Modelling multivariate count data using copulas
Dimitris Karlis, Aristidis K. Nikoloulopoulos

To cite this version:

HAL Id: hal-00537682
https://hal.archives-ouvertes.fr/hal-00537682
Submitted on 19 Nov 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Modelling multivariate count data using copulas

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Communications in Statistics - Simulation and Computation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID:</td>
<td>LSSP-2008-0315.R2</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Original Paper</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>22-Sep-2009</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>Karlis, Dimitris; Athens University of Economics, Statistics Nikoloulopoulos, Aristidis; University of East Anglia, Computing Sciences</td>
</tr>
<tr>
<td>Keywords:</td>
<td>copulas, discrete distributions, super market data, archimedean copulas</td>
</tr>
<tr>
<td>Abstract:</td>
<td>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 paper of Genest and Neshlehova (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.</td>
</tr>
</tbody>
</table>

Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.

LSSP-2008-0315.R2 Nikoloulopoulos and KArlis.zip
Modeling multivariate count data using copulas

Aristidis K. Nikoloulopoulos ∗  Dimitris Karlis †

Abstract
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 paper 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: Kendall’s tau; Archimedean copulas; partially symmetric copulas; mixtures of max-id copulas; market basket count data.

1 Introduction
Multivariate count data occur in several disciplines, like 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, e.t.c. In addition, working jointly with more products can be quite useful to derive new marketing strategies; see for 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

∗A.Nikoloulopoulos@uea.ac.uk, School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK
†karlis@aueb.gr, Department of Statistics, Athens University of Economics and Business, 76 Patission street, 10434 Athens, Greece.

URL: http://mc.manuscriptcentral.com/lssp E-mail: comstat@univmail.cis.mcmaster.ca
that are purchased together. Therefore, working with more products or with one
product but in successive time periods, allow us to reveal the 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 allows only for pos-
itive correlation. Typical extensions are based on mixtures (see for 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 the above usually have marginal distribution of a specific kind and the de-
pendence 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 for e.g., Wedel et al. (2003), and generalized
estimating equations (GEE), see for e.g., Liang and Zeger (1986). The models de-
derived in the paper are based on known parametric families of copulas. However, to
our knowledge this is the first time of using some of them 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 for 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 $Y = (Y_1, \ldots, Y_m)$ with marginal
distributions $F_j$, $j = 1, \ldots, m$. Conversely, if $H$ is an $m$-variate cdf with univari-
ate marginal cdfs $F_1, \ldots, F_m$, then there exists an $m$-variate copula $C$ such for all $y = (y_1, \ldots, y_m)$,

$$H(y_1, \ldots, y_m) = C(F_1(y_1), \ldots, F_m(y_m)).$$

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 $\text{Ran}(F_1) \times \cdots \times \text{Ran}(F_m)$, where $\text{Ran}(F)$ 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 $c = (c_1, \ldots, c_m)$ 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

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

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

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

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 (for 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) exploit 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 1-dimensional integrals; 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; 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 for 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) use 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) predict 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 remaining of the paper proceeds as follows: Section 2 presents briefly the multivariate copula families with a closed form cdf, which will be used in this paper in a self-contained manner. Section 3 describes how copula functions can be used to model dependence on count data. In Section 4 estimation procedures are presented, while in Section 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 Section 6 concluding remarks can be found.
2 Multivariate parametric families of copulas

2.1 Multivariate Archimedean copulas

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

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

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

\[ F(x) = \int_0^\infty G(x) d\Lambda(s) = \phi(-\log G(x)) \]  

(2)

directing to \( G(x) = \exp \left( -\phi^{-1}(F(x)) \right) \), where \( \phi^{-1} \) is the functional inverse of \( \phi \).

Extending this result for bivariate case the following formula is a bivariate cdf,

\[ \int_0^\infty G_1^*(y_1)G_2^*(y_2)d\Lambda(s) = \phi(-\log G_1(y_1)-\log G_2(y_2)) = \phi \left( \phi^{-1}(F_1(y_1)) + \phi^{-1}(F_2(y_2)) \right) \]

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

\[ C(u_1, u_2) = \phi(\phi^{-1}(u_1) + \phi^{-1}(u_2)), \]  

(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,

\[ C(u) = \phi \left( \sum_{j=1}^m \phi^{-1}(u_j) \right). \]  

(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,

\[
\begin{pmatrix}
1 & \tau_\phi & \cdots & \tau_\phi \\
\vdots & \vdots & \ddots & \vdots \\
\tau_\phi & \tau_\phi & \cdots & 1
\end{pmatrix}
\]

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,

\[
C(u) = \phi_1 \left( \phi_1^{-1} \circ \phi_2 \left( \phi_2^{-1}(u_1) + \phi_3^{-1}(u_2) \right) + \phi_1^{-1}(u_3) \right) \tag{5}
\]

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

As the dimension increases there are many possible LT nestings. For the 4-variate case the two possible LT nestings are,

\[
C(u) = \phi_1 \left( \phi_1^{-1} \circ \phi_2 \left( \phi_2^{-1}(u_1) + \phi_3^{-1}(u_2) \right) + \phi_1^{-1}(u_3) \right) \tag{6}
\]

\[
C(u) = \phi_1 \left( \phi_1^{-1} \circ \phi_2 \left( \phi_2^{-1}(u_1) + \phi_3^{-1}(u_2) \right) + \phi_1^{-1}(u_3) \right) \tag{7}
\]

where \(\phi_1, \phi_2, \phi_3\) are LTs and \(\phi_1^{-1} \circ \phi_2, \phi_1^{-1} \circ \phi_3 \in L^*_{\infty}\) defined earlier. For the 4-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 $\phi_1^{-1} \circ \phi_2 \in L_\infty$, 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.

Table 1 about here

### 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

$$
\int_0^\infty \prod_{1 \leq j < k \leq m} C'_{jk}(H_j, H_k) \prod_{j=1}^m H_j^{\nu_j} d\Lambda(s) = \phi \left( - \sum_{1 \leq j < k \leq m} \log C'_{jk}(H_j, H_k) + \sum_{j=1}^m \nu_j \log H_j \right). \tag{8}
$$

The above representation defines a multivariate copula if $H_j$ are chosen appropriately. The univariate margins of (8) are $F_j = \phi \left( -(\nu_j + m - 1) \log H_j \right)$, 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,

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

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

$$C_{jk}(u_j, u_k) = \phi \left( -\log C'_{jk}(e^{-p_j\phi^{-1}(u_j)}, e^{-p_k\phi^{-1}(u_k)}) + (\nu_j + m - 2)p_j \phi^{-1}(u_j) + (\nu_k + m - 2)p_k \phi^{-1}(u_k) \right).$$

(10)

Keeping $\nu_j$ 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 $C'_{rs}(u_r, u_s) = u_ru_s$ (product copula, independence) together with $\nu_r = \nu_s = -1$, for some pairs. This implies that for that pairs of variables we assume the minimum level of dependence as introduced by $\phi$. 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 mention 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 ($\tau$) is associated only with the copula parameters; see for 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), Mesfioui and Tajar (2005). For this reason the marginal distributions play also some role on dependence, and $\tau$ 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

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

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 $Y_2$.

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 Figure 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 $\lambda$) greater than 10 their association with the value of Kendall’s tau is negligible. Moreover as $\lambda$ 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

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

(12)

where \( F_i \) are the marginal cdfs, with parameter vectors \( \alpha_i, i = 1, \ldots, m \) and \( \theta \) 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:

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

(13)

and the joint log-likelihood

\[
L(\theta, \alpha_1, \ldots, \alpha_m) = \sum_{i=1}^{n} \log h(y_{1i}, \ldots, y_{mi}; \alpha_1, \ldots, \alpha_m, \theta),
\]

(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 \( \theta \) 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 FML, 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 supermarket for a given month time period in Belgium. Hygiene category contains articles like hair, gel, shaving foam, bath foam, toilet soap e.t.c., 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 Figure 2 and in Figure 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 marginal distributions. For our data, the negative binomial model (see for 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 $X_i$ and cumulative probability function given by

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

(15)

where $E(y_{ij}) = \mu_{ij} = \exp(X_i \beta_j)$ and $\text{var}(y_{ij}) = \mu_{ij} + \mu_{ij}^2/\sigma_j$. 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.

Tables 2 and 3 about here

5.2 4-variate fitted models

We started our modeling by considering the simplest structure provided by multi-
variate Archimedean copulas. This class assumes the same residual dependence
for each pair. Then we fitted partially symmetric copulas, see subsection 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.,

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

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 ($\theta$) and one ($\theta_{jk}$) for
each marginal $(j, k)$ keeping $\nu$’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
$\nu_3 = \nu_4 = -1$. In this manner we assumed a lower level of dependence for the
(3,4) bivariate margin represented by the parameter $\theta$ of the LT $\phi$ and a higher dependence for the other bivariate margins with the parameters $\theta_{jk}$ representing bivariate dependence exceeding the minimum dependence of the LT $\phi$.

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.

Table 4 about here

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 Kendall’s tau is associated both with marginal and copula parameters, see Figure 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 6-parameters 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 6-parameters 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 paper.

6 Concluding remarks

Modeling multivariate count data based on copulas was described in the present paper. 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 for 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 for e.g., in Nikoloulopoulos and Karlis (2009).
Acknowledgements

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


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 $u_i = 1 - v_i$ and $\tilde{u}_i = -\log u_i$, $i = 1, 2$.

<table>
<thead>
<tr>
<th>Family</th>
<th>$C'(u_1, u_2; \theta)$</th>
<th>LT</th>
<th>$\phi(t; \theta)$</th>
<th>$\theta \in$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gumbel</td>
<td>$e^{-((\tilde{u}_1 + \tilde{u}_2) - 1)^{-1/\theta}}$</td>
<td>LTA</td>
<td>$e^{-t^{1/\theta}}$</td>
<td>$[1, \infty)$</td>
</tr>
<tr>
<td>Mardia-Takahasi</td>
<td>$(u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}$</td>
<td>LTB</td>
<td>$(1 + t)^{-1/\theta}$</td>
<td>$(0, \infty)$</td>
</tr>
<tr>
<td>Joe</td>
<td>$1 - (\tilde{u}_1^{-\theta} + \tilde{u}_2^{-\theta} - \tilde{u}_1^{-\theta}\tilde{u}_2^{-\theta})^{1/\theta}$</td>
<td>LTC</td>
<td>$1 - (1 - e^{-t})^{1/\theta}$</td>
<td>$(1, \infty)$</td>
</tr>
<tr>
<td>Frank</td>
<td>$-\frac{1}{\theta} \log \left{ 1 + \frac{(e^{-\theta u_1 - 1})(e^{-\theta u_2 - 1})}{e^{\theta - 1}} \right}$</td>
<td>LTD</td>
<td>$-\frac{\log(1-(1-e^{-\theta})e^{-t})}{\theta}$</td>
<td>$(0, \infty)$</td>
</tr>
<tr>
<td>Galambos</td>
<td>$u_1 u_2 e^{(\tilde{u}_1^{-\theta} + \tilde{u}_2^{-\theta})^{-1/\theta}}$</td>
<td>$[0, \infty)$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Univariate estimates and standard errors (SE) for the negative binomial regression models for each response variable.
Table 3: Sample Kendall’s tau values for the residuals of the fitted marginal models.

<table>
<thead>
<tr>
<th></th>
<th>food</th>
<th>non-food</th>
<th>hygiene</th>
<th>fresh</th>
</tr>
</thead>
<tbody>
<tr>
<td>food</td>
<td>1.00</td>
<td>0.42</td>
<td>0.27</td>
<td>0.43</td>
</tr>
<tr>
<td>non-food</td>
<td>0.42</td>
<td>1.00</td>
<td>0.23</td>
<td>0.32</td>
</tr>
<tr>
<td>hygiene</td>
<td>0.27</td>
<td>0.23</td>
<td>1.00</td>
<td>0.18</td>
</tr>
<tr>
<td>fresh</td>
<td>0.43</td>
<td>0.32</td>
<td>0.18</td>
<td>1.00</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>AIC</th>
<th>τ_{12}</th>
<th>τ_{13}</th>
<th>τ_{14}</th>
<th>τ_{23}</th>
<th>τ_{24}</th>
<th>τ_{34}</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-parameters mixture of max-id copulas</td>
<td>LTD Gumbel</td>
<td>-31360.713</td>
<td>0.36</td>
<td>0.30</td>
<td>0.39</td>
<td>0.24</td>
</tr>
<tr>
<td>5-parameters mixture of max-id copulas</td>
<td>LTD Gumbel</td>
<td>-31338.59</td>
<td>0.37</td>
<td>0.30</td>
<td>0.39</td>
<td>0.24</td>
</tr>
<tr>
<td>6-parameters mixture of max-id copulas</td>
<td>LTD Gumbel</td>
<td>-31332.74</td>
<td>0.36</td>
<td>0.30</td>
<td>0.39</td>
<td>0.25</td>
</tr>
<tr>
<td>partially symmetric</td>
<td>LTD Gumbel</td>
<td>-31332.74</td>
<td>0.36</td>
<td>0.30</td>
<td>0.39</td>
<td>0.25</td>
</tr>
<tr>
<td>Archimedean</td>
<td>LTD Gumbel</td>
<td>-31387.09</td>
<td>0.35</td>
<td>0.22</td>
<td>0.42</td>
<td>0.22</td>
</tr>
<tr>
<td>LTD</td>
<td>-31483.04</td>
<td>0.32</td>
<td>0.27</td>
<td>0.32</td>
<td>0.27</td>
<td>0.32</td>
</tr>
</tbody>
</table>
Figure 1: Kendall’s tau values computed using Archimedean copulas for a grid of parameter value for each copula (different lines) and Poisson marginal distributions (P(λ)) with the same parameter λ up to 30, higher curves corresponding to higher values of the copula parameter.
Figure 2: Histograms of purchases.
Figure 3: Scatter plots of purchases.