Y$ , and has support on the reals. Given a particular realization of the jittering, the (conditional) density of  $Y^*$  is  $f_{Y^*|V}(y^*|v) = f_Y(\lceil y^* \rceil)$ , where  $\lceil y^* \rceil$  is the ceiling of  $y^*$ , or the smallest integer greater than or equal to  $y^*$ . The (conditional) cdf of  $Y^*$  is  $F_{Y^*|V}(y^*|v) = F_Y(\lceil y^* \rceil - 1) + v f_Y(\lceil y^* \rceil)$ . It is easy to see that the simplified cdf and density of  $Y^*$  take the form  $F_{Y^*|V}(y^*|v) = F_Y(y - 1) + v f_Y(y)$  and  $f_{Y^*|V}(y^*|v) = f_Y(y)$ , respectively.

### 2.1 Surrogate likelihood based on the CE of a discrete random variable

Heinen and Rengifo (2007, 2008) were the first that adopt this approach to form a surrogate likelihood for MVN copula-based models with discrete margins (hereafter HR method). Copula-based models were originally developed for continuous responses where the density is obtained using partial derivatives of the multivariate cdf (see e.g. Nikoloulopoulos et al. (2012)), and hence the numerical calculations are much simpler. The corresponding density for the jittered continuous vector  $\mathbf{Y}^* = (Y_1^*, \dots, Y_d^*)$  given  $n$  independent standard uniforms  $\mathbf{V} = (V_1, \dots, V_d)$  is,

$$h_{\mathbf{Y}^*|\mathbf{V}}(\mathbf{y}^*|\mathbf{v}; \mathbf{x}) = c\left(F_{Y_1^*|V_1}(y_1^*|v_1; \mathbf{x}_1), \dots, F_{Y_d^*|V_d}(y_d^*|v_d; \mathbf{x}_d); \mathbf{R}\right) \prod_{j=1}^d f_{Y_j^*|V_j}(y_j^*|v_j; \mathbf{x}_j),$$

where  $\mathbf{v} = (v_1, \dots, v_d)$  are realizations of the jitters  $\mathbf{V}$  and  $c(u_1, \dots, u_d; \mathbf{R})$  is the  $d$ -variate normal copula density. Since the MVN copula has a closed form density,

$$c(u_1, \dots, u_d; \mathbf{R}) = |\mathbf{R}|^{-1/2} \exp\left[\frac{1}{2} \{\mathbf{q}^\top (\mathbf{I}_d - \mathbf{R}^{-1}) \mathbf{q}\}\right],$$

where  $\mathbf{q} = (q_1, \dots, q_d)$  with  $q_j = \Phi^{-1}(u_j)$ ,  $j = 1, \dots, d$  and  $\mathbf{I}_d$  is the  $d$ -dimensional identity matrix, the authors avoid the multidimensional integration by using the CE of the discrete random variables. The estimated parameters can be obtained by maximizing the surrogate log-likelihood,

$$\ell_{HR}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{R}) = \sum_{i=1}^n \log h_{\mathbf{Y}^*|\mathbf{V}}(y_{i1}^*, \dots, y_{id}^* | v_{i1}, \dots, v_{id}; \mathbf{x}_{i1}, \dots, \mathbf{x}_{id}),$$

over the univariate and copula parameters  $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{R})$ . In order to avoid the noise introduced by the jitters  $\mathbf{V}$ , they use  $m$  jitters. That is, they simulate a vector of independent standard uniforms  $\mathbf{V}_k = (V_{1,k}, \dots, V_{d,k})$  and maximize,

$$\ell_{HR}^{(k)}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{R}) = \sum_{i=1}^n \log h_{\mathbf{Y}^*|\mathbf{V}_k}(y_{i1}^*, \dots, y_{id}^* | v_{i1,k}, \dots, v_{id,k}; \mathbf{x}_{i1}, \dots, \mathbf{x}_{id}),$$

for  $k = 1, \dots, m$ . The HR estimates are the average of the estimates over the  $m$  runs.

## 2.2 Simulated likelihood based on the CE of a discrete random variable

Madsen (2009) and Madsen and Fang (2011) also adopt the CE of  $Y_j$  proposed by Denuit and Lambert (2005), but they make use of importance functions in combination with simulated likelihood. Simulated likelihood for multivariate probit models has been used in econometrics since the early 1990s, and it is one of the recommended methods for generalized linear mixed models, see e.g. Chapter 7 in Demidenko (2004). The simulated likelihood method, applies importance sampling to simulate the likelihood function, but this must be done in a way that the simulated likelihood changes little when the parameters of the model are perturbed a little. This can be accomplished by using the same set of random draws in an appropriate way. For importance sampling to work well, the integrand, say,

$$\int e(z) dz = \int \frac{e(z)}{g(z)} g(z) dz = E_Z \left[ \frac{e(z)}{g(z)} \right],$$

must be converted to a good form of an empirical average based on data simulated from a suitable importance sampling distribution  $g$ . Better choices of  $g$  have small variance for  $w(z) = e(z)/g(z)$ , such that the importance weight  $w$  is bounded. A bad choice can lead to an inflated variance of the integral-estimate and thus to a very poor approximation. Hence, in order to use simulated likelihood in an efficient way the choice of the the importance sampling distribution is crucial.

The simulated likelihood using the jitters (hereafter MF method) is,

$$L_{MF}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{R}) = E_{\mathbf{V}} \left[ \prod_{i=1}^n h_{\mathbf{Y}^* | \mathbf{V}}(y_{i1}^*, \dots, y_{id}^* | v_{i1}, \dots, v_{id}; \mathbf{x}_{i1}, \dots, \mathbf{x}_{id}) \right]. \quad (2)$$

Essentially, the MF importance weight is unbounded. Madsen and Fang (2011) write an exponential with quadratic form  $\frac{1}{2} \{ \mathbf{q}^\top (\mathbf{I}_d - \mathbf{R}^{-1}) \mathbf{q} \}$ , which is not positive definite, so for part of the  $\mathbf{q}$  vector space, it can be arbitrarily negative; hence their integrand can be arbitrarily large. This is a poor choice of an importance function, and it is hard to achieve good accuracy for the numerical integral.

The expected likelihood in (2) can be approximated by averaging over the jitters  $\mathbf{V}_k$ ,  $k = 1, \dots, m$ , that is,

$$\begin{aligned} L_{MF}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{R}) &= \frac{1}{m} \sum_{k=1}^m \left[ \prod_{i=1}^n h_{\mathbf{Y}^* | \mathbf{V}_k}(y_{i1}^*, \dots, y_{id}^* | v_{i1,k}, \dots, v_{id,k}; \mathbf{x}_{i1}, \dots, \mathbf{x}_{id}) \right] \\ &= \frac{1}{m} \sum_{k=1}^m \left( \prod_{i=1}^n \left\{ |\mathbf{R}|^{-1/2} \exp \left[ \frac{1}{2} \{ \mathbf{q}_{i,k}^\top (\mathbf{I}_d - \mathbf{R}^{-1}) \mathbf{q}_{i,k} \} \right] \prod_{j=1}^d f_{Y_j^* | V_j}(y_{ij}^* | v_{ij,k}; \mathbf{x}_{ij}) \right\} \right) \\ &= \frac{1}{m} \prod_{i=1}^n \prod_{j=1}^d f_{Y_j}(y_{ij}; \mathbf{x}_{ij}) \sum_{k=1}^m \left( \prod_{i=1}^n \left\{ |\mathbf{R}|^{-1/2} \exp \left[ \frac{1}{2} \{ \mathbf{q}_{i,k}^\top (\mathbf{I}_d - \mathbf{R}^{-1}) \mathbf{q}_{i,k} \} \right] \right\} \right), \quad (3) \end{aligned}$$

where  $\mathbf{q}_{i,k} = (\Phi^{-1}[F_{Y_1^* | V_{1,k}}(y_{i1}^* | v_{i1,k}; \mathbf{x}_{i1})], \dots, \Phi^{-1}[F_{Y_d^* | V_{d,k}}(y_{id}^* | v_{id,k}; \mathbf{x}_{id})])$ .

The MF estimates of  $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{R})$  are derived by maximizing the  $\ell_{MF} = \log L_{MF}$  with respect to the univariate and copula parameters. For the MF (and HR in subsection 2.1) likelihood, to work well in a numerical optimization routine, the evaluations via simulation have to be smooth (differentiable) when the parameters change by small amounts. In order to accomplish this, the same set of uniform random variables should be used no matter the parameter values in the iterative optimization; see e.g. Bhat and Sidharthan (2011).

### 3 Appropriate methods for handling the likelihood

Estimation of the model parameters  $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{R})$  can be approached by the standard maximum likelihood (ML) method, by maximizing the joint log-likelihood (Joe, 1997),

$$\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{R}) = \sum_{i=1}^n \log h_{\mathbf{Y}}(y_{i1}, \dots, y_{id}; \mathbf{x}_{i1}, \dots, \mathbf{x}_{id}), \quad (4)$$

over the univariate and copula parameters  $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{R})$ , where  $h_{\mathbf{Y}}$  is the joint pmf of the multivariate discrete response vector  $\mathbf{Y} = (Y_1, \dots, Y_d)$ . Song (2007), influencing other authors (e.g. Heinen and Rengifo (2007, 2008); Madsen (2009); Madsen and Fang (2011)), acknowledged that the pmf can be obtained as a finite difference of the cdf in (1). Generally speaking, this is an imprecise statement, since calculating the finite difference among  $2^d$  numerically computed orthant probabilities may result in negative values. The pmf can be alternatively obtained by computing the following rectangle probability,

$$\begin{aligned} h_{\mathbf{Y}}(\mathbf{y}; \mathbf{x}) &= \Pr(Y_1 = y_1, \dots, Y_d = y_d; \mathbf{x}) \\ &= \Pr(y_1 - 1 < Y_1 \leq y_1, \dots, y_d - 1 < Y_d \leq y_d; \mathbf{x}) \\ &= \int_{\Phi^{-1}[F_{Y_1}(y_1-1; \mathbf{x}_1)]}^{\Phi^{-1}[F_{Y_1}(y_1; \mathbf{x}_1)]} \dots \int_{\Phi^{-1}[F_{Y_d}(y_d-1; \mathbf{x}_d)]}^{\Phi^{-1}[F_{Y_d}(y_d; \mathbf{x}_d)]} \phi_{\mathbf{R}}(z_1, \dots, z_d) dz_1 \dots dz_d, \end{aligned} \quad (5)$$

where  $\phi_{\mathbf{R}}$  denotes the standard MVN density with latent correlation matrix  $\mathbf{R}$ .

The computation of MVN rectangle probabilities such as the one in (5) is possible, but requires multidimensional integration. However, there is a special case overlooked in the biostatistics literature (Kiefer, 1982; Madsen and Fang, 2011; Ochi and Prentice, 1984; Song et al., 2009): for positive exchangeable correlation structures, the  $d$ -dimensional integrals conveniently reduce to 1-dimensional integrals (Johnson and Kotz, 1972, p. 48). Hence, MVN rectangle probabilities can be quickly computed to a desired accuracy that is  $10^{-6}$  or less, because 1-dimensional numerical integrals are computationally easier than higher-dimensional numerical integrals. For general correlation structures, there are several papers in the literature that focus on the computation of the MVN rectangle probabilities, and, conveniently, the implementation of the proposed algorithms is available in contributed R packages<sup>1</sup>. Schervish (1984) proposed a locally adaptive numerical integration method but this method, while more accurate, is time consuming and restricted to a low dimension. Therefore, Genz (1992) proposed a Monte Carlo method and Joe (1995) proposed two approximations to multivariate normal probabilities. The first-order approximation makes use of all of the univariate and bivariate marginal probabilities, and the second-order approximation also makes use of trivariate and four-variate marginal probabilities. These advances in computation of MVN probabilities can be used to implement MVN copula models with discrete response data:

- For positive exchangeable dependence structures, if one computes the rectangle MVN probabilities in (4) with the 1-dimensional integral method in Johnson and Kotz (1972), then one is using a numerically accurate likelihood method that is valid for any dimension.
- If one computes the rectangle MVN probabilities in (4) with the methods in Joe (1995), then one is using an approximate likelihood method; see e.g. Joe (1997).

---

<sup>1</sup>Both approximations to MVN rectangle in Joe (1995), the 1-dimensional integral in the exchangeable case, and the method in Schervish (1984), can be computed with the functions *mvnapp*, *exchmvn*, and *pmnorm*, respectively, in the R package *mprobit* (Joe et al., 2011). The methods in Genz (1992) can be computed with the function *pmvnorm* in the R package *mutnorm* (Genz et al., 2012).

- If one computes the rectangle MVN probabilities in (4) via simulation based on the method in Genz (1992), then one is using the simulated likelihood method but with something much better than in Madsen (2009) and Madsen and Fang (2011).

Since both approximations in Joe (1995) are better when the correlations are smaller, we concentrate on the simulated likelihood method based on the computation of the MVN probabilities ala Genz (1992).

### 3.1 Simulated likelihood method via the optimized method of Genz and Bretz (2002)

For an integral in the context of an MVN rectangle probability,

$$P(a_j < Z_j < b_j, j = 1, \dots, d) = \int_{a_1}^{b_1} \cdots \int_{a_d}^{b_d} \phi_{\mathbf{R}}(z_1, \dots, z_d) dz_1 \dots dz_d, \quad (6)$$

for an MVN density  $\phi_{\mathbf{R}}$  with correlation matrix  $\mathbf{R}$ , Genz (1992) uses a sequence of three transformations to transform the original integral into an integral over a unit hypercube,

$$P(a_j < Z_j < b_j, j = 1, \dots, d) = \int_0^1 \cdots \int_0^1 e(v_1, \dots, v_{d-1}) dv_1 \dots dv_{d-1},$$

with,

$$\begin{aligned} e(v_1, \dots, v_{d-1}) &= e_1 e_2(v_1) e_3(v_1, v_2) \dots e_d(v_1, \dots, v_{d-1}), \\ e_j(v_1, \dots, v_{j-1}) &= \varepsilon_j(v_1, \dots, v_{j-1}) - \varepsilon_j(v_1, \dots, v_{j-1}), \\ \varepsilon_j(v_1, \dots, v_{j-1}) &= \Phi\left(\left[a_j - \sum_{k=1}^{j-1} c_{jk} \Phi^{-1}\{\varepsilon_k(v_1, \dots, v_{k-1}) + u_k e_k(v_1, \dots, v_{k-1})\}\right] / c_{jj}\right), \\ \varepsilon_j(v_1, \dots, v_{j-1}) &= \Phi\left(\left[b_j - \sum_{k=1}^{j-1} c_{jk} \Phi^{-1}\{\varepsilon_k(v_1, \dots, v_{k-1}) + u_k e_k(v_1, \dots, v_{k-1})\}\right] / c_{jj}\right); \end{aligned}$$

$\mathbf{C} = (c_{jk} : 1 \leq j < k \leq d)$  is the matrix used for the Cholesky decomposition of  $\mathbf{R}$ .

This sequence of transformations reduces the number of integration variables by one, but, more interestingly, the rectangle integral is converted to a bounded integrand, so that the rectangle probability can be successfully evaluated via importance sampling based on a  $(d-1)$ -variate standard uniform random sample  $\mathbf{V}_k$ ,

$$P(a_j < Z_j < b_j, j = 1, \dots, d) = m^{-1} \sum_{k=1}^m e(\mathbf{v}_k).$$

Genz and Bretz (2002) improve the performance of the crude Monte Carlo methods in Genz (1992) by calling a randomized quasi Monte Carlo method with the use of antithetic variates. They use approximations of the form,

$$P(a_j < Z_j < b_j, j = 1, \dots, d) = m^{-1} \sum_{k=1}^m \frac{1}{2P} \sum_{p=1}^P (e(|2[\mathbf{p}_p + \mathbf{v}_k] - \mathbf{1}|) + e(\mathbf{1} - |2[\mathbf{p}_p + \mathbf{v}_k] - \mathbf{1}|)). \quad (7)$$

In this form,  $[\mathbf{t}]$  denotes the vector obtained by taking the fractional part of each of the components of  $\mathbf{t}$ , and  $\mathbf{p}_p, p = 1, \dots, P$  is a set of quasi-random points.

To sum up, Genz and Bretz (2002) achieve error reduction of Monte Carlo methods with variance reduction methods as (a) transforming to a bounded integrand, (b) using



antithetic variates, and (c) using a randomized quasi Monte Carlo method. The test results in Genz and Bretz (2002, 2009) show that their method is very efficient, compared to other methods in the literature. Note in passing that the method in Genz and Bretz (2002) is “optimized” in the `mtvnorm` R package (Genz et al., 2012). Hence, on the calculation of the approximation in (7), one doesn’t need to worry about the selection, for example, of the number of jitters  $m$ , or the number of quasi points  $P$ .

We implement a simulated likelihood (hereafter SL), where the rectangle MVN probabilities are computed based on the method in Genz and Bretz (2002). Since the estimation of the parameters of the MVN copula-based models is obtained using a quasi-Newton routine (Nash, 1990) applied to the log-likelihood in (4), the use of randomized quasi Monte Carlo simulation to four decimal place accuracy for evaluations of integrals works poorly, because numerical derivatives of the log-likelihood with respect to the parameters are not smooth. In order to achieve smoothness, the same set of uniform random variables should be used for every rectangle probability that comes up in the optimization of the SL. Hence, our implementation allows estimation of parameters from response vectors of dimension much larger than three that used in previous theoretical studies and applications (e.g. Song (2007); Song et al. (2009)).

## 4 Theoretical efficiency

In this section, we perform several theoretical calculations, similarly to Joe (1995, 2008), to investigate the accuracy of the “approximate” likelihood methods (Heinen and Rengifo, 2007, 2008; Madsen, 2009; Madsen and Fang, 2011), and the SL method in Subsection 3.1, which is based on evaluating the multidimensional integrals at the likelihood with the method in Genz and Bretz (2002).

### 4.1 Asymptotics

In this subsection, we study the asymptotics of the HR and SL methods, and we assess the accuracy based on the limit (as the number of clusters increases to infinity) of the maximum surrogate likelihood estimate (HRMLE) and the maximum SL estimate (MSLE). By varying factors such as dimension  $d$ , regression and not regression parameters, the amount of discreteness (binary versus count response), and latent correlation for exchangeable structures, we demonstrate patterns in the asymptotic bias of the HRMLE and MSLE, and assess the performance of HR and SL. For the cases where we compute the probability limit, we will take a constant dimension  $d$  that increases. We will also conveniently use discrete covariates so that we can assume that there are a finite number of distinct values. The idea of the continuous extension is to replace a numerically more difficult MVN rectangle probability calculation with a simpler MVN density value, and hence it is discrete responses that matter and not the type of covariates. The pattern should be similar with continuous covariates but the bias cannot be determined as easily.

Let the  $T$  distinct cases for the discrete response and the covariates be denoted as

$$(\mathbf{y}^{(1)}, \mathbf{x}^{(1)}), \dots, (\mathbf{y}^{(T)}, \mathbf{x}^{(T)}),$$

where  $\mathbf{y}^{(t)} = (y_1^{(t)}, \dots, y_d^{(t)})$ ,  $\mathbf{x}^{(t)} = (\mathbf{x}_1^{(t)}, \dots, \mathbf{x}_d^{(t)})$ ,  $t = 1, \dots, T$ . In a random sample of size  $n$ , let the corresponding frequencies be denoted as  $n^{(1)}, \dots, n^{(T)}$ . Assuming a probability distribution on the covariates, for  $t = 1, \dots, T$ , let  $p^{(t)}$  be the limit in probability of  $n^{(t)}/n$

as  $n \rightarrow \infty$ . For the simulated likelihood in (4), we have the limit,

$$n^{-1} \ell(\boldsymbol{\beta}, \boldsymbol{\gamma}, \rho) \rightarrow \sum_{t=1}^T p^{(t)} \log h_{\mathbf{Y}}(y_1^{(t)}, \dots, y_d^{(t)}; \mathbf{x}_1^{(t)}, \dots, \mathbf{x}_d^{(t)}), \quad (8)$$

where  $h_{\mathbf{Y}}(\mathbf{y}^{(t)}; \mathbf{x}^{(t)})$  is computed using the method in Genz and Bretz (2002). The limit of the MSLE (as  $n \rightarrow \infty$ ) is the maximum of (8); we denote this limit as  $(\boldsymbol{\beta}^{SL}, \boldsymbol{\gamma}^{SL}, \rho^{SL})$ . Note in passing that the limit of the standard MLE (as  $n \rightarrow \infty$ ) is the maximum of (8) where  $h_{\mathbf{Y}}(\mathbf{y}^{(t)}; \mathbf{x}^{(t)})$  is computed with the 1-dimensional integral method in Johnson and Kotz (1972).

The surrogate log-likelihood based on the MVN density with exchangeable dependence structure is,

$$\ell_{HR}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \rho) = \sum_{i=1}^n \log \left[ c(u_{i1}, \dots, u_{id}; \rho) \prod_{j=1}^d f_{Y_j}(y_{ij}; \mathbf{x}_{ij}) \right],$$

where  $c(u_{i1}, \dots, u_{id}; \rho) = \frac{1}{\sqrt{[1+(d-1)\rho](1-\rho)^{d-1}}} e^{\frac{-\rho}{2(1-\rho)[1+(d-1)\rho]} \left( (d-1)\rho \sum_{j=1}^d q_{ij}^2 - 2 \sum_{j < k} q_{ij} q_{ik} \right)}$  (Žežula, 2009) with  $q_{ij} = \Phi^{-1}(u_{ij}) = \Phi^{-1}[F_{Y_j}(y_{ij}-1; \mathbf{x}_{ij}) + v_{ij} f_{Y_j}(y_{ij}; \mathbf{x}_{ij})]$  as realizations of standard normal random variables. Therefore  $n^{-1} \ell_{HR}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \rho)$  is,

$$\begin{aligned} & -\frac{1}{2} \log[1 + (d-1)\rho] - \frac{d-1}{2} \log(1-\rho) - n^{-1} \frac{\rho}{2(1-\rho)[1+(d-1)\rho]} \times \\ & \left[ (d-1)\rho \sum_{i=1}^n \sum_{j=1}^d q_{ij}^2 - 2 \sum_{i=1}^n \sum_{j < k} q_{ij} q_{ik} \right] + n^{-1} \sum_{i=1}^n \sum_{j=1}^d \log f_{Y_j}(y_{ij}; \mathbf{x}_{ij}). \end{aligned}$$

Then as  $n \rightarrow \infty$ , the limit in probability of  $n^{-1} \ell_{HR}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \rho)$  is,

$$\begin{aligned} & \sum_{t=1}^T p^{(t)} \times \left[ -\frac{1}{2} \log[1 + (d-1)\rho] - \frac{d-1}{2} \log(1-\rho) - \frac{\rho}{2(1-\rho)[1+(d-1)\rho]} \times \right. \\ & \left. \left\{ (d-1)\rho \sum_{j=1}^d \xi_j^{(t)} - 2 \sum_{j < k} \zeta_j^{(t)} \zeta_k^{(t)} \right\} + \sum_{j=1}^d \log f_{Y_j}(y_j^{(t)}; \mathbf{x}_j^{(t)}) \right], \quad (9) \end{aligned}$$

where  $\zeta_j^{(t)}, \xi_j^{(t)}, j = 1, \dots, d, t = 1, \dots, T$  are conditional expectations for the truncated normal distribution that have closed forms. Further details are given in the Appendix. The limit of the HRMLE (as  $n \rightarrow \infty$ ) is the maximum of (9); we denote this limit as  $(\boldsymbol{\beta}^{HR}, \boldsymbol{\gamma}^{HR}, \rho^{HR})$ .

We will compute these limiting HRMLE and MSLE in a variety of situations to show clearly if the HR and SL methods are good. By using these limits, we do not need Monte Carlo simulations for comparisons, and we can quickly vary parameter values and see the effects. The  $p^{(t)}$  in (8) and (9) are the model based probabilities  $h_{\mathbf{Y}}(\mathbf{y}^{(t)}; \mathbf{x}^{(t)})$ , and computed with the 1-dimensional integral method in Johnson and Kotz (1972). For marginal models we use Bernoulli( $\mu$ ), Poisson( $\mu$ ), and negative binomial (NB). For the latter model, we use both the NB1( $\mu, \gamma$ ) and NB2( $\mu, \gamma$ ) parametrization in Cameron and Trivedi (1998); the NB2 parametrization is that used in Lawless (1987). For a count response, we get a finite number of  $\mathbf{y}^{(t)}$  vectors by truncation. The truncation point is chosen to exceed 0.999 for total probabilities. Further, we use only one binary covariate, which is the same for each cluster; this scenario is typical for longitudinal data with time-independent covariates. Other discrete (and time-dependent) covariates can be used, but computations are more time consuming because  $T$  is larger.



$d$	$\rho$	$\rho^{HR}$	$\beta_0$	$\beta_0^{HR}$	$\beta_1$	$\beta_1^{HR}$
2	0.3	0.120	-0.5	-0.499	0.5	0.499
2	0.6	0.255	-0.5	-0.497	0.5	0.497
2	0.8	0.368	-0.5	-0.493	0.5	0.493
5	0.3	0.120	-0.5	-0.498	0.5	0.498
5	0.6	0.255	-0.5	-0.491	0.5	0.491
5	0.8	0.368	-0.5	-0.479	0.5	0.479
10	0.3	0.120	-0.5	-0.496	0.5	0.496
10	0.6	0.255	-0.5	-0.484	0.5	0.484
10	0.8	0.369	-0.5	-0.467	0.5	0.467

Table 1: Limiting HRMLE for MVN copula-based models with marginal logistic regression.

Representative results are shown in Tables 1 and 2 for logistic and NB2 regression, with MSLE results omitted because they were identical with MLE up to three or four decimal places. Therefore, the SL method leads to unbiased estimating equations. Regarding the HR method, by varying the latent correlation  $\rho$  and dimension  $d$ , results are similar for the binary and count responses. There is substantial asymptotic downward bias for the HRMLE of the latent correlation ( $\rho^{HR}$ ), and it seems that there is negligible asymptotic bias for the HRMLE for the parameters that are regression coefficients ( $\beta_0^{HR}, \beta_1^{HR}$ ); note that this slightly increases as either  $d$  or  $\rho$  increases. Calculating the limit for HRMLE of  $\gamma$  ( $\gamma^{HR}$ ) for NB2 regression, it can also be seen that there is substantial asymptotic downward bias for the univariate marginal parameters that are not regression coefficients as the latent correlation  $\rho$  increases. The results in Tables 1 and 2 show that the HR method is adequate with regard to the univariate marginal parameters that are regression coefficients.

$d$	$\rho$	$\rho^{HR}$	$\beta_0$	$\beta_0^{HR}$	$\beta_1$	$\beta_1^{HR}$	$\gamma$	$\gamma^{HR}$
2	0.3	0.191	-0.5	-0.498	0.5	0.495	0.5	0.480
2	0.6	0.397	-0.5	-0.492	0.5	0.483	0.5	0.410
2	0.8	0.550	-0.5	-0.481	0.5	0.466	0.5	0.302
3	0.3	0.191	-0.5	-0.497	0.5	0.492	0.5	0.468
3	0.6	0.394	-0.5	-0.484	0.5	0.472	0.5	0.361
3	0.8	0.545	-0.5	-0.466	0.5	0.446	0.5	0.214

Table 2: Limiting HRMLE for MVN copula-based models with marginal NB2 regression. The truncation point is 10.

After evaluating the adequacy of the MF and SL log-likelihood on finding the peak (MLE), we evaluate if the curvature (Hessian) is also correct for the cases where the HRMLE and MSLE are correct. To check this, we also computed the negative inverse Hessian  $H$  of the limit of the surrogate log-likelihood in (9) and the simulated log-likelihood in (8); because these are limits as  $n \rightarrow \infty$  of  $n^{-1}$  times the log-likelihood,  $H$  is the inverse Fisher information, or equivalently, the covariance matrix for sample size  $n$  is approximately  $n^{-1}H$ . For a comparison, we have also calculated the Hessian at the limit for the standard MLE. For simpler comparisons, we convert to standard errors (SE), say for a sample size of  $n = 100$  (that is, square roots of the diagonals of the above matrices divided by  $n$ ). Some representative results are given in Table 3 and Table 4 for an MVN copula-based model with marginal logistic and NB2 regression, respectively, with the MSLE results omitted because they were again identical with MLE up to three or four decimal places.

The results in Tables 3 and 4 show that the HR method slightly underestimates the SE for the regression parameters. Underestimation of the curvature increases as the dimension

$d$	$\rho$	$\beta_0$	$\beta_1$	ML	HR	ML	HR	ML	HR
				SE( $\hat{\beta}_0$ )		SE( $\hat{\beta}_1$ )		SE( $\hat{\rho}$ )	
2	0.3	-0.5	0.5	0.16	0.15	0.22	0.21	0.11	0.07
2	0.6	-0.5	0.5	0.17	0.16	0.24	0.22	0.08	0.06
2	0.8	-0.5	0.5	0.18	0.16	0.26	0.22	0.05	0.06
5	0.3	-0.5	0.5	0.12	0.10	0.17	0.14	0.05	0.03
5	0.6	-0.5	0.5	0.15	0.11	0.21	0.16	0.05	0.03
5	0.8	-0.5	0.5	0.17	0.12	0.23	0.17	0.03	0.03
10	0.3	-0.5	0.5	0.11	0.08	0.15	0.11	0.03	0.02
10	0.6	-0.5	0.5	0.14	0.09	0.19	0.12	0.04	0.02
10	0.8	-0.5	0.5	0.16	0.09	0.22	0.13	0.03	0.02

Table 3: Standard errors (SE) of the limiting HRMLE and MLE for MVN copula-based models with marginal logistic regression.

$d$	$\rho$	$\beta_0$	$\beta_1$	$\gamma$	ML	HR	ML	HR	ML	HR	ML	HR
					SE( $\hat{\beta}_0$ )		SE( $\hat{\beta}_1$ )		SE( $\hat{\gamma}$ )		SE( $\hat{\rho}$ )	
2	0.3	-0.5	0.5	0.5	0.11	0.11	0.15	0.14	0.15	0.14	0.08	0.07
2	0.6	-0.5	0.5	0.5	0.13	0.11	0.16	0.15	0.15	0.13	0.06	0.06
2	0.8	-0.5	0.5	0.5	0.13	0.12	0.17	0.15	0.17	0.12	0.04	0.04
3	0.3	-0.5	0.5	0.5	0.10	0.09	0.13	0.12	0.12	0.11	0.06	0.04
3	0.6	-0.5	0.5	0.5	0.12	0.10	0.15	0.13	0.13	0.10	0.05	0.04
3	0.8	-0.5	0.5	0.5	0.13	0.10	0.16	0.13	0.15	0.09	0.03	0.03

Table 4: Standard errors (SE) of the limiting HRMLE and MLE for MVN copula-based models with marginal NB2 regression. The truncation point is 10.

$d$  and/or the latent correlation  $\rho$  increases. The HR method leads to underestimation of the SE, because it is using information on jitters that are not in the observed data.

To sum up, the asymptotics show that the maximum SL method is as good as maximum likelihood, while the maximum surrogate log-likelihood using the CE leads to approximate asymptotic unbiasedness for (some) univariate marginal parameters and not for the latent correlation parameters, because the jittering is univariate and does not account for dependence. Although we show the details only for exchangeable dependence, we expect the above results to hold in general, as well as to apply to different dependence structures. Section 5 contains small sample efficiency calculations using both exchangeable and AR(1) dependence.

Closing this section, we explain why the the surrogate log-likelihood using the CE fails. It is easiest notationally to indicate what is happening with bivariate discretized normal:

$$Y_1 = j_1, Y_2 = j_2 \quad \text{iff} \quad z_{1,j_1-1} < Z_1 \leq z_{1,j_1}, \quad z_{2,j_2-1} < Z_2 \leq z_{2,j_2}$$

where  $z_{1,j}, z_{2,j}$  are cutpoints, and  $(Z_1, Z_2)$  is bivariate standard normal with correlation  $\rho$ . Then conditioned on  $\{Y_1 = j_1, Y_2 = j_2\}$ ,  $(Z_1, Z_2)$  is dependent with a truncated bivariate normal distribution. This means the jittered variables  $(V_1, V_2)$  have to be dependent to get the correct conditional distribution. So, for jittering to be asymptotically unbiased, a sequential approach would be needed based on the previous estimates of the latent correlations.

## 4.2 Computation of MVN rectangle probabilities

In this section, we describe how the high-dimensional multivariate normal rectangle probability is computed by Madsen (2009) and Madsen and Fang (2011) using importance sampling, and compare it with the naive simulation method, and the method in Genz and Bretz (2002).

For an integral in the context of an MVN rectangle probability in (6), naive simulation uses,

$$m^{-1} \sum \mathbf{1}(a_j < z_j < b_j, j = 1, \dots, d),$$

where  $\mathbf{1}(A)$  denotes the indicator function of the set  $A$  and  $z_1, \dots, z_d$  are  $m$  iid variates from  $\phi_{\mathbf{R}}$ . Importance sampling gives,

$$\int_{a_1}^{b_1} \cdots \int_{a_d}^{b_d} [\phi_{\mathbf{R}}(z_1, \dots, z_d)/g(z_1, \dots, z_d)] g(z_1, \dots, z_d) dz_1 \dots dz_d,$$

where  $g$  is a closed-form density from which it is easy to simulate.

Estimation is via,

$$m^{-1} \sum \phi_{\mathbf{R}}(z_1, \dots, z_d)/g(z_1, \dots, z_d),$$

where  $z_1, \dots, z_d$  are iid simulated from  $g$ . Better choices of  $g$  have small variance for  $w = \phi_{\mathbf{R}}/g$ , such that  $w$  is bounded (in which case one can bound the variance). The penalty for a bad  $g$  can be longer run times than for a general Monte Carlo simulation without importance sampling.

Madsen (2009) and Madsen and Fang (2011) implement the integration in (6) by transforming  $z_1 = \Phi^{-1}(\omega_1), \dots, z_d = \Phi^{-1}(\omega_d)$  to get,

$$\int_{\Phi(a_1)}^{\Phi(b_1)} \cdots \int_{\Phi(a_d)}^{\Phi(b_d)} \frac{\phi_{\mathbf{R}}(\Phi^{-1}(\omega_1), \dots, \Phi^{-1}(\omega_d))}{\phi(\Phi^{-1}(\omega_1)) \cdots \phi(\Phi^{-1}(\omega_d))} d\omega_1 \dots d\omega_d,$$

where  $\phi$  is the standard normal density. The Denuit and Lambert (2005) uniform extension in this case corresponds to evaluating the above integral based on a  $d$ -variate uniform random sample  $\boldsymbol{\Omega}_k = (\Omega_{1,k}, \dots, \Omega_{d,k})$ , where  $\Omega_{j,k}$ ,  $j = 1, \dots, d$  are uniform in the interval  $\Phi(a_j)$  to  $\Phi(b_j)$  for  $j = 1, \dots, d$ . Approximation is via,

$$m^{-1} \sum_{k=1}^m \frac{\phi_{\mathbf{R}}(\Phi^{-1}(\omega_{1,k}), \dots, \Phi^{-1}(\omega_{d,k})) \prod_{j=1}^d (\Phi(b_j) - \Phi(a_j))}{\phi(\Phi^{-1}(\omega_{1,k})) \cdots \phi(\Phi^{-1}(\omega_{d,k}))},$$

where  $\boldsymbol{\omega}_k = (\omega_{1,k}, \dots, \omega_{d,k})$  are realizations of the jitters  $\boldsymbol{\Omega}_k$ .

In Table 5, comparisons of the accuracy of the MF ( $m = 10^3$  and  $m = 10^4$ ), the naive ( $m = 10^4$ ) method, and the method in Genz and Bretz (2002) are presented. The accuracy comparisons are for the computation of the equicorrelated rectangle probability of the form  $\Pr(-a \leq Z_j \leq a, j = 1, \dots, d)$  for dimensions  $d = 5, 10, 20$  and correlations  $\rho = 0.3, 0.6, 0.8$ . These probabilities are also computed with the numerically accurate method in Johnson and Kotz (1972), and the results are identical up to three or four decimal places with the ones found by the method in Genz and Bretz (2002). For these MVN rectangle probabilities, we simulate based on the MF method, keeping track of the values of  $w(\cdot)$  when simulating from the  $d$ -variate uniform random sample. Based on the sample variance of  $w(\cdot)$ , we estimate the achieved accuracy at the number of replications  $m$ , that is,  $\text{SD} = \sqrt{\text{Var}[w(\cdot)]/m} < \text{accuracy}$ . This calculation is simple for the naive method.

In Table 5, it is clear that the MF method (even with  $m = 10^4$ ) gets worse as

1. the dimension  $d$  increases;

$d$	$a$	$\rho$	GB	MF	SD	MF	SD	Naive	SD
				$m = 10^3$			$m = 10^4$		$m = 10^4$
5	1	0.3	0.176	0.176	0.001	0.176	$< 10^{-3}$	0.176	0.004
		0.6	0.266	0.266	0.005	0.267	0.001	0.267	0.004
		0.8	0.391	0.382	0.014	0.395	0.005	0.393	0.005
	2	0.3	0.808	0.823	0.023	0.809	0.006	0.809	0.004
		0.6	0.847	0.863	0.064	0.850	0.016	0.847	0.004
		0.8	0.883	0.808	0.109	0.862	0.033	0.883	0.003
	4	0.3	1.000	1.049	0.055	0.996	0.012	1.000	$< 10^{-3}$
		0.6	1.000	1.025	0.133	0.981	0.032	1.000	$< 10^{-3}$
		0.8	1.000	0.805	0.112	0.915	0.044	1.000	$< 10^{-3}$
10	1	0.3	0.038	0.037	$< 10^{-3}$	0.038	$< 10^{-3}$	0.039	0.002
		0.6	0.110	0.107	0.003	0.111	0.001	0.113	0.003
		0.8	0.267	0.244	0.020	0.270	0.007	0.270	0.004
	2	0.3	0.674	0.641	0.040	0.677	0.010	0.675	0.005
		0.6	0.768	0.741	0.192	0.753	0.033	0.763	0.004
		0.8	0.840	0.647	0.293	0.659	0.063	0.837	0.004
	4	0.3	0.999	0.912	0.111	0.972	0.024	1.000	$< 10^{-3}$
		0.6	0.999	1.001	0.420	0.890	0.057	1.000	$< 10^{-3}$
		0.8	1.000	0.621	0.289	0.624	0.069	1.000	$< 10^{-3}$
20	1	0.3	0.002	0.002	$< 10^{-3}$	0.002	$< 10^{-3}$	0.002	$< 10^{-3}$
		0.6	0.024	0.023	0.001	0.025	$< 10^{-3}$	0.025	0.002
		0.8	0.156	0.121	0.024	0.168	0.010	0.159	0.004
	2	0.3	0.493	0.422	0.016	0.498	0.019	0.496	0.005
		0.6	0.670	0.354	0.064	0.633	0.047	0.671	0.005
		0.8	0.792	0.478	0.451	0.673	0.283	0.792	0.004
	4	0.3	0.999	0.672	0.036	0.959	0.091	0.998	$< 10^{-3}$
		0.6	0.999	0.347	0.107	0.748	0.073	0.999	$< 10^{-3}$
		0.8	0.999	0.514	0.503	0.699	0.341	0.999	$< 10^{-3}$

Table 5: Comparisons of the accuracy of the MF, the naive method, and the method in Genz and Bretz (GB, 2002) for the equicorrelated rectangle probability of the form  $\Pr(-a \leq Z_j \leq a, j = 1, \dots, d)$  for dimensions  $d = 5, 10, 20$  and correlations  $\rho = 0.3, 0.6, 0.8$ . The computed probabilities with the GB method are identical up to three or four decimal places with the ones found by the numerically accurate method in Johnson and Kotz (1972).

2. the latent correlation  $\rho$  increases;
3. the limits (integrated region) increase;

and even the naive method is much better; it has 2 decimal place accuracy for  $m = 10^4$  replications. The use of jitters with  $m = 10^3$ , as in Madsen and Fang (2011), is highly inefficient even in a low dimension. Therefore, the MF method is a very inefficient way to compute a multivariate normal rectangle probability, which means that even  $m = 10^4$  is far from sufficient to achieve a desired accuracy. It is also shown to be difficult to even estimate the SD, because the integrand has no bound.

Regarding equation (3), this is much worse than approximating many  $d$ -dimensional MVN rectangle probabilities separately, that is, separate jitters of each  $d$ -dimensional probability. This is actually a worse way to do simulated likelihood compared with what we mention above.

## 5 Simulations

In this section we performed several simulation studies to assess the performance of the HR, MF, and SL methods. We used structured latent correlation matrices for the MVN copula. For exchangeable dependence, we took  $\mathbf{R}$  as  $(1 - \rho)\mathbf{I}_d + \rho\mathbf{J}_d$ , where  $\mathbf{I}_d$  is the identity matrix of order  $d$  and  $\mathbf{J}_d$  is the  $d \times d$  matrix of 1s. For AR(1) dependence,  $\mathbf{R}$  was taken as  $(\rho^{|j-k|})_{1 \leq j, k \leq d}$ .

For marginal models we used Bernoulli( $\mu$ ), Poisson( $\mu$ ), NB1( $\mu, \gamma$ ), and NB2( $\mu, \gamma$ ) parametrization of the negative binomial distribution. For the covariates and regression parameters, we chose  $p = 2$ ,  $\mathbf{x}_j = (\mathbf{1}, x_j)^\top$ , where  $x_j$  drawn from a  $U[-1, 1]$ , and let  $\mu$  depend on the covariates, that is  $\eta(\mu_j) = (\beta_0 + \beta_1 x_j)$ ,  $j = 1, \dots, d$ , where  $\beta_1 = -\beta_0 = 0.5$ . For the link function  $\eta$ , we took the log link function for Poisson and NB regression, and the logit link function or the probit link function for binary regression. Note also that binary and Poisson regression  $\gamma$  is null, while for NB1 and NB2 regression  $\gamma$  is scalar ( $r = 1$ ).

### 5.1 Assessing the variability due only to jittering

As a first step, we assess the variability of the HR and MF estimates over different sets of jitters  $\mathbf{V}_k$ ,  $k = 1, \dots, m$ . Our goal is to define the value of  $m$  for which different sets of jitters for the same data set reproduce the same results. To this end, we fixed one simulated dataset and used 5 sets of random jitters  $\mathbf{V}_k$ ,  $k = 1, \dots, m$ ; the number of jitters  $m$  was set at  $m = 10, 10^2, 10^3, 10^4$ . The following experiments are typical of the results that were obtained.

We fixed one simulated data set with  $n = 100$  observations from the bivariate normal copula with exchangeable moderate dependence ( $\rho = 0.5$ ) and marginal logistic regression. Table 6 shows the variability of the HR and MF estimators over 5 different sets of jitters as compared with ML estimates and standard errors; for  $d = 2$  there is a numerical accurate calculation of the bivariate normal rectangle probabilities in (4). To investigate if these results hold for higher dimensions, we also fixed one simulated data set with  $n = 100$  observations from the MVN ( $d = 5$ ) copula with strong AR(1) dependence ( $\rho = 0.8$ ), and NB2 regression with large overdispersion  $\gamma = 2$ . Table 7 shows the variability of the HR and MF estimators over 5 different sets of jitters as compared with maximum SL estimates and standard errors. These results and other computations that we have done with latent correlation  $\rho$  varying from 0 to 0.9 in 0.1 increments confirm that for the HR (MF) method,  $m = 100$  ( $m = 1000$ ) is sufficient to produce the same estimates over different sets of jitters. Note that the SEs of the ML and maximum SL estimates are obtained via the square roots of the diagonals of the inverse Hessian computed numerically during the maximization process.

### 5.2 Small-sample efficiency

In this subsection, we gauge the small-sample efficiency of the HR, MF, and SL methods. The number of jitters was set at  $m = 10, 10^2, 10^3, 10^4$ , and the sample size at  $n = 100, 300, 500$ . The following experiments are typical of the results that were obtained. We report simulations from small sample sizes ( $n = 100$ ), since the additional simulation results do not show sensitivity to  $n$ . Note that the theoretical variances of the maximum SL and MF estimates are obtained via the gradients and the Hessian computed numerically during the maximization process, while the HR variances are obtained by averaging the latter over  $m$  runs.

For  $n = 100$ ,  $10^4$  random samples of size  $n$  were generated from the bivariate normal copula with exchangeable moderate dependence ( $\rho = 0.5$ ), and marginal logistic regression. Table 8 contains the parameter values, the bias, variance (Var), and mean square errors

MF ( $m = 10$ )			HR ( $m = 10$ )			
set	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\rho}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\rho}$
1	-0.51	0.40	0.20	-0.51	0.40	0.15
2	-0.54	0.44	0.33	-0.51	0.42	0.18
3	-0.52	0.41	0.27	-0.51	0.41	0.16
4	-0.50	0.45	0.27	-0.51	0.42	0.16
5	-0.51	0.41	0.25	-0.51	0.41	0.21
MF ( $m = 10^2$ )			HR ( $m = 10^2$ )			
set	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\rho}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\rho}$
1	-0.52	0.33	0.39	-0.51	0.41	0.16
2	-0.52	0.45	0.32	-0.51	0.42	0.18
3	-0.52	0.47	0.39	-0.51	0.42	0.18
4	-0.52	0.41	0.33	-0.51	0.42	0.17
5	-0.49	0.46	0.30	-0.51	0.42	0.17
MF ( $m = 10^3$ )			HR ( $m = 10^3$ )			
set	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\rho}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\rho}$
1	-0.52	0.44	0.35	-0.51	0.42	0.17
2	-0.51	0.45	0.38	-0.51	0.42	0.17
3	-0.51	0.45	0.36	-0.51	0.42	0.17
4	-0.49	0.49	0.42	-0.51	0.42	0.17
5	-0.50	0.47	0.32	-0.51	0.42	0.17
MF ( $m = 10^4$ )			HR ( $m = 10^4$ )			
set	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\rho}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\rho}$
1	-0.51	0.45	0.36	-0.51	0.42	0.17
2	-0.52	0.46	0.38	-0.51	0.42	0.17
3	-0.51	0.45	0.38	-0.51	0.42	0.17
4	-0.50	0.46	0.37	-0.51	0.42	0.17
5	-0.51	0.45	0.34	-0.51	0.42	0.17
ML	$\hat{\beta}_0$	SE	$\hat{\beta}_1$	SE	$\hat{\rho}$	SE
	-0.51	0.03	0.47	0.07	0.43	0.02

Table 6: Variability of the MF and HR estimates with different sets of jitters  $\mathbf{V}_k$ ,  $k = 1, \dots, m$ , along with ML estimates and standard errors (SE) for a fixed simulated data set with  $n = 100$  observations from the bivariate normal copula model with moderate dependence ( $\rho = 0.5$ ) and logistic regression. The number of jitters was set at  $m = 10, 10^2, 10^3, 10^4$ .

(MSE) of the MF, HR and ML (for  $d = 2$  there is a numerical accurate calculation of the bivariate probabilities) estimates, along with the average of their theoretical variances ( $\bar{V}$ ).

For  $n = 100$ ,  $10^4$  random samples of size  $n$  were generated from the MVN ( $d = 5$ ) copula with strong AR(1) dependence ( $\rho = 0.8$ ) and NB2 regression with large overdispersion  $\gamma = 2$ . Table 9 contains the parameter values, the bias, variance (Var) and mean square errors (MSE) of the MF, HR and maximum SL estimates, along with the average of their theoretical variances ( $\bar{V}$ ). Because HR and MF likelihood takes much longer with larger  $m$  and  $d$ , we ran fewer replications on a subset for the 5-dimensional case. With an Intel Core Duo  $2 \times 2.53$ GHz processor, the computing times in an R program for SL, MF and HR with  $m = 10^3$  jitters, averaged about 1, 6, and 6 minutes respectively; the time is about 1 hour for HR and MF with  $m = 10^4$  jitters. To this end, simulations have been restricted to



set	MF ( $m = 10$ )				HR ( $m = 10$ )			
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}$	$\hat{\rho}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}$	$\hat{\rho}$
1	-0.44	0.49	1.14	0.50	-0.42	0.48	1.14	0.50
2	-0.39	0.49	1.13	0.53	-0.42	0.46	1.14	0.50
3	-0.45	0.53	1.09	0.50	-0.43	0.50	1.11	0.48
4	-0.43	0.42	1.16	0.53	-0.42	0.48	1.15	0.50
5	-0.39	0.56	1.08	0.52	-0.42	0.50	1.14	0.49
set	MF ( $m = 10^2$ )				HR ( $m = 10^2$ )			
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}$	$\hat{\rho}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}$	$\hat{\rho}$
1	-0.37	0.54	1.13	0.58	-0.42	0.49	1.13	0.49
2	-0.40	0.47	1.20	0.57	-0.42	0.49	1.12	0.50
3	-0.41	0.47	1.17	0.56	-0.42	0.48	1.12	0.49
4	-0.42	0.50	1.11	0.53	-0.42	0.49	1.13	0.49
5	-0.43	0.51	1.17	0.55	-0.42	0.48	1.13	0.49
set	MF ( $m = 10^3$ )				HR ( $m = 10^3$ )			
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}$	$\hat{\rho}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}$	$\hat{\rho}$
1	-0.38	0.48	1.13	0.58	-0.42	0.48	1.13	0.49
2	-0.38	0.46	1.24	0.61	-0.42	0.48	1.13	0.49
3	-0.37	0.48	1.16	0.58	-0.42	0.48	1.13	0.49
4	-0.39	0.48	1.20	0.57	-0.42	0.49	1.12	0.49
5	-0.41	0.49	1.16	0.56	-0.42	0.48	1.13	0.49
SL	$\hat{\beta}_0$	SE	$\hat{\beta}_1$	SE	$\hat{\gamma}$	SE	$\hat{\rho}$	SE
	-0.37	0.15	0.49	0.09	1.96	0.40	0.80	0.03

Table 7: Variability of the MF and HR estimates with different sets of jitters  $\mathbf{V}_k$ ,  $k = 1, \dots, m$ , along with maximum SL estimates and standard errors (SE) for a fixed simulated data set with  $n = 100$  observations from the MVN ( $d = 5$ ) copula model with strong dependence ( $\rho = 0.8$ ) and NB2 regression. The number of jitters was set at  $m = 10, 10^2, 10^3$ .

$m = 10^3$ .

Conclusions from the values in the tables and other computations that we have done are the following:

1. The SL method is highly efficient according to the simulated biases and variances.
2. The HR and MF methods yield estimates that are almost as good as the ML and maximum SL estimates for the regression parameters.
3. The HR and MF methods underestimate the univariate marginal parameters that are not regression coefficients as the latent correlation increases.
4. The efficiency of HR and MF methods is low for the latent correlation, for a wide range of  $\rho$  values. The dependence parameters (latent correlations) are underestimated.
5. For the MF method, it appears that there are improvements with  $m = 10^4$ , so one might wonder if  $m = 10^6$  would be better. This number of jitters is prohibitive with respect to the computational time as the dimension increases. Further, theoretically, there are still problems; see subsection 4.2.
6. Overall, efficiencies of HR do not change considerably, when  $m$  changes.
7. The variances for the MF and HR estimates are underestimated for the non-regression parameters, the intercept, and the regression parameters of discrete covariates.

Method	$m$	$n\text{Bias}$	$n\text{Var}$	$n\text{MSE}$	$nV$	
MF	10	-0.56	3.02	3.02	2.61	
	$10^2$	-0.54	3.02	3.02	2.69	
	$10^3$	-0.58	3.03	3.03	2.75	
	$10^4$	-0.63	3.02	3.03	2.80	
HR	$\beta_0 = -0.5$	10	-0.60	3.00	3.00	2.48
		$10^2$	-0.60	2.99	3.00	2.48
		$10^3$	-0.60	2.99	2.99	2.48
		$10^4$	-0.60	2.99	2.99	2.47
ML		-0.73	3.00	3.00	2.92	
MF	10	0.70	6.54	6.54	6.56	
	$10^2$	0.76	6.49	6.50	6.48	
	$10^3$	0.83	6.45	6.45	6.41	
	$10^4$	0.82	6.41	6.42	6.35	
HR	$\beta_1 = 0.5$	10	0.70	6.57	6.58	6.67
		$10^2$	0.70	6.57	6.57	6.67
		$10^3$	0.70	6.56	6.57	6.67
		$10^4$	0.70	6.56	6.57	6.67
ML		0.85	6.31	6.31	6.15	
MF	10	-20.93	0.77	5.15	0.96	
	$10^2$	-15.17	0.74	3.05	1.00	
	$10^3$	-11.19	0.78	2.03	1.06	
	$10^4$	-8.30	0.87	1.56	1.12	
HR	$\rho = 0.5$	10	-30.42	0.42	9.67	0.90
		$10^2$	-30.40	0.37	9.61	0.90
		$10^3$	-30.39	0.36	9.59	0.90
		$10^4$	-30.39	0.36	9.59	0.90
ML		-0.73	1.84	1.84	1.82	

Table 8: Small sample of sizes  $n = 100$  simulations ( $10^4$  replications) and resultant biases and mean square errors (MSE) and variances (Var), along with average theoretical variances scaled by  $n$ , for the MF, HR, and ML of the regression and copula parameters for the bivariate normal copula model with moderate dependence and logistic regression. The number of jitters was set at  $m = 10, 10^2, 10^3, 10^4$ .

8. The variances for the MF and HR estimates are overestimated for the regression parameters of continuous covariates that do not vary a lot.

These conclusions justify why Madsen (2009) and Madsen and Fang (2011) do not report latent correlations to their applications and studies of efficiency.

## 6 The toenail infection data

In this section we re-analyze the toenail infection data in Madsen and Fang (2011). The data were obtained from a randomized, double-blind, parallel group, multicenter study for the comparison of two oral treatments for toenail dermatophyte onychomycosis (De Backer et al., 1996; Molenberghs and Verbeke, 2005). Subjects were followed during 12 weeks (3 months) of treatment and followed further, up to a total of 48 weeks (12 months). Measurements

Method	$m$	$n\text{Bias}$	$n\text{Var}$	$n\text{MSE}$	$nV$	
MF	10	-5.94	2.48	2.84	1.10	
	$10^2$	-5.89	2.49	2.83	1.15	
	$10^3$	-5.82	2.50	2.84	1.19	
HR	$\beta_0 = -0.5$	10	-5.88	2.43	2.78	1.03
	$10^2$	-5.89	2.43	2.77	1.04	
	$10^3$	-5.88	2.43	2.77	1.04	
SL		-1.36	2.33	2.35	2.28	
MF	10	-0.72	1.09	1.10	1.32	
	$10^2$	-0.55	1.07	1.08	1.28	
	$10^3$	-0.35	1.05	1.05	1.26	
HR	$\beta_1 = 0.5$	10	-1.06	1.03	1.04	1.39
	$10^2$	-1.07	1.02	1.03	1.39	
	$10^3$	-1.07	1.02	1.03	1.39	
SL		0.12	0.89	0.89	0.83	
MF	10	-78.13	8.36	69.41	6.51	
	$10^2$	-77.77	8.13	68.62	6.51	
	$10^3$	-77.18	8.01	67.57	6.54	
HR	$\gamma = 2$	10	-77.27	8.71	68.42	6.62
	$10^2$	-77.24	8.68	68.34	6.62	
	$10^3$	-77.24	8.68	68.33	6.62	
SL		-1.32	19.03	19.05	18.59	
MF	10	-28.70	0.20	8.44	0.12	
	$10^2$	-25.96	0.18	6.92	0.11	
	$10^3$	-23.95	0.16	5.90	0.11	
HR	$\rho = 0.8$	10	-33.23	0.19	11.23	0.13
	$10^2$	-33.22	0.18	11.21	0.13	
	$10^3$	-33.21	0.18	11.21	0.13	
SL		0.88	0.11	0.12	0.11	

Table 9: Small sample of sizes  $n = 100$  simulations ( $10^4$  replications) and resultant biases and mean square errors (MSE) and variances (Var), along with average theoretical variances scaled by  $n$ , for the HR, MF, and SL of the regression and copula parameters for the MVN ( $d = 5$ ) copula model with strong dependence and NB2 regression. The number of jitters was set at  $m = 10, 10^2, 10^3$ .

were taken at baseline, every month during treatment, and every 3 months afterwards, resulting in a maximum of 7 measurements per subject. The observations are coded as 1 if the subject’s infection was severe and 0 otherwise. The question of interest was whether the percentage of severe infections decreased over time, and whether that evolution was different for the two treatment groups. In accordance with Madsen and Fang (2011), we use the 224 subjects observed at all seven time points, though the SL method does not depend on a constant “cluster” size  $d$ . In this data example, Madsen and Fang (2011) assumed an exchangeable structure for the MVN copula with logistic regressions, and hence this can be easily analyzed with the standard ML method. ML estimation is possible for an exchangeable structure, since the 7-dimensional rectangle reduces to 1-dimensional integral.

However, for longitudinal data, it is common practice to use various parametric correlation matrices, so we have also calculated the estimates under an AR(1), and Markov

Dependence Link	AR(1)				Markov			
	probit		logit		probit		logit	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Intercept	-0.418	0.116	-0.642	0.189	-0.385	0.119	-0.587	0.194
Treatment	-0.006	0.160	0.009	0.262	-0.011	0.165	-0.006	0.268
Time	-0.099	0.016	-0.188	0.032	-0.111	0.016	-0.208	0.033
Trt.×time	-0.022	0.023	-0.054	0.048	-0.021	0.024	-0.048	0.048
$\rho$	0.923	0.015	0.924	0.015	0.952	0.010	0.952	0.010
$\ell$	-412.58		-410.06		-405.71		-403.49	

Table 10: Maximum SL log-likelihoods ( $\ell$ 's), estimates and standard errors (SE) for the the toenail infection data.

dependence with the SL method. Using a Markov model, the AR(1) model is generalized to times that are unequally-spaced, and this is the case for toenail infection data. For Markov dependence,  $\mathbf{R}$  is taken as  $(\rho^{|t_j-t_k|})_{1 \leq j, k \leq d}$ . Further, both logit and probit links are used for the marginal binary regressions. In Table 10, we report the resulting maximum SL log-likelihoods ( $\ell$ 's), estimates and standard errors (SE) of the MVN copula-based models with binary regression. The SL log-likelihoods show that Markov dependence is marginally better than AR(1) dependence, and both are far better than exchangeable dependence (Sabo and Chaganty, 2011, Table 1). Further, logistic regression is slightly better than probit regression.

Based on our analysis and the exchangeable analysis, the standard errors show the time effect to be highly significant, and the treatment by time interaction insignificant; hence there is no (significant) difference in evolution between both treatment groups. However, for an AR(1) or Markov structure, the  $p$ -value for the treatment by time interaction is smaller than its counterpart for the exchangeable analysis. Generally speaking, this implies that ignoring the actual correlation structure in the data could lead to invalid conclusions, although this was not crucial in this example.

## 7 Discussion

In this paper we have studied two “approximate” likelihood estimation methods based on the CE of a discrete random variable. For the binary, Poisson, and negative binomial regression models with the MVN copula, we have shown that these methods lead to substantial downward bias for the estimates of the latent correlation and the univariate marginal parameters that they are not regression coefficients; for the latter parameters only for strong dependence.

We have shown that the simulated likelihood in Madsen (2009) and Madsen and Fang (2011) is a very inefficient way to compute a multivariate normal rectangle probability, since the importance weight is not bounded, and even the naive method is much better. The inefficiency of their method also yields to the claim that the GEE (Liang and Zeger, 1986) is more efficient than the ML. A moment based estimate cannot be better than a maximum (simulated) likelihood estimate, and this has to do with the inefficiency of the MF method. See Sabo and Chaganty (2011) and Song et al. (2011) for further criticism.

We have implemented a simulated likelihood method, where the rectangles are converted to bounded integrands via the method in Genz and Bretz (2002), and hence the computational and statistical efficiency of simulated likelihood is substantially improved,

and actually is as good as maximum likelihood as shown for dimension 10 or lower. We expect our findings to hold in higher dimensions. Although there is an issue of computational burden as the dimension and the sample size increase, this will become marginal, as computing technology is advancing rapidly.

## Appendix. Derivation of $\xi_j^{(t)}$ and $\zeta_j^{(t)}$

Split the  $n$  observations into the distinct sets  $n^{(t)} = \{i : \mathbf{y}_i = \mathbf{y}^{(t)}, \mathbf{x}_i = \mathbf{x}^{(t)}\}$ . As  $n \rightarrow \infty$ , then with convergence in probability,

$$\begin{aligned} n^{-1} \sum_{i \in n^{(t)}} q_{ij}^2 &= n^{-1} \sum_{i \in n^{(t)}} [\Phi^{-1}\{F_{Y_j}(y_{ij} - 1; \mathbf{x}_{ij}) + v_{ij} f_{Y_j}(y_{ij}; \mathbf{x}_{ij})\}]^2 \\ &\rightarrow p^{(t)} \int_0^1 [\Phi^{-1}\{F_{Y_j}(y_j^{(t)} - 1; \mathbf{x}_j^{(t)}) + v f_{Y_j}(y_j^{(t)}; \mathbf{x}_j^{(t)})\}]^2 dv \\ &:= p^{(t)} \xi_j^{(t)} \end{aligned}$$

and

$$\begin{aligned} n^{-1} \sum_{i \in n^{(t)}} q_{ij} q_{ik} &= n^{-1} \sum_{i \in n^{(t)}} \Phi^{-1}\{F_{Y_j}(y_{ij} - 1; \mathbf{x}_{ij}) + v_{ij} f_{Y_j}(y_{ij}; \mathbf{x}_{ij})\} \times \\ &\quad \Phi^{-1}\{F_{Y_k}(y_{ik} - 1; \mathbf{x}_{ik}) + v_{ik} f_{Y_k}(y_{ik}; \mathbf{x}_{ik})\} \\ &\rightarrow p^{(t)} \int_0^1 \Phi^{-1}\{F_{Y_j}(y_j^{(t)} - 1; \mathbf{x}_j^{(t)}) + v_j f_{Y_j}(y_j^{(t)}; \mathbf{x}_j^{(t)})\} dv_j \times \\ &\quad \int_0^1 \Phi^{-1}\{F_{Y_k}(y_k^{(t)} - 1; \mathbf{x}_k^{(t)}) + v_k f_{Y_k}(y_k^{(t)}; \mathbf{x}_k^{(t)})\} dv_k \\ &:= p^{(t)} \zeta_j^{(t)} \zeta_k^{(t)}. \end{aligned}$$

Note that  $\xi_j^{(t)}, \zeta_j^{(t)}$ ,  $j = 1, \dots, d, t = 1, \dots, T$  are conditional expectations that have closed forms. For example with  $Z \sim N(0, 1)$ ,

$$\begin{aligned} \zeta_j^{(t)} &= \int_0^1 \Phi^{-1}\{F_{Y_j}(y_j^{(t)} - 1; \mathbf{x}_j^{(t)}) + v_j f_{Y_j}(y_j^{(t)}; \mathbf{x}_j^{(t)})\} dv_j \\ &= \frac{1}{f_{Y_j}(y_j^{(t)}; \mathbf{x}_j^{(t)})} \int_{\Phi^{-1}\{F_{Y_j}(y_j^{(t)} - 1; \mathbf{x}_j^{(t)})\}}^{\Phi^{-1}\{F_{Y_j}(y_j^{(t)}; \mathbf{x}_j^{(t)})\}} z \phi(z) dz \\ &= E[Z | \Phi^{-1}\{F_{Y_j}(y_j^{(t)} - 1; \mathbf{x}_j^{(t)})\} \leq Z \leq \Phi^{-1}\{F_{Y_j}(y_j^{(t)}; \mathbf{x}_j^{(t)})\}] \end{aligned}$$

and

$$\xi_j^{(t)} = E[Z^2 | \Phi^{-1}\{F_{Y_j}(y_j^{(t)} - 1; \mathbf{x}_j^{(t)})\} \leq Z \leq \Phi^{-1}\{F_{Y_j}(y_j^{(t)}; \mathbf{x}_j^{(t)})\}].$$

## Acknowledgements

Special thanks to Professor Harry Joe, University of British Columbia, for helpful comments and suggestions.

## References

- Ashford, J. R. and Sowden, R. R. (1970). Multivariate probit analysis. *Biometrics*, 26:535–546.
- Bhat, C. and Sidharthan, R. (2011). A simulation evaluation of the maximum approximate composite marginal likelihood (MACML) estimator for mixed multinomial probit models. *Transportation Research Part B*, 45:940–953.
- Cameron, A. C. and Trivedi, P. K. (1998). *Regression Analysis of Count Data*. Cambridge University Press.
- De Backer, M., De Keyser, P., De Vroey, C., and Lesaffre, E. (1996). A 12-week treatment for dermatophyte toe onychomycosis: Terbafine 250 mg/day vs. Itraconazole 200 mg/day a double-blind comparative trial. *British Journal of Dermatology*, 134:16–17.
- Demidenko, E. (2004). *Mixed Models: Theory and Applications*. John Wiley & Sons, Hoboken, New Jersey.
- Denuit, M. and Lambert, P. (2005). Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis*, 93(1):40–57.
- Frank, M. J. (1979). On the simultaneous associativity of  $F(x, y)$  and  $x + y - F(x, y)$ . *Aequationes Mathematicae*, 19(2-3):194–226.
- Genz, A. (1992). Numerical computation of the multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1:141–150.
- Genz, A. and Bretz, F. (2002). Methods for the computation of the multivariate  $t$ -probabilities. *Journal of Computational and Graphical Statistics*, 11:950–971.
- Genz, A. and Bretz, F. (2009). *Computation of Multivariate Normal and  $t$  Probabilities*. Springer.
- Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. (2012). *mvtnorm: Multivariate Normal and  $t$  Distributions*. R package version 0.9-9992.
- Hausman, J. and Wise, D. (1978). A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica*, 45:319–339.
- Heinen, A. and Rengifo, E. (2007). Multivariate autoregressive modelling of time series count data using copulas. *Journal of Empirical Finance*, 14:564–583.
- Heinen, A. and Rengifo, E. (2008). Multivariate reduced rank regression in non-Gaussian contexts using copulas. *Computational Statistics & Data Analysis*, 52:2931–2944.
- Joe, H. (1995). Approximations to multivariate normal rectangle probabilities based on conditional expectations. *Journal of the American Statistical Association*, 90(431):957–964.
- Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall, London.
- Joe, H. (2008). Accuracy of laplace approximation for discrete response mixed models. *Computational Statistics and Data Analysis*, 52(12):5066–5074.
- Joe, H., Chou, L. W., and Zhang, H. (2011). *mprobit: Multivariate probit model for binary/ordinal response*. R package version 0.9-3.
- Johnson, N. L. and Kotz, S. (1972). *Continuous Multivariate Distributions*. Wiley, New York.
- Kiefer, N. M. (1982). Testing for dependence in multivariate probit models. *Biometrika*, 69:161–166.
- Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics*, 15(3):209–225.
- Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22.
- Madsen, L. (2009). Maximum likelihood estimation of regression parameters with spa-



- tially dependent discrete data. *Journal of Agricultural, Biological, and Environmental Statistics*, 14:375–391.
- Madsen, L. and Fang, Y. (2011). Joint regression analysis for discrete longitudinal data. *Biometrics*, 67:1171–1175.
- Meester, S. and MacKay, J. (1994). A parametric model for cluster correlated categorical data. *Biometrics*, 50:954–963.
- Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer.
- Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, 43:551–560.
- Nash, J. (1990). *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation*. Hilger, New York. 2nd edition.
- Nikoloulopoulos, A. K., Joe, H., and Chaganty, N. R. (2011). Weighted scores method for regression models with dependent data. *Biostatistics*, 12:653–665.
- Nikoloulopoulos, A. K., Joe, H., and Li, H. (2012). Vine copulas with asymmetric tail dependence and applications to financial return data. *Computational Statistics & Data Analysis*, 56:659–673.
- Nikoloulopoulos, A. K. and Karlis, D. (2009). Finite normal mixture copulas for multivariate discrete data modeling. *Journal of Statistical Planning and Inference*, 139:3878–3890.
- Ochi, Y. and Prentice, R. L. (1984). Likelihood inference in a correlated probit regression model. *Biometrika*, 71:531–543.
- Sabo, R. and Chaganty, N. R. (2011). Letter to the editor of biometrics on “Joint regression analysis for discrete longitudinal data” by Madsen and Fang. *Biometrics*, 67:1669–1670.
- Schervish, M. (1984). Algorithm AS 195. multivariate normal probabilities with error bound. *Applied Statistics*, 33:81–94.
- Song, P. X.-K. (2007). *Correlated Data Analysis: Modeling, Analytics, and Application*. Springer, NY.
- Song, P. X.-K., Li, M., and Yuan, Y. (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics*, 65:60–68.
- Song, P. X.-K., Li, M., and Yuan, Y. (2011). Rejoinder to “Joint regression analysis for discrete longitudinal data” by Madsen and Fang. *Biometrics*, 67:1175–1176.
- Žežula, I. (2009). On multivariate Gaussian copulas. *Journal of Statistical Planning and Inference*, 139:3942–3946.