

Likelihood Inference for Exchangeable Continuous Data with Covariates and Varying Cluster Sizes; Use of the Farlie–Gumbel–Morgenstern Model

Catalina Stefanescu ^{a,*} , Bruce W. Turnbull ^b

^a*Management Science and Operations, London Business School, London, UK*

^b*School of Operations Research, Cornell University, Ithaca, NY 14853*

Abstract

This article investigates the Farlie–Gumbel–Morgenstern class of models for exchangeable continuous data. We show how the model specification can account for both individual and cluster level covariates, we derive insights from comparisons with the multivariate normal distribution, and we discuss maximum likelihood inference when a sample of independent clusters of varying sizes is available. We propose a method for maximum likelihood estimation which is an alternative to direct numerical maximization of the likelihood that sometimes exhibits non-convergence problems. We describe an algorithm for generating samples from the exchangeable multivariate Farlie–Gumbel–Morgenstern distribution with any marginals, using the structural properties of the distribution. Finally, we present the results of a simulation study designed to assess the properties of the maximum likelihood estimators, and we illustrate the use of the FGM distributions with the analysis of a small data set from a developmental toxicity study.

Key words: Exchangeable continuous data, Maximum likelihood, Correlation, Accept–reject simulation

* Corresponding author. Management Science and Operations, London Business School, Regent’s Park, London NW1 4SA, United Kingdom. Tel: +44 20 7000 8846, Fax: +44 20 7000 7001.

Email address: cstefanescu@london.edu (Catalina Stefanescu).

1 Introduction

Correlated continuous data arise frequently in many applications, such as developmental toxicity experiments, group randomized clinical trials or cluster sample surveys. We consider a subclass of multivariate continuous distributions in which the responses may be assumed to be exchangeable. A sequence of continuous random variables Y_1, Y_2, \dots is exchangeable if

$$F_{Y_1, \dots, Y_r}(y_1, \dots, y_r) = F_{Y_1, \dots, Y_r}(y_{\pi(1)}, \dots, y_{\pi(r)})$$

for any r , any (y_1, \dots, y_r) and any permutation π of $1, 2, \dots, r$, where F_{Y_1, \dots, Y_r} is the joint cumulative distribution function of Y_1, \dots, Y_r . A large literature exists on the analysis of clustered data, however, few papers have focused specifically on the exchangeable case. Most of the previous approaches involve the mean response and the second order correlation. However, models with higher order interactions are often of interest. In particular, a saturated model in which arbitrary interactions of all orders are allowed, can be used to test adequacy of fit of more parsimonious nested submodels. For the case of binary data, such models have been developed by Bowman and George (1995) and Stefanescu and Turnbull (2003) who propose estimation by maximum likelihood. Estimation becomes more challenging when the data consist of clusters of unequal sizes, which is typical of applications to clinical trials and family studies. In such cases it is important to specify model parameters in such a way that they have a consistent interpretation whatever the cluster size — the *reproductive* or *interpretability* property as termed by Prentice (1988) and Stefanescu and Turnbull (2003), respectively. Essentially, this requires that marginalizing over one response variable leaves unchanged the form of the model and those parameters not involving that variable. Not all multivariate models have this property — see e.g. Prentice (1988), Matthews, Finkelstein and Betensky (2005).

In the case of continuous data, the family of Farlie–Gumbel–Morgenstern (FGM) distributions (Kotz et al., 2000, Ch. 44.13) has considerable appeal for model building. It provides a convenient way of constructing a joint distribution with any specified marginals. The correlation structure can be specified separately in terms of parameters that represent two- and higher-way association, analogous to the approach of Bowman and George (1995) for binary

data. The reproductive property is satisfied. The exchangeable FGM model can also be considered as a polynomial copula — Embrechts et al. (2003), Nelsen (1999).

A number of papers have discussed the theoretical properties of the FGM model, in particular the implied patterns of dependence and the characterizations in terms of different marginal distributions — see Drouot Mari and Kotz (2004) for a summary of the relevant literature. A drawback that limits applicability is that it permits a restricted range for the correlation — see, for example, Huang and Kotz (1999). However, even though only weak dependence can be modelled, this dependence can be either positive or negative, unlike frailty models such as the “positive dependent by mixture” (PDM) family (Shaked, 1975).

In some studies, the assumption of exchangeability is valid for the homogeneous case when no covariates are present. However, in the heterogeneous case it may be reasonable to assume that the clustered responses are exchangeable only after accounting for the presence of explanatory variables, which may act either at cluster level or individual level — see Section 2. We shall say that the response variables Y_1, \dots, Y_r are exchangeable after adjustment for corresponding covariates $\mathbf{x}_1, \dots, \mathbf{x}_r$ if their joint distribution function satisfies:

$$F(y_1, \dots, y_r; \mathbf{x}_1, \dots, \mathbf{x}_r) = F(y_{\pi(1)}, \dots, y_{\pi(r)}; \mathbf{x}_{\pi(1)}, \dots, \mathbf{x}_{\pi(r)}) \quad (1)$$

for any y_1, \dots, y_r and any permutation π of the indices $1, 2, \dots, r$. The effects of covariates on the marginal means are often of interest. Almost always, their inclusion in the analysis is important in order to avoid bias.

Applications of the bivariate FGM model have been discussed in the credit risk literature (e.g. Blum et al., 2002) and in hydrology (Long and Krzysztofowicz, 1992). Most of these studies focus just on the bivariate case, and none of them discusses the introduction of covariates. In this paper we focus on the exchangeable FGM model and discuss its use for the statistical analysis of clustered exchangeable continuous data with covariates and varying cluster sizes.

The paper is structured as follows: Section 2 discusses the specification of the multivariate exchangeable FGM model. We show how both cluster and individual level covariates may be easily included in the analysis, and note the bounds on the association parameters. Section 3 provides a comparison of the

multivariate FGM and the multivariate normal distribution. In Section 4 we discuss estimation of the FGM model parameters using a maximum likelihood approach. In Section 5, we describe an algorithm for generating samples from the exchangeable multivariate FGM distribution and discuss the results of a simulation study designed to assess the properties of the maximum likelihood estimators. We also illustrate the use of the FGM distributions with the analysis of a small data set from a developmental toxicity study. Several directions for future research are outlined in a concluding section.

2 Model Specification

Let $\mathbf{Y} = (Y_1, \dots, Y_r)$ be a cluster of r continuous random variables. One form of the r -dimensional FGM distribution considered by Johnson and Kotz (1975) is defined by the joint cumulative distribution function (cdf):

$$\begin{aligned} F_{\mathbf{Y}}(\mathbf{y}) &= F_{Y_1, \dots, Y_r}(y_1, \dots, y_r) \\ &= \prod_{i_1=1}^r F_{Y_{i_1}}(y_{i_1}) \left[1 + \sum_{1 \leq i_1 < i_2 \leq r} \alpha_{i_1 i_2} \{1 - F_{Y_{i_1}}(y_{i_1})\} \{1 - F_{Y_{i_2}}(y_{i_2})\} \right. \\ &\quad \left. + \dots + \alpha_{12 \dots r} \prod_{i=1}^r \{1 - F_{Y_i}(y_i)\} \right], \end{aligned} \quad (2)$$

where $\{F_{Y_i}\}$ are the marginal distribution functions. In order for this to be a valid multivariate cdf, i.e. nondecreasing in each argument, the constants $\{\alpha_{i_1 \dots i_l}\}_{l=2, \dots, n}$ must satisfy the constraints

$$1 + \sum_{1 \leq i_1 < i_2 \leq r} \epsilon_{i_1} \epsilon_{i_2} \alpha_{i_1 i_2} + \dots + \epsilon_1 \epsilon_2 \dots \epsilon_r \alpha_{12 \dots r} \geq 0, \quad (3)$$

for any $\epsilon_i = \pm 1$ (Kotz et al., 2000, p.54). The FGM system provides a convenient way of constructing the joint distribution with any specified marginals. It clearly satisfies the reproductive property: the distribution of any s -dimensional marginal distribution ($s \leq r$) is also of the FGM form with the same values of a_2, \dots, a_s . Through the choice of the coefficients $\{\alpha_{i_1 \dots i_s}\}$, the model is able to capture the higher level correlation structure, similarly to the approaches of Bowman and George (1995) and Stefanescu and Turnbull (2003) for binary data.

We introduce symmetry into the association structure by setting

$$\begin{aligned}
a_2 &= \alpha_{i_1 i_2}, & 1 \leq i_1 < i_2 \leq r; \\
a_3 &= \alpha_{i_1 i_2 i_3}, & 1 \leq i_1 < i_2 < i_3 \leq r; \\
&\dots, \\
a_r &= \alpha_{12\dots r}.
\end{aligned}$$

Now (2) reduces to

$$\begin{aligned}
F_{\mathbf{Y}}(\mathbf{y}) &= F_{Y_1, \dots, Y_r}(y_1, \dots, y_r) \\
&= \prod_{i_1=1}^r F_{Y_{i_1}}(y_{i_1}) [1 + a_2 \sum_{1 \leq i_1 < i_2 \leq r} \{1 - F_{Y_{i_1}}(y_{i_1})\} \{1 - F_{Y_{i_2}}(y_{i_2})\} \\
&\quad + \dots + a_r \prod_{i=1}^r \{1 - F_{Y_i}(y_i)\}]. \tag{4}
\end{aligned}$$

Following Kotz et al. (2000, p.54), the constraints (3) imply that the $\{a_j; 1 \leq j \leq r\}$ must lie in the polytope given by

$$\prod_{i=1}^r (1 + \epsilon_i a) \geq 0,$$

for any $\epsilon_i = \pm 1$, where a^j is to be interpreted as a_j and $a_1 = 0$. This can be re-written as

$$1 + a_2 \phi_{2,r}(\epsilon_1, \dots, \epsilon_r) + a_3 \phi_{3,r}(\epsilon_1, \dots, \epsilon_r) + \dots + a_r \phi_{r,r}(\epsilon_1, \dots, \epsilon_r) \geq 0, \tag{5}$$

for all $\epsilon_i = \pm 1$, $i = 1, \dots, r$, where $\phi_{i,r}(\epsilon_1, \dots, \epsilon_r)$ is the i -th elementary symmetric function of $\epsilon_1, \dots, \epsilon_r$ defined by

$$\phi_{i,r}(\epsilon_1, \dots, \epsilon_r) = \sum_{1 \leq j_1 < \dots < j_i \leq r} \epsilon_{j_1} \epsilon_{j_2} \dots \epsilon_{j_i}. \tag{6}$$

Note that $\phi_{i,r}(\epsilon_1, \dots, \epsilon_r)$ is the same for any permutation of $(\epsilon_1, \dots, \epsilon_r)$, hence the constraints (5) do not depend on the permutations of $(\epsilon_1, \dots, \epsilon_r)$. Since any given combination $(\epsilon_1, \dots, \epsilon_r)$ can be uniquely characterized by the number of positive ϵ 's (number of $\epsilon_i = 1$) which can take values between 0 and r , it follows that we have at most $r + 1$ distinct constraints (5). For example, for $r = 2$ the constraints are

$$\begin{aligned}
1 + a_2 &\geq 0, \\
1 - a_2 &\geq 0,
\end{aligned}$$

for $r = 3$ they consist in

$$\begin{aligned} 1 + 3a_2 + a_3 &\geq 0, \\ 1 - a_2 - a_3 &\geq 0, \\ 1 - a_2 + a_3 &\geq 0, \\ 1 + 3a_2 - a_3 &\geq 0, \end{aligned}$$

and for $r = 4$ they are given by

$$\begin{aligned} 1 + 6a_2 - 4a_3 + a_4 &\geq 0, \\ 1 + 2a_3 - a_4 &\geq 0, \\ 1 - 2a_2 + a_4 &\geq 0, \\ 1 - 2a_3 - a_4 &\geq 0, \\ 1 + 6a_2 + 4a_3 + a_4 &\geq 0. \end{aligned}$$

Bairamov and Eryilmaz (2004) have considered a special case of (4) in which $a_3 = \dots = a_r = 0$. In this case they show that tight bounds for the admissible range for the parameter a_2 are given by

$$-\frac{1}{\binom{r}{2}} \leq a_2 \leq \frac{1}{\lfloor \frac{r}{2} \rfloor},$$

where $[x]$ denotes the integer part of x .

We now introduce covariates into the model through the marginal cdfs $\{F_{Y_i}; 1 \leq i \leq r\}$. For a cluster of r continuous response random variables (Y_1, \dots, Y_r) , suppose \mathbf{x}_i represents a p -vector of covariate values to Y_i for each $i; 1 \leq i \leq r$. If component j ($1 \leq j \leq p$) of covariate \mathbf{x} acts at cluster level, then those corresponding components of $\mathbf{x}_1, \dots, \mathbf{x}_r$ are all equal. For individual-level covariates, the corresponding components will typically be different. In particular, the first component of each \mathbf{x} will typically be one, indicating the presence of an intercept term. We account for the presence of covariates by taking the $\{F_{Y_i}\}$ to be members of a location–scale family, i.e.

$$F_{Y_i}(y) = G\left(\frac{1}{\sigma}(y - \boldsymbol{\beta}\mathbf{x}_i)\right) \quad (7)$$

for some standard distribution G . For example, G could be specified parametrically such as the standard normal $N(0, 1)$, or it can be left unspecified leading to a semi-parametric model. Finally, with F_{Y_i} defined by (7) we have a model

that satisfies (1), and the response variables Y_1, Y_2, \dots, Y_r are exchangeable after adjustment for covariates $\mathbf{x}_1, \dots, \mathbf{x}_r$.

Since the joint distribution is assumed to be continuous, we can write the joint density of the FGM model as

$$\begin{aligned}
& f_{Y_1, \dots, Y_r}(y_1, \dots, y_r; \mathbf{x}_1, \dots, \mathbf{x}_r) \\
&= \prod_{i=1}^r \frac{1}{\sigma} g\left(\frac{y_i - \beta \mathbf{x}_i}{\sigma}\right) \left[1 + \sum_{1 \leq i_1 < i_2 \leq r} a_2 \left\{1 - \frac{2}{\sigma} G\left(\frac{y_{i_1} - \beta \mathbf{x}_{i_1}}{\sigma}\right)\right\} \left\{1 - \frac{2}{\sigma} G\left(\frac{y_{i_2} - \beta \mathbf{x}_{i_2}}{\sigma}\right)\right\}\right. \\
&\quad \left. + \dots + a_r \prod_{i=1}^r \left\{1 - \frac{2}{\sigma} G\left(\frac{y_i - \beta \mathbf{x}_i}{\sigma}\right)\right\}\right], \tag{8}
\end{aligned}$$

where g is the derivative of G .

In Section 4 we describe inference procedures for the case when vectors of responses \mathbf{Y} are available from a number of independent clusters of possibly differing sizes.

3 Comparison with the multivariate normal distribution

The FGM distributions have the advantage of being able to account for higher orders of association between components of the same cluster.¹ In particular, they allow for higher order dependence even when the second-order dependence is zero, which is not always the case for other distributions, such as the multivariate normal.

Indeed, as an example let us assume that (Y_1, Y_2, Y_3) follow the trivariate normal distribution with standard marginals and with a second-order correlation coefficient equal to zero. This implies that $Y_1, Y_2,$ and Y_3 are uncorrelated, hence independent. Therefore $E[Y_1 | Y_2 = y_2, Y_3 = y_3] = E[Y_1] = 0$, and all higher order correlations between $Y_1, Y_2,$ and Y_3 are zero.

On the other hand, let us assume that (Y_1, Y_2, Y_3) have a trivariate FGM distribution with standard normal marginals and with no second-order correlation, so that $a_2 = 0$. Consider the case when $a_3 \neq 0$, so that third-order

¹ For definitions and discussions of higher order association parameters, see Bahadur (1961).

correlation is nonzero. The joint density of Y_1 , Y_2 , and Y_3 is

$$f_{Y_1, Y_2, Y_3}(y_1, y_2, y_3) = \prod_{i=1}^3 \varphi(y_i) [1 + a_3 \prod_{i=1}^3 \{1 - 2\Phi(y_i)\}], \quad (9)$$

where $\varphi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and cumulative distribution functions, respectively. After some algebra and integration by parts, it follows that the conditional mean of Y_1 is given by

$$E[Y_1 | Y_2 = y_2, Y_3 = y_3] = -\frac{a_3}{\sqrt{2\pi}} \left[1 + \frac{1}{\sqrt{2}} \right] \times \varphi(y_2) [1 - 2\Phi(y_2)] \cdot \varphi(y_3) [1 - 2\Phi(y_3)], \quad (10)$$

which is generally different from the unconditional mean $E[Y_1] = 0$, and different from the conditional means $E[Y_1 | Y_2]$ and $E[Y_1 | Y_3]$ which are also zero. (To see that the $E[Y_1 | Y_2] = E[Y_1 | Y_3] = 0$, notice that expression (10) is an odd function of y_2 and y_3 .) For illustration, Figure 1 plots the conditional mean $E[Y_1 | Y_2, Y_3]$ for different values of Y_2 and Y_3 , for a trivariate FGM distribution with standard normal marginals and with association parameters $a_2 = 0$ and $a_3 = 0.8$.

[Figure 6 about here.]

4 Model Estimation

Suppose that K independent clusters of varying sizes $\{r_1, r_2, \dots, r_K\}$ are available for inference. Let $(y_{k1}, \mathbf{x}_{k1}), \dots, (y_{kr_k}, \mathbf{x}_{kr_k})$ denote the responses and covariates in the k -th cluster. Denote the maximum cluster size by $R = \max\{r_k; 1 \leq k \leq K\}$.

If we wish to combine information from clusters of varying sizes for the purpose of inference, it is natural to look for model formulations in which the parameters conserve their meaning for different cluster sizes — for example, $\text{Corr}(Y_1, Y_2)$ is constant irrespective of r_k ($r_k = 2, \dots, R$). Prentice (1988) calls this the *reproductive* property and Stefanescu and Turnbull (2003) term it the *interpretability* assumption. The exchangeable FGM model given by (4) and (7) satisfies naturally the interpretability assumption, as long as the constraints (5) are satisfied for the maximum cluster size R . Indeed, notice

that if (a_2, a_3, \dots, a_r) satisfy (5) for r , then $(a_2, a_3, \dots, a_{r-1})$ satisfy (5) for $r - 1$. The reverse, however, is not true; i.e. if $(a_2, a_3, \dots, a_{r-1})$ satisfy the constraints (5) for $r - 1$, then $(a_2, a_3, \dots, a_{r-1}, a_r = 0)$ do not necessarily satisfy the constraints (5) for r .

The likelihood of the parameters $\boldsymbol{\beta}$, σ^2 and (a_2, a_3, \dots, a_R) is a product of K cluster likelihoods:

$$L(\boldsymbol{\beta}, \sigma^2, (a_2, \dots, a_R) \mid \{\mathbf{y}_k\}, \{\mathbf{x}_k\}) = \prod_{k=1}^K f_{Y_{k1}, \dots, Y_{kr_k}}(y_1, \dots, y_{r_k}; \mathbf{x}_1, \dots, \mathbf{x}_{r_k}), \quad (11)$$

where $f_{Y_1, \dots, Y_r}(y_1, \dots, y_r; \mathbf{x}_1, \dots, \mathbf{x}_r)$ is the joint density function given by (8). If G is specified as a given standard distribution, such as the standard normal, then the log-likelihood derived from (11) can be maximized with a standard optimization algorithm. Standard errors may be obtained by numerically computing the Hessian, as long as the estimates \hat{a}_i are not on the boundary of the parameter space. In our experience, however, using a standard optimization routine, such as *fmincon* in MATLAB, leads to occasional problems of non-convergence and instability. Therefore, in the rest of this section we propose an alternative method of maximizing the likelihood. This method was used for the simulation experiments reported in Section 5, and we have experienced no instability problems in any of the thousands of simulation runs.

The approach that we propose is based on an iterative scheme that relies on the particular structure of the FGM joint density function $f_{Y_1, \dots, Y_r}(y_1, \dots, y_r)$ which is linear in the unknown parameters $\mathbf{a} = \{a_2, \dots, a_r\}$. At each iteration, the parameters $\boldsymbol{\beta}$ and σ^2 of the marginal distributions are first held fixed and so the marginal univariate $G(\cdot)$ and its derivative $g(\cdot)$ in (8) are known and only the FGM parameters $\mathbf{a} = \{a_2, \dots, a_r\}$ are estimated. The optimal $\boldsymbol{\beta}$ and σ^2 are then determined through a grid search over a $(p + 1)$ -dimensional grid, where each grid point evaluation involves a constrained optimization over the a_i with a linear objective function. Note that in the semi-parametric case where the form of the continuous density g is unspecified, the marginal distribution can also be estimated nonparametrically, for example with a kernel density estimate.

The constrained optimization step is based on the following change of variables. Let $\tilde{y}_k = f_{\mathbf{Y}}(\mathbf{y}_k, \mathbf{a})$ for $k = 1, \dots, K$, and define also $\tilde{y}_{K+1} = a_2$, $\tilde{y}_{K+2} = a_3, \dots, \tilde{y}_{K+r-1} = a_r$. The objective function in (11) is then sim-

ply $\prod_{k=1}^K \tilde{y}_k$. It involves only the first K of the decision variables and it is log-concave (in the y 's). There are K linear equality constraints of the form $\tilde{y}_k = f_{\mathbf{Y}}(\mathbf{y}_k, \tilde{y}_{K+1}, \dots, \tilde{y}_{K+r-1})$ for $k = 1, \dots, K$, which are given by the change of variables. In addition, there are also $r + 1$ linear inequality constraints between $\tilde{y}_{K+1}, \dots, \tilde{y}_{K+r-1}$ given by (5). This reparametrization makes it particularly straightforward to compute the derivatives and the Hessian of the objective function with respect to the decision variables, which in practice is helpful for the stability of the numerical routines.

The estimation method that we outlined above is particularly easy to implement, and a study of its performance through simulation experiments with normal margins is described in the next section. As remarked in Section 1, the exchangeable FGM model can also be considered as a polynomial copula. Other estimation approaches for parameters for general copula models proposed in the literature include the IFM (inference function for margins) method developed by Joe and Xu (1996), the two-stage parametric maximum likelihood approach described by Shih and Louis (1995), and the omnibus estimator discussed by Genest and Werker (2002).

5 Applications: Simulation Study and Examples

In this section, we first describe an algorithm for generating random samples from the exchangeable FGM distribution. We then discuss the design and the results of a small simulation study which we conducted in order to investigate the performance of the maximum likelihood estimators computed with the methodology from Section 4. Finally, we illustrate the use of the FGM distributions with the analysis of a data set from a developmental toxicity study. In this section we shall focus on the homogeneous case when there are no covariates.

5.1 Generating exchangeable FGM variates

Random samples from the exchangeable FGM distribution with any marginals may be simulated using an accept-reject approach (Devroye, 1986). Let f be the marginal density function which is common to all components, since in this section we assume that there are no covariates. The simulation algorithm based on the accept-reject method can be implemented as follows:

Step 1. Compute the bound

$$M = 1 + \binom{r}{2} |a_2| + \dots + \binom{r}{r-1} |a_{r-1}| + |a_r|.$$

Step 2. Generate $\mathbf{Y} = Y_1, \dots, Y_r$ independently from the common univariate marginal density $f(\cdot)$.

Step 3. Generate a standard uniform $U \sim U[0, 1]$.

Step 4. Compute the instrumental joint density

$$h(\mathbf{y}) = h(y_1, \dots, y_r) = \prod_{i=1}^r f(y_i).$$

Step 5. Compute the joint density $f_{\mathbf{Y}}(\mathbf{y})$ as given by (8) with $\boldsymbol{\beta} = \mathbf{0}$.

Note that $f_{\mathbf{Y}}(\mathbf{y})$ in the FGM model satisfies $f_{\mathbf{Y}}(\mathbf{y}) \leq M \cdot h(\mathbf{y})$.

Step 6. Accept \mathbf{Y} if and only if

$$U \leq \frac{f_{\mathbf{Y}}(\mathbf{Y})}{M h(\mathbf{Y})}.$$

The bound M is not tight, particularly for moderate or large values of r . However, each iteration of Step 2 is so fast that the overall method is still computationally efficient.

5.2 Simulation examples

In order to assess the performance of the maximum likelihood estimators we generated data from exchangeable FGM distributions with no covariates, under two simulation scenarios. In both scenarios the marginal distributions are standard normal ($\beta_0 = 0$, $\sigma^2 = 1$), and we computed the maximum likelihood estimates using the optimization procedure described in Section 4 (iterative grid search). We repeated the process for 10000 iterations. Table 1 reports the average bias and the mean squared error (MSE) for all estimated parameters.

[Table 1 about here.]

Under the first scenario the cluster size is $R = 3$, and the true FGM parameters are $a_2 = 0$ and $a_3 = .8$. At each iteration the sample data contains

$K = 300$ clusters generated using the accept–reject algorithm outlined in Section 5.1. The bias and mean square errors are larger for the FGM parameter estimates than for the estimates $\widehat{\beta}_0$ and $\widehat{\sigma}^2$, implying that, as expected, the association structure is more difficult to estimate than the marginal parameters.

At each iteration under the first scenario we have also fitted to the generated data set a multivariate normal distribution with equal correlation ρ . The average of the estimate $\hat{\rho}$ over the 10000 iterations is -0.0014, suggesting that the true value of ρ is zero and thus that the data are independent. It is apparent, however, that this is not the case, since third-order association is positive even though second-order association is zero. Consistent with the discussion in Section 3, using the FGM distribution rather than the multivariate normal as a model for the data in this example has the advantage of not leading to the incorrect inference of independence.

Under the second scenario the maximum cluster size is $R = 4$, and the true FGM parameters are $a_2 = .3$, $a_3 = .2$ and $a_4 = .1$. At each iteration the sample data contains $K = 300$ clusters generated using the accept–reject algorithm, with 100 clusters of size 2, 100 clusters of size 3, and 100 clusters of size 4. As in the previous scenario with a smaller cluster size, the bias and mean square error are larger for the FGM parameter estimates than for the marginal parameter estimates $\widehat{\beta}_0$ and $\widehat{\sigma}^2$. The mean square error also increases with the order of the FGM association parameters, implying that higher order association parameters are more difficult to estimate accurately than lower order parameters, especially since there were fewer of the larger clusters in the sample.

As an additional simulation investigation, we also checked whether the FGM distribution can identify zero correlation when data are uncorrelated. To this purpose, we generated samples from the trivariate normal distribution with zero correlation and standard normal marginals, to which we fitted the trivariate exchangeable FGM distribution. Over 100 iterations, the averages of the estimated FGM parameters were $\hat{a}_2 = 0.0014$ and $\hat{a}_3 = -0.0115$, suggesting that the true values are zero and that indeed the data are uncorrelated.

5.3 Example: Developmental Toxicity Data

In this section we illustrate the use of the FGM distribution with the analysis of a data set from a developmental toxicity study conducted at the National Center for Toxicological Research. The data is a subset of the data used by Ahn and Chen (1997). In general, teratological experiments involve treating a pregnant animal with some compound of interest and measuring responses on the fetuses in the litter. The litters constitute the clusters, and the responses from the same cluster are usually correlated. The data in our example come from a study of exposure to the herbicide 2,4,5-trichlorophenoxyacetic acid, where one of the developmental endpoints was the fetal weight recorded in the litters. One outbred and four inbred strains of mice were exposed to several dose levels, ranging from 0 to 60 mg/kg/day. In this example, we focus on the 89 mice litters from the A/JAX inbred strain who were not exposed to the compound (the dose level was 0 mg/kg/day). We restrict attention to the first four mice from each litter, so that the maximum cluster size is 4. The response variable is the fetal weight for each mouse. Since the litters in our data set were not exposed to the chemical compound, we expect that the correlation between fetal weights would be due to shared genetic factors.

Table 2 reports the results from fitting the FGM distribution with maximum cluster size $R = 4$. The table gives maximum likelihood estimates of the marginal parameters β and σ , and of the FGM association parameters a_2 , a_3 and a_4 , with standard errors in parentheses. We computed the standard errors of the estimates using the nonparametric bootstrap approach (Efron and Tibshirani, 1998), implemented as follows. We drew a simple random sample with replacement of the available litters, and for each of the selected litters we included the corresponding fetal weights in the sample data. We then recorded the model estimates computed from the bootstrapped data, and repeated the process 500 times resulting in 500 sets of estimates. The bootstrap standard error of each parameter is then computed as the sample standard deviation across these 500 sets of estimates.

[Table 2 about here.]

The standard errors of the estimates from Table 2 strongly suggest that the FGM association parameters a_2 , a_3 and a_4 are different from zero, hence there is indeed significant intra-litter correlation between fetal weights.

6 Discussion

In this paper we investigated the exchangeable specification of the multivariate FGM model. The exchangeable FGM model is suitable for the analysis of multivariate data with small to moderate dependence and potentially varying cluster sizes, especially when interest lies not only in the estimation of marginal effects and bivariate correlations but also of higher order association parameters. We proposed a method for maximum likelihood estimation when a sample of clusters of varying sizes is available. This method is an alternative to direct numerical maximization of the likelihood that sometimes leads to non-convergence and instability problems. Our estimation approach performs well, as outlined by simulation experiments.

Several generalizations of the FGM model have been proposed in the literature, motivated by the restricted degree of association that the model is able to capture (Farlie (1960), Huang and Kotz (1984), Peristiani (1991), Huang and Kotz (1999), Bairamov, Kotz and Bekci (2001)). A summary of the extensions of the FGM model designed to increase the maximum value of the correlation coefficient is given in Drouet Mari and Kotz (2004).

One potential generalization of our multivariate exchangeable FGM model is to use nonparametric instead of normal margins. When no covariates are available, estimation for this model may be done either by considering the empirical cumulative estimation function or a kernel smoothed estimate. When covariates are available, the regression parameters β may be estimated by least-squares, then a kernel smoothed estimate of the error density can be obtained from the residuals. These and other generalizations are the subject of further research.

References

- [1] H. Ahn and J.J. Chen, Tree-structured logistic model for over-dispersed binomial data with application to modeling developmental effects. *Biometrics* 53 (1997) 435–455.
- [2] R.R. Bahadur, A representation of the joint distribution of responses to n dichotomous items. In: H. Solomon (Ed.), *Studies in Item Analysis and Prediction*, Stanford University Press, California, 1961, pp.158–168.
- [3] I. Bairamov and S. Eryilmaz, Characterization of symmetry and ex-

- ceedance models in multivariate FGM distributions. *Journal of Applied Statistical Science* 13 (2004) 87–99.
- [4] I. Bairamov, S. Kotz, and M. Bekci, New generalized Farlie–Gumbel–Morgenstern distributions and concomitants of order statistics. *Journal of Applied Statistics* 28 (2001) 521–536.
- [5] P. Blum, A. Dias, and P. Embrechts, The ART of dependence modelling: The latest advances in correlation analysis. In: M. Lane (Ed.), *Alternative Risk Strategies*, Risk Waters Group, London, 2002, pp.339–356.
- [6] D. Bowman and E. George, A saturated model for analyzing exchangeable binary data: Applications to clinical and developmental toxicity studies. *Journal of the American Statistical Association* 90 (1995) 871–879.
- [7] L. Devroye, *Non–Uniform Random Variate Generation*. Springer, New York, 1986.
- [8] D. Drouot Mari and S. Kotz, *Correlation and Dependence*. Imperial College Press, London, 2004.
- [9] B. Efron, and R.J. Tibshirani, *An Introduction to the Bootstrap*. Chapman and Hall, Boca Raton, 1998.
- [10] P. Embrechts, F. Lindskog, and A. McNeil, Modelling dependence with copulas and applications to risk management. In: S.T. Rachev (Ed.), *Handbook of Heavy Tailed Distributions in Finance*, Elsevier, 2003, pp.329–384.
- [11] D.J.G. Farlie, The performance of some correlation coefficients for a general bivariate distribution. *Biometrika* 47 (1960) 307–323.
- [12] C. Genest and B. Werker, Conditions for the asymptotic semiparametric efficiency of an omnibus estimator of dependence parameters in copula models. In: C.M. Cuadras and J.A. Rodriguez Lallena (Eds.), *Proceedings of the Conference on Distributions with Given Marginals and Statistical Modelling*, 2002.
- [13] J.S. Huang and S. Kotz, Correlation structure in iterated Farlie–Gumbel–Morgenstern distributions. *Biometrika* 71 (1984) 633–636.
- [14] J.S. Huang and S. Kotz, Modifications of the Farlie–Gumbel–Morgenstern distributions. A tough hill to climb. *Metrika* 49 (1999) 135–145.
- [15] H. Joe and J.J. Xu, The estimation method of inference functions for margins for multivariate models. Technical report no.166, Department of Statistics, University of British Columbia, 1996.
- [16] N. Johnson and S. Kotz, On some generalized Farlie–Gumbel–

- Morgenstern distributions. *Communications in Statistics* 4 (1975) 415–427.
- [17] S. Kotz, B. N. Balakrishnan and N. Johnson, *Continuous Multivariate Distributions. Volume 1: Models and Applications* (2 ed.). Wiley, New York, 2000.
- [18] D. Long and R. Krzysztofowicz, Farlie–Gumbel–Morgenstern bivariate densities: Are they applicable in hydrology? *Stochastic Hydrology and Hydraulics* 6 (1992) 47–54.
- [19] A.G. Matthews, D.M. Finkelstein, and R.A. Betensky, Analysis of familial aggregation in the presence of varying family sizes. *Applied Statistics* 54 (2005) 847–862.
- [20] R.B. Nelsen, *An Introduction to Copulas*. Springer, New York, 1999.
- [21] S. Peristiani, The F–system distribution as an alternative to multivariate normality: An application in multivariate models with qualitative dependent variables. *Communications in Statistics — Theory and Methods* 20 (1991) 147–163.
- [22] R.L. Prentice, Correlated binary regression with covariates specific to each binary observation. *Biometrics* 44 (1988) 1033–1048.
- [23] W.R. Schucany, W.C. Parr, and J.E. Boyer, Correlation structure in Farlie–Gumbel–Morgenstern distributions. *Biometrika* 65 (1978) 650–653.
- [24] M. Shaked, A note on the exchangeable generalized Farlie–Gumbel–Morgenstern distributions. *Communications in Statistics* 4 (1975) 711–721.
- [25] J.H. Shih, and T.A. Louis, Inferences on the association parameter in copula models for bivariate survival data. *Biometrics* 51 (1995) 1384–1399.
- [26] C. Stefanescu and B. W. Turnbull, Likelihood inference for exchangeable binary data with varying cluster sizes. *Biometrics* 59 (2003) 18–24.

Fig. 1. Plot of the conditional mean $E[Y_1 | Y_2, Y_3]$ for different values of Y_2 and Y_3 , for a trivariate FGM distribution with standard normal marginals and with association parameters $a_2 = 0$ and $a_3 = 0.8$.

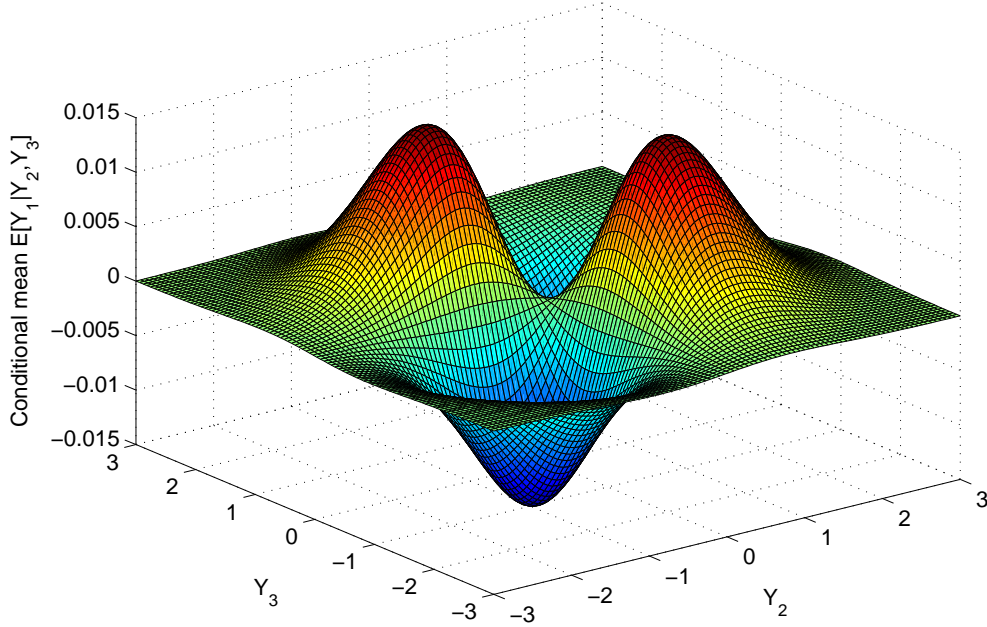


Table 1

Simulation study for maximum likelihood estimates. Average bias and mean squared error (MSE) based on 10000 iterations. $K = 300$ clusters of varying sizes were generated from a multivariate exchangeable FGM model with standard normal marginal distributions ($\beta_0 = 0, \sigma^2 = 1$).

Maximum		$\widehat{\sigma}^2$	$\widehat{\beta}_0$	\widehat{a}_2	\widehat{a}_3	\widehat{a}_4
cluster size						
$R = 3$	Bias	-0.0009	0.0000	0.0051	-0.0662	–
	MSE	[0.0006]	[0.0011]	[0.0070]	[0.0486]	–
$R = 4$	Bias	-0.0006	-0.0004	-0.0040	-0.0222	-0.0120
	MSE	[0.0007]	[0.0015]	[0.0115]	[0.0512]	[0.2193]

Table 2

Illustration example for maximum likelihood estimates. The FGM distribution with maximum cluster size $R = 4$ has been fitted to fetal weight data from 89 mice litters (data adapted from Ahn and Chen (1997)). The table reports maximum likelihood estimates of the marginal parameters β and σ , and of the FGM parameters a_2 , a_3 and a_4 , with standard errors in parentheses.

Parameter	$\hat{\sigma}$	$\hat{\beta}$	\hat{a}_2	\hat{a}_3	\hat{a}_4
MLE	7.9391	57.7265	0.8076	0.1914	0.6163
	[0.3625]	[0.8048]	[0.1253]	[0.1129]	[0.2288]