A family of multivariate distributions with prefixed marginals

Isidro R. Cruz_Medina*, F. Garcia_Paez and J. R. Pablos_Tavares.

Recursos Naturales, Instituto Tecnológico de Sonora
Cinco de Febrero 818, Cd. Obregón Son. CP 85 000 México.

Abstract

When dealing with multivariate data, it is convenient to have a stock of adaptable families flexible enough to be constructed from a specific set of univariate marginals. Such a family of multivariate distributions is constructed in this paper by a transformation of the multivariate normal distribution. It is shown that the marginal and conditional distributions belong to the same family which has a simple expression. The dependence parameter for this family, in the bivariate case, is the well known product moment correlation coefficient. In contrast, most bivariate copulas have parameters difficult to interpret. An application to hydrological data shows that the bivariate version of the proposed family compares well with bivariate copulas.

Key words: Copulas, Multivariate distribution; Marginal and conditional distribution; Bivariate distributions.

1. Introduction

Multivariate distributions are needed to model complex events in everyday science applications. Although the multinormal distribution has been extensively used for the analysis of multivariate data, there are many important situations where the univariate components of complex phenomena are clearly non-normal. In areas like hydrology (Yue et al, 2001), life testing and biostatistics (Park, 2004), for example, variables are positive

* Corresponding autor: E-mail address: rcruz@itson.mx
and skewed and distributions like gamma, lognormal, inverse Gaussian and others are commonly used. In most cases it may be difficult to completely specify the joint distribution of the available random vector; however, it may be possible to specify the marginal distribution of their components and their correlation structure. The construction of bivariate and multivariate distributions functions is a subject which has been studied intensively. In particular, the construction of bivariate distributions with specified marginals has been discussed by Moran (1969), Farlie (1960), Plackett (1965), Johnson & Tenenbein (1981), Genest & MacKay (1986) and Genest (1987). Marshall & Olkin (1988), Koehler & Symanowski (1995) and Arnold et al., (2006) have presented methods for constructing multivariate distributions. The general method for constructing multivariate distributions was provided by the concept of copula given by Sklar (1959) with the theorem.

**Theorem.** If \( H \) is a distribution function on \( R^p \), with one-dimensional marginal distributions functions \( F_1, F_2, \ldots, F_p \). There is a copula \( C \) such that

\[
H(X) = C(F_1(X_1), F_2(X_2), \ldots, F_p(X_p)) \tag{1}
\]

If \( F \) is continuous, then the copula \( C \) satisfying (1) is unique and is given by

\[
C(U) = H(F_1^{-1}(U_1), F_2^{-1}(U_2), \ldots, F_p^{-1}(U_p)) \tag{2}
\]

for \( U \in (0, 1)^p \) where \( F_i^{-1}(V) = \inf \{ W : F_i(W) \geq V \} \), \( i=1,2,\ldots,p \).

where

\[
X^T = (X_1, X_2, \ldots, X_p) \quad \text{and} \quad U^T = (U_1, U_2, \ldots, U_p) \tag{3}
\]

are the transposes of column vectors \( X \) and \( U \) respectively.

Conversely, If \( C \) is a copula on \([0,1]^p\) and \( F_1, F_2, \ldots, F_p \) are distribution functions on \( R^p \), then the function \( H \) defined by (1) is a distribution function on \( R^p \) with one dimensional distributions \( F_1, F_2, \ldots, F_p \).
The proof of this theorem can be seen in Nelsen (2006), who describes different methods for constructing copulas. In this paper, the inversion method is applied to obtain a multivariate family. The inversion method begins with a multivariate distribution function $G$, with margins $G_1, G_2, \ldots, G_p$, to obtain the copula:

$$C(U) = G(G_1^{-1}(u_1), G_2^{-1}(u_2), \ldots, G_p^{-1}(u_p))$$

with this copula, a new multivariate distribution with arbitrary margins $F_1, F_2, \ldots, F_p$, can be constructed using Sklar’s theorem:

$$H(X) = C(F_1(x_1), F_2(x_2), \ldots, F_p(x_p))$$

The inversion method is quite popular, it was used by Moran (1969), Koehler & Symanowski (1995) and Arnold et al., (2006) to obtain multivariate distributions. In this paper, the multivariate normal distribution is used as the G function in the inversion method, this distribution has the property that the dependence structure is completely specified with the covariance matrix. In section 2, the probability density function of this family, which is called the Generalized Moran Family (GMF), is obtained. In section 3, the conditional and marginal distributions of any subset of variables are derived and it is shown that they belong to the same GMF. In section 4, an application of this family is presented and some methods of estimation are discussed.

2. A family of multivariate distributions

For the construction of the multivariate GMF, the following well known results are used.

1. Dependence among variables in a multivariate standard normal distribution is completely specified by the correlation matrix (this property makes the multivariate normal distribution ideal for expressing the dependence among variables with other distributions).
2. The cumulative distribution function (CDF) for any continuous univariate distribution has a uniform distribution function \( u(0, 1) \), in the interval \((0, 1)\). This property allows to transform any distribution onto another. Given a continuous distribution function \( f(x) \), its cumulative distribution function \( F(x) \) has uniform \( u(0, 1) \) distribution. The application of the quantile function \( G^{-1}(u) \) to this uniform distribution gives a variable \( y \) with distribution function \( g(y) \).

3. The change of variable theorem (Casella, 1990) allows one to obtain a multivariate distribution by transformations on an original set of variables.

Let's start with a \( p \)-dimensional random vector \( Z \), with multivariate standard normal distribution (Mardia et al, 1979) given by (4):

\[
f_Z(Z) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2} Z^T \Sigma^{-1} Z\right]
\]

(4)

Where \( Z \) is a column vector, which components \( Z_i \), follows a standard normal distribution, \( \Sigma^{-1} \) is the inverse of the covariance (correlation) matrix given by (5).

\[
\Sigma = \begin{pmatrix}
1 & \rho_{12} & \cdots & \rho_{1p} \\
\rho_{12} & 1 & \cdots & \rho_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
\rho_{1p} & \rho_{2p} & \cdots & 1
\end{pmatrix}
\]

(5)

Let \( X \) be the column vector defined in (3), whose components \( x_i \), follow possibly different known marginal distributions with PDFs given by (6):

\[
f_i(X_i; \theta_i) \text{ for } i = 1, 2, \ldots p
\]

(6)

where \( \theta_i \) is a column vector of \( t \) parameters. For example, for the gamma family, \( t=2 \), and \( \theta_i = (\alpha_i, \lambda_i) \); \( \alpha_i \) and \( \lambda_i \) are the scale and shape parameters of the gamma
distribution. The corresponding CDFs of variables $X_i$ are $F_i(X_i; \theta_i)$ given by equation (7):

$$F_i(X_i; \theta_i) = \int_{-\infty}^{X_i} f_i(t; \theta_i) \, dt, \quad i = 1, 2, \ldots p$$

(7)

Vector $X$ can be obtained from vector $Z$ with the element-wise one-to-one transformation:

$$X_i = F_i^{-1}(\Phi(Z_i)), \quad i = 1, 2, \ldots p$$

(8)

where $\Phi(Z_i)$ represents the standard normal CDF given by (9).

$$\Phi(Z_i) = (2\pi)^{-1/2} \int_{-\infty}^{Z_i} \exp\left(-1/2 t^2\right) dt$$

(9)

The inverse transformation of equation (9) is:

$$Z_i = \Phi^{-1}(F_i(X_i)), \quad i = 1, 2, \ldots p$$

(10)

Applying the change of variable theorem, the joint PDF of vector $X$ is given by (11).

$$h(X) = f_Z(Z) |J|$$

(11)

where $Z^T = [\Phi^{-1}(F_1(X_1)), \Phi^{-1}(F_2(X_2)), \ldots, \Phi^{-1}(F_p(X_p))], and J is the Jacobian of the inverse transformation given by:

$$J = \prod_{i=1}^{p} \frac{\partial Z_i}{\partial X_i}$$

(12)

This equality holds because function $Z_i$ has only one argument: Variable $X_i$, $\frac{\partial Z_i}{\partial X_i}$ is obtained applying the well known formulas for the derivatives of a composition of functions and for the inverse function.

$$\frac{\partial Z_i}{\partial X_i} = \frac{\partial \Phi^{-1}(F_i(X_i))}{\partial X_i} = \frac{f_i(X_i; \theta_i)}{\varphi(\Phi^{-1}(F_i(X_i)))} =$$
\[ = (2\pi)^{-1/2} \exp \left\{ \frac{1}{2} \mathbf{Z}_i^2 \right\} f_i(X_i; \theta_i), \quad i = 1, 2, \ldots, p \]

where \( \varphi \) stands for the PDF of the standard normal distribution and \( Z_i = \Phi^{-1}(F_i(X_i)) \).

Therefore, the PDF of the GMF has the compact expression:

\[
h(X) = |\Sigma|^{-1/2} \exp \left[ -\frac{1}{2} \mathbf{Z}^T (\Sigma^{-1} - I) \mathbf{Z} \right] \prod_{i=1}^p f_i(X_i; \theta_i)
\]

(13)

where \( I \) stands for the identity matrix of order \( p \).

3. Marginal and conditional distributions

3.1 Marginal distributions

If the \( p \)-dimensional vector \( X \) belongs to the GMF we will say that

\[
X \sim M \left( 0, \Sigma, \left\{ f_i(X_i; \theta_i) \right\}_{i = 1, \ldots, p} \right)
\]

(14)

where the first element, the zero \( p \)-component vector is the mean and \( \Sigma \), is the covariance matrix of the multivariate normal distribution used in the transformation, the third set represents the marginal PDF’s of the GMF. If the random vector \( X \) is partitioned in two vectors \( X_1 \) and \( X_2 \) with \( r \) and \( q = (p-r) \) components, the construction of the family suggests that: \( X_1 \sim M \left( 0, \Sigma_{11}, \left\{ f_i(X_i; \theta_i) \right\}_{i = 1, \ldots, r} \right) \), where \( \Sigma_{11} \) is defined in the corresponding partitioning of matrix \( \Sigma \):

\[
\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}
\]

Proof: By definition:

\[
h_{X_1}(X_1) = \int_{X_2} h(X) dX_2
\]

(15)

For the evaluation of this integral, the \( q \)-component vector \( X_2 \) will be transformed into the vector \( Z_2 \) of standard normal components, where vector \( Z^T = (Z_1^T, Z_2^T) \) is partitioned as vector \( X_2 \). The Jacobian of this transformation (8) is:
\[
J = \prod_{i=r+1}^{p} \frac{\partial X_i}{\partial Z_i} = (2\pi)^{-q/2} \text{Exp} \left\{ -\frac{\sum_{i=r+1}^{p} Z_i^2}{2} \left[ \prod_{i=r+1}^{p} f_i(X_i; \theta_i) \right] \right\}^{-1} \tag{16}
\]

therefore,
\[
\mathbf{h}_{X_1}(X_1) = \int_{\mathbb{Z}_2} (2\pi)^{-q/2} |\Sigma|^{-1/2} \text{Exp} \left\{ -\frac{1}{2} Z^T \Sigma^{-1} Z + \frac{\sum_{i=r+1}^{p} Z_i^2}{2} \right\} \prod_{i=r+1}^{p} f_i(X_i; \theta_i) \, d\mathbf{Z}_2
\tag{17}
\]

The expressions for the determinant and the inverse of a partitioned matrix (Graybill, 1969):

\[
|\Sigma| = |\Sigma_{11}| \begin{vmatrix} \Sigma_{22} \end{vmatrix}
\]

\[
\Sigma^{-1} = \begin{bmatrix}
\Sigma_{11}^{-1} + \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1} & \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \\
\Sigma_{21} \Sigma_{11}^{-1} & \Sigma_{22}^{-1}
\end{bmatrix}
\tag{18}
\]

where:

\[
\Sigma_{22,1} = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}
\]

Allow to express the quadratic form in (17) as:

\[
Z^T \Sigma^{-1} Z = Z_1^T \Sigma_{11}^{-1} Z_1 + (Z_2 - \Sigma_{21} \Sigma_{11}^{-1} Z_1)^T \Sigma_{22,1}^{-1} (Z_2 - \Sigma_{21} \Sigma_{11}^{-1} Z_1)
\tag{19}
\]

Therefore:

\[
\mathbf{h}_{X_1}(X_1) = \begin{vmatrix} \Sigma_{11} \end{vmatrix}^{-1/2} \text{Exp} \left\{ -\frac{1}{2} Z_1^T \Sigma_{11}^{-1} Z_1 + \frac{\sum_{i=r+1}^{p} Z_i^2}{2} \right\} \prod_{i=r+1}^{p} f_i(X_i; \theta_i) \times
\]

\[
\times \int_{\mathbb{Z}_2} (2\pi)^{-q/2} |\Sigma_{22,1}|^{-1/2} \text{Exp} \left\{ -\frac{1}{2} (Z_2 - \Sigma_{21} \Sigma_{11}^{-1} Z_1)^T \Sigma_{22,1}^{-1} (Z_2 - \Sigma_{21} \Sigma_{11}^{-1} Z_1) \right\} d\mathbf{Z}_2
\]

The marginal distribution of \( X_1 \) is

\[
\mathbf{M} \left\{ \Sigma_{21} \Sigma_{11}^{-1} Z_1, \Sigma_{11}, \left\{ f_i(X_i; \theta_i) \right\}_{i=1,\ldots,r} \right\}
\]

because the integral in the second line of this expression is equal to one. \( \mathbf{Z}_2 \) has
multivariate normal distribution with a mean equal to $\Sigma_{21} \Sigma_{11}^{-1} Z_i$ and covariance matrix $\Sigma_{22,1}$ (conditional distribution of $f(Z_2 | Z_i)$).

3.2 Conditional distributions

Results in the last section show that the conditional distribution of $X_2 / X_1$ is:

$$M \left( \Sigma_{21} \Sigma_{11}^{-1} Z_i, \Sigma_{22,1}, \{ f_i(X_i, \theta_i) \}_{i=r+1, \ldots, p} \right), \text{ this is:}$$

$$h_{X_2/1} (X_2 / X_1) =$$

$$\left| \Sigma_{22,1} \right|^{-1/2} \exp \left[ -\frac{1}{2} (Z_2 - \Sigma_{21} \Sigma_{11}^{-1} Z_i)^\top \Sigma_{22,1}^{-1} (Z_2 - \Sigma_{21} \Sigma_{11}^{-1} Z_i) + \frac{\sum_{i=r+1}^p Z_i^2}{2} \right] \times$$

$$\prod_{i=r+1}^p f_i(X_i; \theta_i)$$

(20)

3.3 Parameter estimation

Although parameters of the GMF family could be estimated by the algorithm given by Karian et al (2000), it is advisable to obtain maximum likelihood (ML) estimators because of their well known properties. Marginal densities (which are GMF parameters) can be estimated by ML. By the invariance property of ML estimation (Cassela & Berger, 1990), the ML estimator of $\Sigma$ (which is a function of the marginal densities), is the correlation matrix of the normalized variables (10). Therefore the ML estimators of the GMF family can be obtained with the algorithm:

(i) Select the marginal distributions with a statistic method (Kolmogorov-Smirnov, Anderson-Darling, $\chi^2$, etc.).

(ii) Obtain the ML estimators of the marginal distributions.

(iii) Estimate $\Sigma$ as the correlation matrix of the normalized variables (10).
3.3 Some bivariate density contours

Density contours are displayed for selected bivariate distributions to show some possible distributional shapes provided by this family. In each pair of density contours, the figure on the left, presents the joint density with independent components ($\rho = 0$) while the figure on the right presents the joint density with a correlation coefficient equal to 0.60 between their components. Figure 1 shows bivariate densities with gamma marginals and Figure 2 shows bivariate densities with Weibull marginals.

Figure 1. Bivariate density with gamma marginals $G(2.5,10)$ in $X$ and $G(1.5, 8)$ in $Y$.

Figure 2. Bivariate density with Weibull marginals $W(2, 5)$ in $X$ and $W(3, 8)$ in $Y$. 
These figures, obtained with Mathematica (Wolfram, 1998), show that the GMF could be useful for expressing a linear relationship between variables with the great advantage that only the correlation coefficient is needed to express the dependence among each pair of variables. If the relationship is not linear, another copula may be more appropriate.

4. Application to a hydrological study

Runoff modeling and forecasting is of great importance because it helps to estimate water availability and planning diverse human activities like agriculture. In the particular case of the Yaqui river which irrigates around 200,000 ha in the Yaqui Valley in northwest México, monthly runoff can be divided in three non correlated periods: a) July, August and September (JAS), b) October and November, and c) December to June.

July’s monthly runoff is not autocorrelated and runoff estimates for this month can only be obtained from its probability distribution, Augusts’ runoff is correlated with July’s runoff, therefore a bivariate distribution can be fit and the runoff conditional distribution of August given July can be used to improve the estimates of Augusts’ runoff. Bivariate distributions from two Archimedean copulas (Gumbel-Hougaard and Cook-Johnson families), the generalized lambda distribution (Karian and Dudewics) and the GMF, were fit to runoff data consisting in the records of the last 48 years.

Applying the algorithm for ML estimation, a gamma distribution G[4.1025, 114.5286] was estimated for July’s runoff. For Augusts’ runoff the lognormal distribution LogN[6.5707, 0.49513] was chosen. The Gumbel-Hougaard [21] and the Cook-Johnson copulas [22], were estimated by the relationship between the Kendall’s coefficient between July’s and August’s runoff, and the parameter of the copula (Zhang and Singh, 2007; Nelsen, 2006). The parameter Ψ for the generalized lambda distribution [23] was
estimated with the procedure given in Karian and Dudewics (2000). For the GMF, the correlation coefficient (equal to 0.4715) was obtained from the transformed runoff variables (10).

Gumbel-Hougaard copula.

\[ H(X_1, X_2) = \exp \left\{ - \left[ (\log(F_1(X_1)))^\theta + (\log(F_2(X_2)))^\theta \right]^{1/\theta} \right\} \] (21)

where \( \theta \), is the parameter of the copula, related with the equation \( \tau = 1 - \theta^{-1} \), with \( \tau \), the Kendall’s coefficient of correlation.

Cook-Johnson copula.

\[ H(X_1, X_2) = \left[ (F_1(X_1))^{-\theta} + (F_2(X_2))^{-\theta} \right]^{-1/\theta} \] (22)

where \( \theta \), is the parameter of the copula, related with the equation \( \tau = \theta / (\theta + 2) \), with \( \tau \), the Kendall’s coefficient of correlation.

Generalized lambda (GL) copula.

\[ H(X_1, X_2) = \frac{S - \sqrt{S^2 - 4(\Psi - 1)^2 F_1(X_1) F_2(X_2)}}{2(\Psi - 1)} \] (23)

where \( S = 1 + (F_1(X_1) - F_2(X_2)) (\Psi - 1) \) and \( \Psi \) is the parameter of the copula, which was estimated with the procedure given in Karian and Dudewics (2000).

The empirical joint distribution of runoff variables was obtained (Zhang and Singh, 2007) with equation (24).

\[ H(x_{ij}, x_{j}) = P(X_1 \leq x_{ij}, X_2 \leq x_j) = \frac{\sum_{i=1}^n \sum_{j=1}^m N_{mn} - 0.44}{N + 0.12} \] (24)

where \( N \) is the sample size; \( N_{mn} \) is the number of \( (x_{ij}, x_j) \) counted as \( x_{ij} \leq x_i \) and \( x_{2j} \leq x_j \), \( i = 1, 2, \ldots, N \).
P-P plots obtained with the four fitted copulas show that their fit are similar. Copulas were compared with the root mean square error (RMSE), the Kolmogorov-Smirnov (KS) test and the AIC criterion, Table 1 shows these statistics.

<table>
<thead>
<tr>
<th>Copula</th>
<th>RMSE</th>
<th>Kolmogorov-Smirnov P-value</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gumbel-H</td>
<td>0.0291</td>
<td>0.8794</td>
<td>1350.89</td>
</tr>
<tr>
<td>Cook-J.</td>
<td>0.0372</td>
<td>0.5909</td>
<td>1351.16</td>
</tr>
<tr>
<td>GL</td>
<td>0.0336</td>
<td>0.6393</td>
<td>1350.26</td>
</tr>
<tr>
<td>GMF</td>
<td>0.0317</td>
<td>0.7829</td>
<td>1348.26</td>
</tr>
</tbody>
</table>

Table 1. Goodness of fit statistics for the fitted copulas.

The RMSE was obtained with the cumulative probabilities as:

\[
RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left[ H_S(x_{1i}, x_{2i}) - H_C(x_{1i}, x_{2i}) \right]^2}
\]

where subindex \( S \), in the CDF, \( H_S(x_{1i}, x_{2i}) \) stands for the empirical distribution (\( S=E \)) or the distribution given by the selected copula (\( S=C \)).

The AIC criterion:

\[
AIC = -2 \ln L(\hat{\theta}) + 2k
\]

where \( L(\hat{\theta}) \) is ML function for the model and \( k \), is the number of fitted parameters. The ML function was evaluated for the GMF, for the rest of the models; the parameter of the copula was estimated by means of the relationship between the copula parameter and the Kendall’s coefficient.

The best fitted distribution using the AIC criterion is the GMF, but the RMSE shows that the GMF and the Gumbel-Hougaard copula have the best fit. This example
shows that the GMF competes well with other copulas and gives a flexible family for fitting multivariate distributions with fixed marginal distributions.

References


