Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

Modeling Multivariate Count Data Using Copulas

Aristidis K. Nikoloulopoulos & Dimitris Karlis

Communications in Statistics—Simulation and Computation®, 39: 172–187, 2010

ISSN: 0361-0918 print/1532-4141 online
DOI: 10.1080/03610910903391262

Modeling Multivariate Count Data Using Copulas

ARISTIDIS K. NIKOLOULOPOULOS[1]
AND DIMITRIS KARLIS[2]

[1]School of Computing Sciences, University of East Anglia,
Norwich, UK
[2]Department of Statistics, Athens University of Economics and Business,
Athens, Greece

Multivariate count data occur in several different disciplines. However, existing models do not offer great flexibility for dependence modeling. Models based on copulas nowadays are widely used for continuous data dependence modeling. Modeling count data via copulas is still in its infancy; see the recent article of Genest and Nešlehová (2007). A series of different copula models providing various residual dependence structures are considered for vectors of count response variables whose marginal distributions depend on covariates through negative binomial regressions. A real data application related to the number of purchases of different products is provided.

Keywords Archimedean copulas; Kendall's tau; Market basket count data; Mixtures of max-id copulas; Partially symmetric copulas.

Mathematics Subject Classification 62H20; 62P10.

1. Introduction

Multivariate count data occur in several disciplines, such as epidemiology, marketing, criminology, industrial statistics, among others. In marketing, modeling the number of purchases of different products has been of special interest as it has various implications like predicting sales in the future, examining the behavior and the typology of buyers, creating marketing strategies, etc. In addition, working jointly with more products can be quite useful to derive new marketing strategies; see, e.g., Brijs et al. (2004).

Accordingly, if only one product is considered then we loose valuable information related to moving to different brands, using substitutes or finding related products that are purchased together. Therefore, working with more products, or with one product but in successive time periods, allows us to reveal the

Received December 19, 2008; Accepted September 30, 2009
Address correspondence to Dimitris Karlis, Department of Statistics, Athens University of Economics and Business, 76 Patission Street, Athens 10434, Greece; E-mail: karlis@aueb.gr

existing structure in the buying behavior and perhaps predict in a better way the expected income from each customer; see Hoogendoorn and Sickel (1999). However, flexible models for such data are not widely available and usually are hard to be fitted in real data.

Most of the existing models start from the multivariate Poisson model; see Johnson et al. (1997). The multivariate Poisson distributions allow only for positive correlation. Typical extensions are based on mixtures (see, e.g., Chib and Winkelmann (2001) and Karlis and Xekalaki (2005) and the references therein) to allow for flexible correlation structure and overdispersed marginal distributions. However, a certain limitation is that since the correlation structure comes from a multivariate mixing distribution, the possible choices are very limited and perhaps they lead to very specific marginal models. On the other hand, models based on other discrete distributions can be also constructed as for example the bivariate negative binomial model of Winkelmann (2000) or models based on conditional distributions (Berkhout and Plug, 2004). Such models suffer from the difficulty to generalize to other families of marginal distributions.

All of the above usually have marginal distributions of a specific kind and the dependence structure offered is limited. For this reason, we proceed by considering the use of copula based models. The literature on copulas used for count data is limited. We aim at contributing in this area by considering copula-based models for multivariate counts that allow for both flexible dependence structure and flexible marginal distributions. The specification in this way of the multivariate discrete distribution provides complete inference, i.e., maximum likelihood estimation and calculation of joint and conditional probabilities. The latter is not provided by other methods such as log-linear models, see, e.g., Wedel et al. (2003), and generalized estimating equations (GEE), see, e.g., Liang and Zeger (1986). The models derived in this article are based on known parametric families of copulas. However, to our knowledge this is the first time some of them were used for count data dependence modeling and their comparison reveals interesting implications on their use for real data.

By definition, an m-variate copula $C(u_1, \ldots, u_m)$ is a cumulative distribution function (cdf) with uniform marginals on the interval $(0, 1)$; see, e.g., Joe (1997) or Nelsen (2006). If $F_j(y_j)$ is the cdf of a univariate random variable Y_j, then $C(F_1(y_1), \ldots, F_m(y_m))$ is an m-variate distribution for $\mathbf{Y} = (Y_1, \ldots, Y_m)$ with marginal distributions F_j, $j = 1, \ldots, m$. Conversely, if H is an m-variate cdf with univariate marginal cdfs $F_1, \ldots, F_m$, then there exists an m-variate copula C such for all $\mathbf{y} = (y_1, \ldots, y_m)$,

$$H(y_1, \ldots, y_m) = C(F_1(y_1), \ldots, F_m(y_m)). \tag{1}$$

If $F_1, \ldots, F_m$ are continuous, then C is unique; otherwise, there are many possible copulas as emphasized by Genest and Nešlehová (2007), but all of these coincide on the closure of $\mathrm{Ran}(F_1) \times \cdots \times \mathrm{Ran}(F_m)$, where $\mathrm{Ran}(F)$ denotes the range of F. This result, known as Sklar's theorem, indicates the way that multivariate cdfs and their univariate cdfs can be connected. While the derivation of joint density is easy for the continuous case through partial derivatives, it is not so simple in the case of discrete data. In the latter case, the probability mass function (pmf) $h(\cdot)$ is obtained using finite differences as indicated in the following proposition:

Proposition 1.1. *Consider a discrete integer-valued random vector $(Y_1, \ldots, Y_m)$ with marginals $F_1, \ldots, F_m$ and joint cdf given by the copula representation $H(y_1, \ldots, y_m) = C(F_1(y_1), \ldots, F_m(y_m))$. Let $\mathbf{c} = (c_1, \ldots, c_m)$ be vertices where each c_k is equal to either y_k or $y_k - 1$, $k = 1, \ldots, m$. Then the joint pmf $h(\cdot)$ of the discrete random variables $Y_1, \ldots, Y_m$ is given by*

$$h(y_1, y_2, \ldots, y_m) = \sum \mathrm{sgn}(\mathbf{c}) C(F_1(c_1), \ldots, F_m(c_m)),$$

where the sum is taken over all vertices $\mathbf{c}$, *and* $sgn(\mathbf{c})$ *is given by,*

$$sgn(\mathbf{c}) = \begin{cases} 1, & \text{if } c_k = y_k - 1 \text{ for an even number of } k's. \\ -1, & \text{if } c_k = y_k - 1 \text{ for an odd number of } k's. \end{cases}$$

From the above it is evident that for calculating the joint probability function one needs to evaluate the copula repeatedly. Therefore, in practice, in order to be able to use copula models for multivariate count data, one needs to specify copulas with computationally feasible form of the cdf.

Multivariate elliptical (e.g., normal) copulas, see Fang et al. (2002) and Abdous et al. (2005), provide flexible structure (allowing both positive and negative dependence), but they do not have a closed form cdf. Therefore, computational problems appear for $m > 2$ in the derivation of pmf which involves computation of the copula in several different points and hence repeated multivariate numerical integration. Van Ophem (1999) and Lee (2001) exploited the use of bivariate normal copula to model count data, while Song (2000, 2007) defined multivariate dispersion models through multivariate normal copula. Computational problems for the multivariate case are not mentioned, as the author concentrate his demonstration on exchangeable dependence where multidimensional probabilities are one-dimensional integrals; see Joe (1995).

The remaining literature for copulas and discrete data is concentrated to copulas with a closed form cdf. There are few papers using Archimedean copulas with discrete data; see Meester and MacKay (1994), Lee (1999), Trégouët et al. (1999), Cameron et al. (2004), and McHale and Scarf (2007). Therein Frank copula used to model mainly bivariate discrete data (i.e., count and binary data), allowing both positive and negative residual dependence. For multivariate data, Frank copula allow only exchangeable structure with a narrower range of negative residual dependence as the dimension increases; see, e.g., Joe (1997).

Joe (1993) defined the partially symmetric copulas extending Archimedean to a class with a non exchangeable structure. Zimmer and Trivedi (2006) and Paiva and Kolev (2009) used a trivariate partially symmetric Frank copula to model discrete data.

The pioneering work of Joe and Hu (1996), defining multivariate parametric families of copulas that are mixtures of max-id bivariate copulas, has remained almost completely overseen for modeling multivariate count data. Recently, Nikoloulopoulos and Karlis (2008) predicted dependent binary outcomes using this class of copulas, which allows flexible dependence among the random variables and has a closed form cdf and thus computations are rather easy. Herein, we will present its superiority in contrast with the other existing classes of copulas and propose its use for multivariate count data modeling.

The remainder of this article proceeds as follows. Section 2 presents briefly the multivariate copula families with a closed form cdf, which will be used in this article in a self-contained manner. Section 3 describes how copula functions can be used to model dependence on count data. In Sec. 4, estimation procedures are presented, while in Sec. 5 a real data application, concerning market basket count data, is provided. In fact, the copula functions are used to describe the dependence of error terms in negative binomial regression models for marginals considered. Finally, in Sec. 6 concluding remarks can be found.

2. Multivariate Parametric Families of Copulas

2.1. *Multivariate Archimedean Copulas*

Let Λ be a univariate cdf of a positive random variable ($\Lambda(0) = 0$), and let ϕ be the Laplace transform (LT) of Λ,

$$\phi(t) = \int_0^{\infty} e^{-ts} d\Lambda(s), \quad t \geq 0.$$

For an arbitrary univariate cdf F, there exists a unique cdf G, such that

$$F(x) = \int_0^{\infty} G^s(x) d\Lambda(s) = \phi(-\log G(x)) \tag{2}$$

directing to $G(x) = \exp(-\phi^{-1}(F(x)))$, where ϕ^{-1} is the functional inverse of ϕ. Extending this result for bivariate case the following formula is a bivariate cdf:

$$\begin{aligned}\int_0^{\infty} G_1^s(y_1) G_2^s(y_2) d\Lambda(s) &= \phi(-\log G_1(y_1) - \log G_2(y_2)) \\ &= \phi(\phi^{-1}(F_1(y_1)) + \phi^{-1}(F_2(y_2))),\end{aligned}$$

where now $G_j(x) = \exp(-\phi^{-1}(F_j(x)))$, $j = 1, 2$. The bivariate copula

$$C(u_1, u_2) = \phi(\phi^{-1}(u_1) + \phi^{-1}(u_2)), \tag{3}$$

is the well-known Archimedean copula with generator the inverse function of the LT.

The multivariate Archimedean copula is a simple extension of (3) to the m-variate case

$$C(\mathbf{u}) = \phi\left(\sum_{j=1}^{m} \phi^{-1}(u_j)\right). \tag{4}$$

This multivariate copula is permutation-symmetric in the m arguments, thus it is a distribution for exchangeable $U(0, 1)$ random variables with Kendall's tau association matrix,

$$\begin{pmatrix} 1 & \tau_\phi & \cdots & \tau_\phi \\ \vdots & \vdots & \vdots & \vdots \\ \tau_\phi & \tau_\phi & \cdots & 1 \end{pmatrix}.$$

Table 1
Max-id bivariate copulas and LTs. There is a correspondence between the LT used and the copulas, i.e., LTA corresponds to a Gumbel copula and so on. Note that $\overline{u_i} = 1 - u_i$ and $\widetilde{u_i} = -\log u_i$, $i = 1, 2$

Family	$C'(u_1, u_2; \theta)$	LT	$\phi(t; \theta)$	$\theta \in$
Gumbel	$e^{-(\widetilde{u_1}^{\theta}+\widetilde{u_2}^{\theta})^{1/\theta}}$	LTA	$e^{-t^{1/\theta}}$	$[1, \infty)$
Mardia-Takahasi	$(u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}$	LTB	$(1+t)^{-1/\theta}$	$(0, \infty)$
Joe	$1 - (\overline{u_1}^{\theta} + \overline{u_2}^{\theta} - \overline{u_1}^{\theta}\overline{u_2}^{\theta})^{1/\theta}$	LTC	$1 - (1 - e^{-t})^{1/\theta}$	$[1, \infty)$
Frank	$-\frac{1}{\theta}\log\left\{1 + \frac{(e^{-\theta u_1}-1)(e^{-\theta u_2}-1)}{e^{-\theta}-1}\right\}$	LTD	$-\frac{\log(1-(1-e^{-\theta})e^{-t})}{\theta}$	$(0, \infty)$
Galambos	$u_1 u_2 e^{(\widetilde{u_1}^{-\theta}+\widetilde{u_2}^{-\theta})^{-1/\theta}}$			$[0, \infty)$

One can see in the latter matrix that there is a common LT for all bivariate marginals. Therefore, all pairs of variables have the same association, which is rather restrictive in practice. Finally, as LTs one can use the choices LTA to LTD in Table 1.

2.2. *Partially Symmetric Copulas*

Joe (1993) extended multivariate Archimedean copulas to a more flexible class of copulas using nested LTs, the so-called partially-symmetric m-variate copulas with $m-1$ dependence parameters. The multivariate form has a complex notation, so we present the trivariate and 4-variate extensions of (4) to help the exposition. The trivariate form is given by

$$C(\mathbf{u}) = \phi_1\big(\phi_1^{-1} \circ \phi_2\big(\phi_2^{-1}(u_1) + \phi_2^{-1}(u_2)\big) + \phi_1^{-1}(u_3)\big), \tag{5}$$

where ϕ_1, ϕ_2 are LTs and $\phi_1^{-1} \circ \phi_2 \in \text{Ł}_\infty^\star = \{\omega : [0, \infty) \longrightarrow [0, \infty) | \omega(0) = 0, \omega(\infty) = \infty, (-1)^{j-1}\omega^j \geq 0, j = 1, \ldots, \infty\}$. From the above formula is clear that (5) has (1,2) bivariate margin of the form (3) with LT ϕ_2, and (1,3), (2,3) bivariate margins of the form (3) with LT ϕ_1.

As the dimension increases, there are many possible LT nestings. For the four-variate case the two possible LT nestings are:

$$C(\mathbf{u}) = \phi_1\big(\phi_1^{-1} \circ \phi_2\big(\phi_2^{-1} \circ \phi_3\big(\phi_3^{-1}(u_1) + \phi_3^{-1}(u_2)\big) + \phi_2^{-1}(u_3)\big) + \phi_1^{-1}(u_4)\big) \tag{6}$$

$$C(\mathbf{u}) = \phi_1\big(\phi_1^{-1} \circ \phi_2\big(\phi_2^{-1}(u_1) + \phi_2^{-1}(u_2)\big) + \phi_1^{-1} \circ \phi_3\big(\phi_3^{-1}(u_3) + \phi_3^{-1}(u_4)\big)\big), \tag{7}$$

where ϕ_1, ϕ_2, ϕ_3 are LTs and $\phi_1^{-1} \circ \phi_2, \phi_1^{-1} \circ \phi_3 \in \text{Ł}_\infty^\star$ defined earlier. For the four-variate case of the forms (6) and (7) all the trivariate margins have form (5) and all the bivariate have form (3). The Kendall's tau association matrix for the copula given in (5) is

$$\begin{pmatrix} 1 & \tau_{\phi_2} & \tau_{\phi_1} \\ \tau_{\phi_2} & 1 & \tau_{\phi_1} \\ \tau_{\phi_1} & \tau_{\phi_1} & 1 \end{pmatrix},$$

where $\tau_{\phi_2} > \tau_{\phi_1}$.

In the same manner, the Kendall's tau association matrix for the copula of the form (6) is

$$\begin{pmatrix} 1 & \tau_{\phi_3} & \tau_{\phi_2} & \tau_{\phi_1} \\ \tau_{\phi_3} & 1 & \tau_{\phi_2} & \tau_{\phi_1} \\ \tau_{\phi_2} & \tau_{\phi_2} & 1 & \tau_{\phi_1} \\ \tau_{\phi_1} & \tau_{\phi_1} & \tau_{\phi_1} & 1 \end{pmatrix},$$

where $\tau_{\phi_1} < \tau_{\phi_2} < \tau_{\phi_3}$, and the Kendall's tau association matrix of the copula given in (7) is

$$\begin{pmatrix} 1 & \tau_{\phi_2} & \tau_{\phi_1} & \tau_{\phi_1} \\ \tau_{\phi_2} & 1 & \tau_{\phi_1} & \tau_{\phi_1} \\ \tau_{\phi_1} & \tau_{\phi_1} & 1 & \tau_{\phi_3} \\ \tau_{\phi_1} & \tau_{\phi_1} & \tau_{\phi_3} & 1 \end{pmatrix},$$

where $\tau_{\phi_1} < \tau_{\phi_2}, \tau_{\phi_3}$.

From the above one can realize that bivariate copulas associated with LTs that are more nested, are larger in concordance than those that are less nested. For example, for (7) the (1,2) and (3,4) bivariate margins are more dependent (concordant) than the remaining four bivariate margins.

To make these results applicable, some choices of LTs, in which the property $\phi_1^{-1} \circ \phi_2 \in Ł_\infty^\star$, is satisfied are LTA to LTD (see Table 1). Note in passing that there are still restrictions on the dependence structure allowed by such copulas and the Archimedean copula is a subcase of partially symmetric copula when all LTs are from the same family.

2.3. *Copulas via Mixtures of max-id Bivariate Copulas*

Joe and Hu (1996) considered the mixture of univariate cdfs H_j and max-id bivariate copulas C'_{jk} of the form

$$\begin{aligned} &\int_0^\infty \prod_{1\le j<k\le m} C'^{s}_{jk}(H_j, H_k) \prod_{j=1}^m H_j^{\nu_j s}\, d\Lambda(s) \\ &\quad = \phi\Bigg(-\sum_{1\le j<k\le m} \log C'_{jk}(H_j, H_k) + \sum_{j=1}^m \nu_j \log H_j\Bigg). \end{aligned} \tag{8}$$

The above representation defines a multivariate copula if H_j are chosen appropriately. The univariate margins of (8) are $F_j = \phi(-(\nu_j + m - 1)\log H_j)$, so substituting $H_j(u_j) = e^{-p_j\phi^{-1}(u_j)}$ and $p_j = (\nu_j + m - 1)^{-1}$, $j = 1, \ldots, m$ in (8) is a multivariate copula distribution with a closed form cdf,

$$C(\mathbf{u}) = \phi\Bigg(-\sum_{j<k} \log C'_{jk}(e^{-p_j\phi^{-1}(u_j)}, e^{-p_k\phi^{-1}(u_k)}) + \sum_{j=1}^m \nu_j p_j \phi^{-1}(u_j)\Bigg). \tag{9}$$

It is well known that the mixing operation introduces dependence, so this new copula has a dependence structure that comes from the form of C'_{jk} and the form of the mixing distribution $\Lambda(\cdot)$ which is characterized by its LT $\phi(\cdot)$.

Another interesting interpretation is that the LT ϕ introduces the minimal dependence between the random variables, while the copulas C'_{jk} provide some additional pairwise dependence beyond the minimal dependence. The parameters v_j are included in order that the parametric family of multivariate copulas (9) is closed under margins. Taking $H_r \longrightarrow 1, \forall r \in \{1, \ldots, m\}\backslash\{j, k\}$ in (8) the (j, k) bivariate marginal copula of (9) is,

$$C_{jk}(u_j, u_k) = \phi\big(-\log C'_{jk}(e^{-p_j\phi^{-1}(u_j)}, e^{-p_k\phi^{-1}(u_k)}) \\ + (v_j + m - 2)p_i\phi^{-1}(u_i) + (v_k + m - 2)p_k\phi^{-1}(u_k)\big). \tag{10}$$

Keeping v_j as known parameters, the copula of the form (9) is a family with $\frac{m(m-1)}{2} + 1$ parameters with flexible dependence. One may simplify the form of the copula by assuming $C'_{rs}(u_r, u_s) = u_r u_s$ (product copula, independence) together with $v_r = v_s = -1$, for some pairs. This implies that for that pairs of variables we assume the minimum level of dependence as introduced by ϕ. This allows to simplify the model and reduce the number of parameters to be estimated. Note, in passing, that the model in its full form allows for different association for each pair of variables.

Some of the choices of bivariate max-id copulas are Galambos, Gumbel, Frank, Joe, and Mardia-Takahasi (also known as Clayton or Kimeldorf-Sampson copula). These, together with some LT, can be seen in Table 1 (LTA to LTD). Their combination results in a variety of parametric families of the form (9) with flexible positive dependence structure. Of course this provides a rich pool of candidate models and deserves some model selection technique. Note that we can construct multivariate copulas that are mixtures of common $C'_{jk}(\cdot) = C'(\cdot; \theta_{jk})$ and not common max-id copulas to provide the most flexible dependence according to data on hand.

2.4. *Negative Dependence*

As we have already mentioned, multivariate Archimedean copulas provide a narrower range of dependence as the dimension increases; see also McNeil and Nešlehová (2009) for a thorough treatment. Furthermore, partially symmetric and mixtures of max-id copulas provide only positive dependence by definition; see Joe (1997). What about if the data on hand have negative dependence?

Negative dependence can be introduced by applying decreasing transformations to the oppositely ordered variables. If $(U_1, \ldots, U_m) \sim C$ where C is a copula with positive dependence, one could always get some negative dependence for a subset of variables, by supposing C^* is the copula of $(U_1, \ldots, U_k, 1 - U_{k+1}, \ldots, 1 - U_m)$.

3. Copulas and Dependence for Count Data

The dependence between random variables is completely described by their joint distribution, which can be represented by (1). For continuous random variables, dependence as measured by Kendall's tau (τ) is associated only with the copula parameters; see, e.g., Nelsen (2006). This is, however, not the case for discrete data because the probability of a tie is positive; see Denuit and Lambert (2005) and Mesfioui and Tajar (2005). For this reason the marginal distributions play also some role on dependence, and τ does not attain ± 1 values. Here, we provide a formula for Kendall's tau; see Nikoloulopoulos (2007) for the derivation. For normalized versions one can refer to Goodman and Kruskal (1954) or Nešlehová (2007).

Lemma 3.1. *Let $Y_i, i = 1, 2$ be integer-valued discrete random variables whose joint distribution is H, with marginal cdfs F_i, pmfs $f_i, i = 1, 2$ and copula C. Then the population version of Kendall's tau for Y_1 and Y_2 is given by*

$$\tau(Y_1, Y_2) = \sum_{y_1=0}^{\infty} \sum_{y_2=0}^{\infty} h(y_1, y_2)\{4C(F_1(y_1 - 1), F_2(y_2 - 1)) - h(y_1, y_2)\} + \sum_{y_1=0}^{\infty} \left(f_1^2(y_1) + f_2^2(y_1)\right) - 1, \tag{11}$$

where

$$h(y_1, y_2) = C(F_1(y_1), F_2(y_2)) - C(F_1(y_1 - 1), F_2(y_2)) - C(F_1(y_1), F_2(y_2 - 1)) + C(F_1(y_1 - 1), F_2(y_2 - 1))$$

is the joint pmf of Y_1 and Y_2.

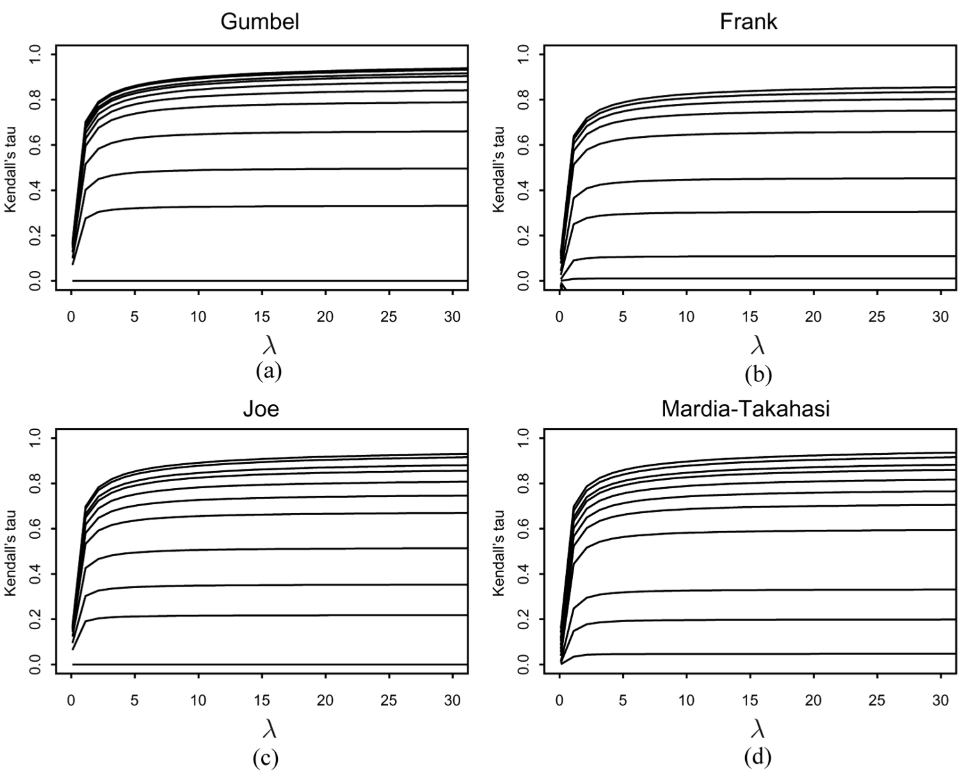

Figure 1. Kendall'tau values computed using Archimedean copulas for a grid of parameter value for each copula (different lines) and Poisson marginal distributions ($P(\lambda)$) with the same parameter λ up to 30, higher curves corresponding to higher values of the copula parameter.

This representation of Kendall's tau is equivalent to the one derived in Denuit and Lambert (2005). In fact, it provides us with better insight, since the marginal probability functions f_i, $i = 1, 2$ are clearly involved in the formulas and make clear the dependence of Kendall's tau on the marginal distributions.

In Fig. 1, Kendall's tau values have been plotted for Archimedean copulas (for each copula the lines correspond to different values of its parameter). We have used Poisson marginal distributions with the same parameter for each marginal. The plot depicts Kendall'tau values against this common Poisson parameter. Higher curves corresponding to higher values of the copula parameter.

From the plot we can see that for marginal parameters (denoted by λ) greater than 10 their association with the value of Kendall's tau is negligible. Moreover, as λ tends to infinity, the upper bound of Kendall's tau is 1, due to absence of ties.

4. Estimation of a Multivariate Copula Based Model

Consider a multivariate copula based parametric model for the random vector $\mathbf{y}$ with m elements and distribution function H provided by the copula representation

$$H(\mathbf{y}; \alpha_1, \ldots, \alpha_m, \theta) = C(F_1(y_1; \alpha_1), \ldots, F_m(y_m; \alpha_m); \theta), \tag{12}$$

where F_i are the marginal cdfs, with parameter vectors α_i, $i = 1, \ldots, m$ and θ is the vector of copula parameter. The pmf $h(\mathbf{y}; \alpha_1, \ldots, \alpha_m, \theta)$ of the specified cdf H in (12) can be obtained using Proposition 1.1.

Consider the m log-likelihoods functions for the univariate marginal distributions:

$$L_{y_j}(\alpha_j) = \sum_{i=1}^{n} \log f_j(y_{ij}; \alpha_j), \quad j = 1, \ldots, m \tag{13}$$

and the joint log-likelihood

$$L(\theta, \alpha_1, \ldots, \alpha_m) = \sum_{i=1}^{n} \log h(y_{i1}, \ldots, y_{im}; \alpha_1, \ldots, \alpha_m, \theta), \tag{14}$$

where f_i, $i = 1, \ldots, m$ are the marginal pmfs and n is the sample size.

Efficient estimation of the model parameters is succeeded by the inference function of margins (IFM), which consists of a two-step approach. At the first step of this method the univariate log-likelihoods (13) are maximized independently of the copula parameter and at the second step the joint log-likelihood (14) maximized over θ with univariate parameters fixed as estimated at the first step of the method. Estimation by IFM method becomes more popular as the dimension increases and computational problems arise. The problem of fitting multivariate data is decomposed into two smaller problems: fitting the marginal distributions separately from fitting the existing dependence structure. Asymptotic efficiency of the IFM has been studied by Joe (2005) for a number of multivariate models. All of these comparisons suggest that the IFM method is highly efficient compared to standard maximum likelihood, except for extreme cases near the Fréchet bounds.

5. Application

5.1. *The Data*

Transactional market basket data provide excellent opportunities for a retailer to segment the customer population into different groups based on differences in their purchasing behavior. The data refer to the frequency of purchases of products or product categories within the retail store and, as a result, they are extremely useful for modeling consumer purchase behavior. Moreover, they reflect the dependencies that exist between purchases in different product categories.

We used the scanner data in Brijs et al. (1999). The data refer to the number of food, non food, hygiene, and fresh purchases from loyalty card holders of a large super market for a given month time period in Belgium. Hygiene category contains articles like hair, gel, shaving foam, bath foam, toilet soap, etc., while fresh category contains vegetables, fruit, meat, cheese, and bakery items. It is a special category because it contains all the food items that are not prepacked but are served by personnel behind a counter. In Figs. 2 and 3 one can see the large tails and the present dependence on the data, respectively.

The use of copulas allows to specify the marginal distributions in a more flexible way as we do not need to specify the entire model at once and hence the marginal distributions can be selected separately. This can help considerably on selecting improved models even with different marginal distributions.

Therefore, before choosing the appropriate copula family to capture the dependence between the residuals of the marginal model we specify the univariate

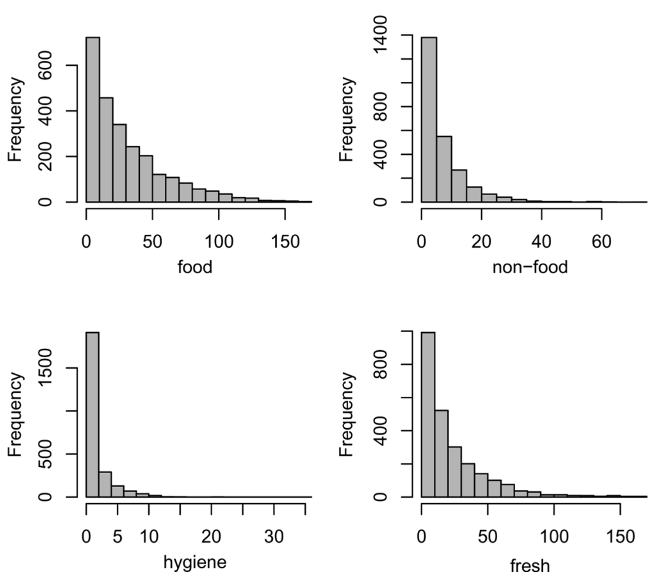

Figure 2. Histograms of purchases.

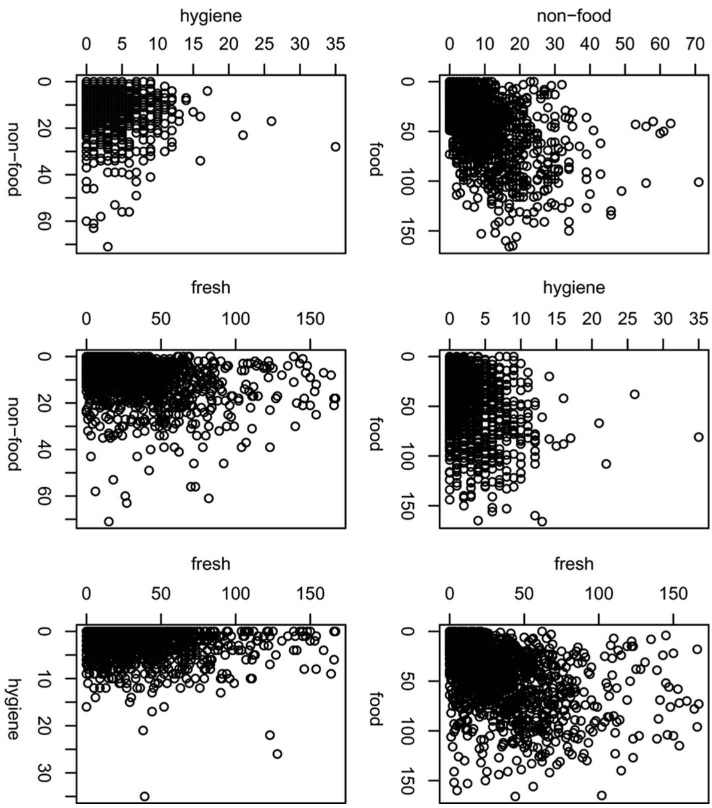

Figure 3. Scatter plots of purchases.

Table 2
Univariate estimates and standard errors (SE) for the negative binomial regression models for each response variable

	Food		Non-food		Hygiene		Fresh	
Covariate	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
(intercept)	2.99	0.08	1.43	0.09	−0.23	0.14	2.78	0.09
pet	0.05	0.04	0.11	0.05	0.09	0.07	−0.02	0.05
car	−0.09	0.08	−0.07	0.09	−0.16	0.15	−0.04	0.09
club	0.00	0.04	0.06	0.04	0.08	0.07	−0.09	0.05
freezer	−0.04	0.08	−0.04	0.09	0.16	0.14	−0.12	0.09
microwave	−0.07	0.05	0.00	0.06	0.01	0.09	0.00	0.06
garden	0.26	0.07	0.17	0.08	0.09	0.13	0.26	0.08
age < 18	0.09	0.02	0.12	0.03	0.11	0.04	0.07	0.03
$18 \leq$ age < 45	0.10	0.02	0.09	0.02	0.18	0.04	0.10	0.03
$45 \leq$ age < 65	0.09	0.03	0.12	0.03	0.08	0.05	0.13	0.03
age ≥ 65	0.03	0.04	0.08	0.05	−0.15	0.07	0.08	0.05
σ	1.06	0.03	0.97	0.03	0.47	0.02	0.82	0.02

marginal distributions. For our data, the negative binomial model (see, e.g., Lawless, 1987) is considered, allowing for the large over-dispersion found in the data. For each observation $i = 1, \ldots, 2472$, each marginal is specified conditional on covariates $\mathbf{X}_i$ and cumulative probability function given by

$$F_j(y_{ij} \,|\, \mathbf{X}_i, \beta_j) = \sum_{k=0}^{y_{ij}} \frac{\Gamma(\sigma_j + k)}{\Gamma(\sigma_j)\Gamma(k+1)} \frac{\mu_{ij}^k \sigma_j^{\sigma_j}}{(\mu_{ij} + \sigma_j)^{\sigma_j + k}}, \quad i = 1, \ldots, 2472 \quad j = 1, 2, 3, 4, \tag{15}$$

where $E(y_{ij}) = \mu_{ij} = \exp(\mathbf{X}_i\beta_j)$ and $\mathrm{var}(y_{ij}) = \mu_{ij} + \mu_{ij}^2/\sigma_j$. A similar approach is used by Cameron et al. (2004) and Paiva and Kolev (2009). The covariate information refers to if the customers use their car to go for shopping in the supermarket, if they have pets, if they have freezer, if they have microwave, and if they have garden in their home. Moreover, the number of members of the family belonging to four different age subcategories: (a) 0–18 years, (b) 18–45 years, (c) 45–65 years, and (d) more than 65 years, was recorded to account for the household composition.

The univariate regression parameters, which estimated at the first step of the IFM method, fitting separate negative binomial regression models for each response variable, are shown in Table 2. As a preliminary data analysis we calculated the sample Kendall's tau values for the residuals of the fitted marginal models; see Table 3. Greater dependence existed between food and non food data and between food and fresh data, while the lowest between hygiene and fresh data. For the remaining marginals the strength of residual dependence was quite the average of lowest and greater dependence and at quite the similar strength between purchases.

5.2. *Four-Variate Fitted Models*

We started our modeling by considering the simplest structure provided by multivariate Archimedean copulas. This class assumes the same residual dependence for each pair. Then we fitted partially symmetric copulas, see Sec. 2.2, and particularly the LT nesting given by (6) with some reordering of the u_j's to capture the dependence in our data, i.e.,

$$C(\mathbf{u}) = \phi_1\big(\phi_1^{-1} \circ \phi_2\big(\phi_2^{-1} \circ \phi_3\big(\phi_3^{-1}(u_1) + \phi_3^{-1}(u_4)\big) + \phi_2^{-1}(u_2)\big) + \phi_1^{-1}(u_3)\big).$$

Table 3
Sample Kendall's tau values for the residuals of the fitted marginal models

	Food	Non-food	Hygiene	Fresh
Food	1.00	0.42	0.27	0.43
Non-food	0.42	1.00	0.23	0.32
Hygiene	0.27	0.23	1.00	0.18
Fresh	0.43	0.32	0.18	1.00

The Kendall's tau correlation matrix for the copula of this form is

$$\begin{pmatrix} 1 & \tau_{\phi_2} & \tau_{\phi_1} & \tau_{\phi_3} \\ \tau_{\phi_2} & 1 & \tau_{\phi_1} & \tau_{\phi_2} \\ \tau_{\phi_1} & \tau_{\phi_1} & 1 & \tau_{\phi_1} \\ \tau_{\phi_3} & \tau_{\phi_2} & \tau_{\phi_1} & 1 \end{pmatrix},$$

where $\tau_{\phi_1} < \tau_{\phi_2} < \tau_{\phi_3}$.

The next step was to fit copulas that are mixtures of max-id copulas of the form (9) with seven dependence parameters, one for the LT (θ) and one (θ_{jk}) for each marginal (j, k) keeping v's fixed and zero. Note here that one parameter is redundant. Finally, based on preliminary data analysis (Table 3), we simplified the models and numerical computations setting $C'_{34} = \Pi$ (independence copula) and $v_3 = v_4 = -1$. In this manner, we assumed a lower level of dependence for the (3,4) bivariate margin represented by the parameter θ of the LT ϕ and a higher dependence for the other bivariate margins with the parameters θ_{jk} representing bivariate dependence exceeding the minimum dependence of the LT ϕ.

5.3. *Results*

We fitted all copula-based models considered in the previous section using negative binomial regression models for the margins. For each class of copulas, several families are considered by choosing different LTs and/or max-id copulas. Table 4 provides the best model for each class, summarizing our findings.

To compare the models we report the AIC, which was calculated as the maximized log-likelihood minus the number of model parameters, to account for the different number of parameters that each type of models has. In addition, we report the average Kendall's tau values for each pair of variables to account for the strength of the residual dependence imposed among purchases.

There are some interesting findings from Table 4. Starting from the simpler model, the multivariate Archimedean copulas, which offer only exchangeable residual dependence, the results contradict with the data and thus they provide the worst AIC. Remember here that for count data with small mean values the

Table 4
Results from the best models from each parametric family of copulas

		AIC	$\bar{\tau}_{12}$	$\bar{\tau}_{13}$	$\bar{\tau}_{14}$	$\bar{\tau}_{23}$	$\bar{\tau}_{24}$	$\bar{\tau}_{34}$
four-parameter mixture of max-id copulas								
LTD	Gumbel	−31360.713	0.36	0.30	0.39	0.24	0.29	0.25
five-parameter mixture of max-id copulas								
LTD	Gumbel	−31338.59	0.37	0.30	0.39	0.24	0.28	0.24
six-parameter mixture of max-id copulas								
LTD	Gumbel	−31332.74	0.36	0.30	0.39	0.25	0.28	0.24
partially symmetric								
LTD		−31387.09	0.35	0.22	0.42	0.22	0.35	0.22
Archimedean								
LTD		−31483.04	0.32	0.27	0.32	0.27	0.32	0.27

Kendall's tau is associated both with marginal and copula parameters; see Fig. 1. To this end, the reason for the lower residual dependence on marginals (1,3), (2,3), and (3,4) is due to the fact that hygiene response has small mean value. The next most complicated model, the partially symmetric copulas, with three copula parameters, provide more flexibility and therefore better fit than Archimedean copulas.

Finally, the six-parameter mixture of max-id copulas are better from the partially symmetric copulas, because they provide flexible residual dependence, meaning that the number of bivariate marginals is equal to the number of dependence parameters. Note that more parsimonious models also fitted by removing parameters with estimated values near the boundary of the parameter space, and assuming product copulas for these pairs of variables. While such models involve less parameters we found that the AIC value was worst than the six-parameter mixture of max-id copulas, as one can read in Table 4.

5.4. *Managerial Implications*

We briefly mention in this subsection the interesting findings from the managerial point of view. First of all probabilities of any kind can be easily obtained based on Proposition 1.1. Thus, one can easily check for non buyers, i.e., persons that will not buy any of the products. Secondly, for a new customer one may calculate the expectation of purchases for each product category and hence deriving the expected amount to be spent by this new customer. Such numbers can be quite helpful for decision making. In addition, the cross products correlations help to examine effective marketing strategies by putting, for example, together products with large correlation in order to promote them. Note also that conditional probabilities and expectations can be deduced by simple calculus and thus predictions for joint number of purchases can be made. The models presented are quite flexible and relatively easy to apply with real data in order to facilitate such decisions. We do not pursue further this issue on the present article.

6. Concluding Remarks

Modeling multivariate count data based on copulas was described in the present article. We presented and fitted a series of different multivariate copulas of varying dependence structure indicating the importance of mixtures of max-id copulas in terms of flexibility. We showed how to account for residual dependence among count responses, given the explanatory variables. In our illustration the data were positively associated, but generally real multivariate discrete data exist with some negative associations, see, e.g., the cases in Aitchinson and Ho (1989) and Chib and Winkelmann (2001). From the latter class, as we have mentioned one could always get some negative dependence by applying decreasing transformations on some subset of the random variables but this is restrictive in general, because this construction cannot model negative dependence among many random variables.

The multivariate elliptical copulas inherit the dependence structure of elliptical distributions, i.e., allow for negative dependence, but lack a closed form cdf; this means likelihood inference might be difficult as multidimensional integration is required for the multivariate probabilities.

Ongoing research is focused on defining a new multivariate parametric family of copulas with computationally feasible form of the cdf and a wide range of dependence, see, e.g., Nikoloulopoulos and Karlis (2009).

Acknowledgments

This work is part of first author's Ph.D. thesis under the supervision of the second author at the Athens University of Economics and Business. The first author would like to thank the National Scholarship Foundation of Greece for financial support. The authors want to thank a referee for his constructive comments.

References

Abdous, B., Genest, C., Rémillard, B. (2005). Dependence properties of meta-elliptical distributions. In: Duchesne, P., Rémillard, B., eds. *Statistical Modelling and Analysis for Complex Data Problems*. Dordrecht: Kluwer, pp. 1–15.

Aitchinson, J., Ho, C. (1989). The multivariate Poisson-log normal distribution. *Biometrika* 75:621–629.

Berkhout, P., Plug, E. (2004). A bivariate Poisson count data model using conditional probabilities. *Statistica Neerlandica* 58:349–364.

Brijs, T., Karlis, D., Swinnen, G., Vanhoof, K., Wets, G., Manchanda, P. (2004). A multivariate Poisson mixture model for marketing applications. *Statistica Neerlandica* 58(3):322–348.

Brijs, T., Swinnen, G., Vanhoof, K., Wets, G. (1999). Using association rules for product assortment decisions: A case study. In: *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*. San Diego, CA, pp. 254–260.

Cameron, A. C., Li, T., Trivedi, P. K., Zimmer, D. M. (2004). Modelling the differences in counted outcomes using bivariate copula models with application to mismeasured counts. *The Econometrics Journal* 7(2):566–584.

Cameron, A. C., Trivedi, P. K. (1998). *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.

Chib, S., Winkelmann, R. (2001). Markov chain Monte Carlo analysis of correlated count data. *Journal of Business & Economic Statistics* 19(4):428–435.

Denuit, M., Lambert, P. (2005). Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis* 93(1):40–57.

Fang, H.-B., Fang, K.-T., Kotz, S. (2002). The meta-elliptical distributions with given marginals. *Journal of Multivariate Analysis* 82:1–16.

Genest, C., Nešlehová, J. (2007). A primer on copulas for count data. *The Astin Bulletin* 37:475–515.

Goodman, L., Kruskal, W. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association* 49:732–764.

Hoogendoorn, A. W., Sickel, D. (1999). Description of purchase incidence by multivariate heterogeneous Poisson processes. *Statistica Neerlandica* 53(1):21–35.

Joe, H. (1993). Parametric families of multivariate distributions with given margins. *Journal of Multivariate Analysis* 46:262–282.

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*. London: Chapman & Hall.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis* 94(2):401–419.

Joe, H., Hu, T. (1996). Multivariate distributions from mixtures of max-infinitely divisible distributions. *Journal of Multivariate Analysis* 57(2):240–265.

Johnson, N., Kotz, S., Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. New York: Wiley.

Karlis, D., Xekalaki, E. (2005). Mixed Poisson distributions. *International Statistical Review* 73:35–58.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics* 15(3):209–225.

Lee, A. (1999). Modelling rugby league data via bivariate negative binomial regression. *Australian and New Zealand Journal of Statistics* 41(2):141–152.

Lee, L.-F. (2001). On the range of correlation coefficients of bivariate ordered discrete random variables. *Econometric Theory* 17(1):247–256.

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

McHale, I., Scarf, P. (2007). Modelling soccer matches using bivariate discrete distributions with general dependence structure. *Statistica Neerlandica* 61(4):432–445.

McNeil, A. J., Nešlehová, J. (2009). Multivariate Archimedean copulas, d-monotone functions and l_1-norm symmetric distributions. *Annals of Statistics* 37:3059–3097.

Meester, S., MacKay, J. (1994). A parametric model for cluster correlated categorical data. *Biometrics* 50:954–963.

Mesfioui, M., Tajar, A. (2005). On the properties of some nonparametric concordance measures in the discrete case. *Journal of Nonparametric Statistics* 17(5):541–554.

Nelsen, R. B. (2006). *An Introduction to Copulas*. New York: Springer-Verlag.

Nešlehová, J. (2007). On rank correlation measures for non-continuous random variables. *Journal of Multivariate Analysis* 98:544–567.

Nikoloulopoulos, A. K. (2007). Application of Copula Functions in Statistics. Ph.D. thesis, Department of Statistics, Athens University of Economics.

Nikoloulopoulos, A. K., Karlis, D. (2008). Multivariate logit copula model with an application to dental data. *Statistics in Medicine* 27:6393–6406.

Nikoloulopoulos, A. K., Karlis, D. (2009). Finite normal mixture copulas for multivariate discrete data modeling. *Journal of Statistical Planning and Inference* 139:3878–3890.

Paiva, D., Kolev, N. (2009). Copula-based regression models: a survey. *Journal of Statistical Planning and Inference* 139(11):3847–3856.

Song, P. X.-K. (2000). Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics* 27(2):305–320.

Song, P. X.-K. (2007). *Correlated Data Analysis: Modeling, Analytics, and Application*. New York: Springer.

Trégouët, D. A., Ducimetière, P., Bocquet, V., Visvikis, S., Soubrier, F., Tiret, L. (1999). A parametric copula model for analysis of familial binary data. *American Journal of Human Genetics* 64(3):886–893.

Van Ophem, H. (1999). A general method to estimate correlated discrete random variables. *Econometric Theory* 15(2):228–237.

Wedel, M., Böckenholt, U., Kamakura, W. (2003). Factor models for multivariate count data. *Journal of Multivariate Analysis* 87:356–369.

Winkelmann, R. (2000). Seemingly unrelated negative binomial regression. *Oxford Bulletin of Economics and Statistics* 62(4):553–560.

Zimmer, D., Trivedi, P. (2006). Using trivariate copulas to model sample selection and treatment effects: application to family health care demand. *Journal of Business & Economic Statistics* 24(1):63–76.