The Use of Copulas to Model Conditional Expectation for Multivariate Data

Käärik, Meelis; Selart, Anne; Käärik, Ene
University of Tartu, Institute of Mathematical Statistics
J. Liivi 2
50409 Tartu, Estonia
E-mail: Meelis.Kaarik@ut.ee; Anne.Selart@ut.ee; Ene.Kaarik@ut.ee

1. Introduction

Copula is one of the most useful tools for handling multivariate distributions with dependent components. Copulas can be considered as certain dependency functions for constructing multivariate distributions from their corresponding marginal distributions. Background information on copulas is covered in a number of papers starting from Nelsen (1998).

There are two main statistical advantages of modelling the dependence of multivariate data by copulas. Firstly, copulas allow us to use arbitrary marginals (that even need not to come from same distributional family). Secondly, the copula technique allows us to separate the modelling of marginals from modelling of the dependence structure. In our previous studies we focused on the Gaussian copula case, which is a natural starting point, in current research we expand the class of copulas, including t-copula, and also go beyond the elliptical class of copulas, examining few related skewed copulas (e.g. skew-normal) for modelling asymmetric data.

Although initially designed for the imputation problem of missing data, the setup allows wide implementation in many fields, including insurance (especially credibility models) and microarray data.

We focus on the case where each marginal distribution F_i is continuous and differentiable. If the copula C and marginals $F_1, \ldots, F_k$ are differentiable, then the joint density $f(x_1, \ldots, x_k)$ corresponding to the joint distribution function $F(x_1, \ldots, x_k)$ can be written by canonical representation as a product of the marginal densities and the copula density $f(x_1, \ldots, x_k) = f_1(x_1) \cdot \ldots \cdot f_k(x_k) \cdot c(F_1, \ldots, F_k)$, where $f_i(x_i)$ is the density corresponding to F_i and the copula density c is defined as derivative of the copula. Taking into account the univariate marginals and joint density defined by copula we can write the conditional density in the following form:

$$f(x_k|x_1, \ldots, x_{k-1}) = f_k(x_k)\frac{c(F_1, \ldots, F_k)}{c(F_1, \ldots, F_{k-1})} \ ,$$

where $c(F_1, \ldots, F_k)$ and $c(F_1, \ldots, F_{k-1})$ are corresponding copula densities.

We start from the Gaussian copula which is examined in Käärik and Käärik (2009, 2010) and introduce t-copula and their possible extentions like skew-normal copula and skew t-copula.

2. Framework

We consider multivariate correlated data in broader sense including repeated measurements over time (longitudinal data) or over space or clustered data. Incomplete data occur when some measurements are missing and so some part of data is absent. We are interested in finding the most probable values for the gaps in data using existing observations and following certain model based on conditional distribution. Similar setup is applicable to various general prediction problems where the aim is to predict the future value of an outcome variable based on the history and the values of correlated variables.

We use the idea of imputing missing value based on conditional distribution conditionally to

all observed variables. Considering symmetrical distributions the value that would be observed most likely (argmax of the conditional density function) is the conditional mean and we use the conditional mean as the imputed (or predicted) value.

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ be the random vector with correlated components Y_j. Consider data with n subjects from $\boldsymbol{Y}$ as the $n \times m$ matrix $\mathbf{Y} = (\mathrm{Y}_1, \ldots, \mathrm{Y}_m)$, $\mathrm{Y}_j = (\mathrm{y}_{1j}, \ldots, \mathrm{y}_{nj}), j = 1, \ldots, m$. Assume that we have k completely observed variables and $m - k$ partially observed variables. For simplicity we suppress the subscript index for individual, thus y_j is the value of data variable Y_j for subject i that has to be predicted or imputed.

We focus on imputing (predicting) the variable Y_{k+1} using the complete (observed) part $Y_1, \ldots, Y_k$. If necessary (particularly in imputation case), we can apply the same method iteratively from $k+1$ to m by algorithm given in (Käärik and Käärik, 2010).

Accordingly to the setup above we can partition the correlation matrix $\mathbf{R}_{k+1}$ (corresponding to $Y_1, \ldots, Y_{k+1}$) as follows

$$\mathbf{R}_{k+1} = \begin{pmatrix} \mathbf{R}_k & \mathbf{r} \\ \mathbf{r}^T & 1 \end{pmatrix},$$

where $\mathbf{R}_k$ is the correlation matrix of the observed part and $\mathbf{r} = (r_{1,k+1}, \ldots, r_{k,k+1})^T$ is the vector of correlations between the observed part and Y_{k+1}.

If the marginal distributions of $(Y_1, \ldots, Y_k)$ and Y_{k+1} are continuous and known and correlation matrix is estimated directly from data we can use copula model for specifying the joint and conditional distributions.

3. Conditional expectation model

Our main aim is to generalize the obtained results for Gaussian copula (including joint and conditional distribution functions and conditional expectations for different dependence structures and related prediction algorithms) to a broader class of copulas. Similar problems are also widely studied by Leong and Valdez (2005), who proved several results related to conditional distributions for different copulas. Although their main focus lies on insurance problems like claims prediction and credibility theory, the general setup for imputation problem is very close.

3.1. Gaussian copula approach

One of the most important examples of copulas is the normal or Gaussian copula. By definition, the k-variate Gaussian copula with k Gaussian marginals corresponds to the k-variate Gaussian distribution. Hence, the marginal distributions of k-variate normal copula are assumed to be continuous and can substantially differ from normal ones and can, in principle, be different.

Definition 1. *Let $\mathbf{R}_k$ be a symmetric, positive definite matrix with $diag(\mathbf{R}_k) = (1, 1, \ldots, 1)^T$ and Φ_k be the k-variate normal distribution function with correlation matrix $\mathbf{R}_k$. Then the multivariate Gaussian copula is defined as $C^N(u_1, \ldots, u_k; \mathbf{R}_k) = \Phi_k(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_k); \mathbf{R}_k)$, where $u_j \in (0, 1), j = 1, \ldots, k$, and Φ^{-1} is the inverse of the univariate standard normal distribution function.*

Using the Gaussian copula we obtain the conditional probability density function (see Käärik and Käärik, 2009)

$$(1) \quad f_{Z_{k+1}|Z_1,\ldots,Z_k}(z_{k+1}|z_1, \ldots, z_k; \mathbf{R}_{k+1}) = \frac{1}{\sigma_k} \varphi \left(\frac{z_{k+1} - \mathbf{r}^T \mathbf{R}_k^{-1} \mathbf{z}_k}{\sigma_k} \right),$$

where $Z_j = \Phi^{-1}[F_j(Y_j)],\ j = 1, \ldots, k+1$ are standard normal, $\mathbf{z}_k = (z_1, \ldots, z_k)^T$, $\sigma_k^2 = 1 - \mathbf{r}^T\mathbf{R}_k^{-1}\mathbf{r}$ and φ is univariate standard normal density.

As a result we have the (conditional) probability density function of a normal random variable with expectation $\mathbf{r}^T\mathbf{R}_k^{-1}\mathbf{z}_k$ and variance $1 - \mathbf{r}^T\mathbf{R}_k^{-1}\mathbf{r}$.

To use arbitrary marginals, we need to apply the following three-step procedure
(1) Use the normalizing transformation $Z_j = \Phi^{-1}(F_j(Y_j))$, $j = 1, \ldots, k+1$.
(2) Predict the value using conditional expectation from (1).
(3) Use the inverse transformation $Y_{k+1} = F_{k+1}^{-1}[\Phi(Z_{k+1})]$ to predict the value in initial scale.

3.2. Student's *t*-copula approach

A well known alternative to the multivariate normal distribution is the multivariate t-distribution which is underlying to the t-copula. The t-copula gives easily computable conditional distribution and is usable for example to generate credibility predictors (Frees and Wang, 2005).

Definition 2. *Let $\mathbf{R}_k$ be a symmetric, positive definite matrix with $diag(\mathbf{R}_k) = (1, 1, \ldots, 1)^T$ and Ψ_k be the k-variate Student's t-distribution function with correlation matrix $\mathbf{R}_k$. Then the Student's t-copula is defined as $C^t(u_1, \ldots, u_k; \nu, \mathbf{R}_k) = \Psi_k(\Psi^{-1}(u_1, \nu), \ldots, \Psi^{-1}(u_k, \nu); \nu, \mathbf{R}_k)$, where $u_j \in (0, 1), j = 1, \ldots, k$, and Ψ^{-1} is the inverse of the univariate standard t-distribution function and ν is the number of degrees of freedom.*

The main difference between the Student's and Gaussian copula lies in the probability of extreme events. A Gaussian copula has zero tail dependence, that means that the probability that variables are in their extremes is asymptotically zero unless linear correlation coefficient is equal to one, while the Student's t has symmetric, but nonzero tail dependence. Of course for moderate values of the correlation coefficient, the Student's copula with large number of degrees of freedom may be close to the Gaussian copula.

Analogously to Gaussian copula approach we first transform our original random variables $Y_1, \ldots, Y_{k+1}$ (which can have arbitrary distributions!) by "studentizing transformation", i.e. $T_j = \Psi^{-1}[F_j(Y_j)] \sim t(0, 1, \nu),\ j = 1, \ldots, k+1$, then find the formula for prediction for obtained t-distributed random variables, and finally transform the predicted values back to original scale via $Y_{k+1} = F_{k+1}^{-1}[\Psi(T_{k+1})]$.

For prediction we can phrase a result similar to the one for Gaussian copula approach:

Lemma 1. *Let $(T_1, \ldots, T_{k+1})$ be a "studentized" random vector, i.e. $T_j \sim t(0, 1, \nu)$. Then the conditional random variable $T_{k+1}|T_1, \ldots, T_k$ is also t-distributed, with $\nu + k$ degrees of freedom and*

$$E(T_{k+1}|T_1 = t_1, \ldots, T_k = t_k) = \mathbf{r}^T\mathbf{R}_k^{-1}\mathbf{t}_k, \tag{2}$$

$$Var(T_{k+1}|T_1 = t_1, \ldots, T_k = t_k) = \frac{\nu}{\nu + k - 2}\left(1 - \mathbf{r}^T\mathbf{R}_k^{-1}\mathbf{r}\right)\left(1 + \frac{1}{\nu}\mathbf{t}_k^T\mathbf{R}_k^{-1}\mathbf{t}_k\right). \tag{3}$$

This result is concordant with the results in (Kotz, Nadarajah, 2004), more detailed and thorough overview of the topic can be found in (Leong, Valdez, 2005).

The copula pertained to the multivariate generalized t-distribution is generalized t-copula (GT-copula) (Mendes and Arslan, 2006). Another generalization of t-copula using mixture constructions gives the skewed t-copula (Demarta and McNeil, 2005) both generalized copulas allow to model a wide variety of skew and heavy tailed datasets.

3.3. Skew-normal copula approach

Skew-normal distribution is an extension of the normal distribution. The symmetry of the normal distribution is distorted with one extra parameter called the shape parameter. The scalar

version of the skew-normal distribution was first introduced by Azzalini (1985) and generalized to the multivariate case in Azzalini & Dalla Valle (1996) and Azzalini & Capitanio (1999).

The skew-normal copula is associated with skew-normal distribution, the advantage is that copula allows any kind of marginals. The multivariate skew-normal distribution has skew-normal marginals and is often useful for fit data with 'normal-like' shape of the empirical distribution, but with lack of symmetry and has an extra parameter to regulate skewness.

There are different ways to parametrize the skew-normal distribution (see, e.g., Azzalini and Capitanio, 1999; Vernic, 2005), as in our approach the parameters of marginal distributions and their relation to multivariate distribution is of high importance, we choose a paramatrization by the marginal distribution parameters similar to Azzalini and Dalla Valle (1996).

Definition 3. *Let $\mathbf{R}_k$ be a symmetric, positive definite matrix with $diag(\mathbf{R}_k) = (1, 1, \ldots, 1)^T$. We say that a k-variate random variable $\boldsymbol{Z} = (Z_1, \ldots, Z_k)$ has skew-normal distribution with skewness parameter $\boldsymbol{\lambda}_k = (\lambda_1, \ldots, \lambda_k)$, $\boldsymbol{Z} \sim SN_k(\mathbf{R}_k, \boldsymbol{\lambda}_k)$, if its probability density function is given by*

$$f(\mathbf{z}_k) = 2\varphi_k(\mathbf{z}_k; \mathbf{R}_k)\Phi(\boldsymbol{\alpha}_k^T \mathbf{z}_k), \tag{4}$$

where

$$\boldsymbol{\alpha}_k = \frac{\Delta_k \Omega_k^{-1} \boldsymbol{\lambda}_k}{\sqrt{1 + \boldsymbol{\lambda}_k^T \Omega_k^{-1} \boldsymbol{\lambda}_k}}, \quad \Omega_k = \Delta_k \mathbf{R}_k \Delta_k - \boldsymbol{\lambda}_k \boldsymbol{\lambda}_k^T, \quad \Delta_k = \operatorname{diag}\left(\sqrt{1 + \lambda_1^2}, \ldots, \sqrt{1 + \lambda_k^2}\right) \tag{5}$$

and φ_k and Φ are the k-dimensional standard normal density and the univariate standard normal distribution function, respectively. Then the univariate marginals of $\boldsymbol{Z}$ also have univariate skew-normal distribution, $Z_j \sim SN_1(1, \lambda_j)$.

Definition 4. *Let H denote the distribution function corresponding to $SN_k(\mathbf{R}_k, \boldsymbol{\lambda}_k)$ and let $G_1, G_2, \ldots, G_k$ denote its marginals. The corresponding skew-normal copula is defined by*

$$C^{SN}(u_1, u_2, \ldots, u_k; \mathbf{R}_k, \boldsymbol{\lambda}_k) = H(G_1^{-1}(u_1), G_2^{-1}(u_2), \ldots, G_k^{-1}(u_k))$$

with respective copula density:

$$c^{SN}(u_1, u_2, \ldots, u_k; \mathbf{R}_k, \boldsymbol{\lambda}_k) = \frac{2^{1-k}}{\sqrt{|\mathbf{R}_k|}} \frac{\exp\{-\frac{1}{2}\mathbf{z}_k^T(\mathbf{R}_k^{-1} - I_k)\mathbf{z}_k\}\Phi(\boldsymbol{\alpha}_k^T \mathbf{z}_k)}{\prod_{i=1}^k \Phi(\lambda_i G_i^{-1}(u_i))},$$

where $\mathbf{z}_k = (G_1^{-1}(u_1), G_2^{-1}(u_2), \ldots, G_k^{-1}(u_k))^T$.

Our setup remains similar to the Gaussian copula and t-copula case: assume we have our original random variables $Y_1, \ldots, Y_{k+1}$ with distribution functions $F_1, \ldots, F_{k+1}$. We first transform the original variables to skew-normal random variables by $Z_j = G_j^{-1}[F_j(Y_j)]$, i.e. $Z_j \sim SN_1(1, \lambda_j)$, then apply the prediction formula (using conditional distribution), and finally transform the predicted values back to original scale via $Y_{k+1} = F_{k+1}^{-1}[G_{k+1}(Z_{k+1})]$.

To find the conditional density function for skew-normal distribution we can apply formula (1) for the standard normal density in (4), the result becomes:

$$f_{Z_{k+1}|Z_1,\ldots,Z_k}(z_{k+1}|z_1, \ldots, z_k; \mathbf{R}_{k+1}, \boldsymbol{\lambda}_{k+1}) = \frac{\exp\{-\frac{(z_{k+1} - \mathbf{r}^T \mathbf{R}_k^{-1} \mathbf{z}_k)^2}{2(1 - \mathbf{r}^T \mathbf{R}_k^{-1} \mathbf{r})}\}}{\sqrt{2\pi(1 - \mathbf{r}^T \mathbf{R}_k^{-1} \mathbf{r})}} \frac{\Phi(\boldsymbol{\alpha}_{k+1}^T \mathbf{z}_{k+1})}{\Phi(\boldsymbol{\alpha}_k^T \mathbf{z}_k)}, \tag{6}$$

where $\boldsymbol{\alpha}_{k+1}$ is calculated in a similar manner to $\boldsymbol{\alpha}_k$ in (5). It is important to notice that the components of vectors $\boldsymbol{\alpha}_k$ and $\boldsymbol{\alpha}_{k+1}$ may be completely different.

The expectation and variance of conditional distribution (6) for special case where $k = 1$ is given by Azzalini and Dalla Valle (1996). It can also be shown, that when introducing an additional parameter for skew-normal distribution, the conditional distribution also belongs to this new class of skew-normal distributions (see, e.g., Vernic, 2005).

4. Inference for different correlation structures

To implement conditional expectation we have to specify the structure of correlation matrix. The natural start is from simple correlation structure, depending on one parameter only. We introduce the following three structures of correlation matrix $\mathbf{R}_k$:
(1) The compound symmetry (CS) or the constant correlation structure, when the correlations between all variables are equal, $r_{ij} = \rho,\ i, j = 1, \ldots, k, i \neq j$.
(2) The first order autoregressive correlation structure (AR) or serial correlation as a traditional time-series representation, $r_{ij} = \rho^{|j-i|},\ \ i, j = 1, \ldots, k, i \neq j$.
(3) The 1-banded Toeplitz (BT) correlation structure, when only two adjoining variables are dependent, $r_{ij} = \rho,\ |i - j| = 1;\ \ r_{ij} = 0,\ |i - j| > 1,\ i, j = 1, \ldots, k$.
The banded Toeplitz structure is more general than the serial correlation model and could be adopted when we assume that there is some Markovian structure. A Toepliz matrix is constant along all diagonals parallel to the main diagonal.

The case of Gaussian copula according to above-named correlation structures is thoroughly examined in Käärik and Käärik (2009, 2010). Correspondingly to three-step procedure we start with normalizing transformation $Z_j = \Phi^{-1}[F_j(Y_j)],\ \ j = 1, \ldots, k+1$, then $z_1, \ldots, z_k$ are observed values for subject we have to predict (according row from data matrix). The results for different correlation structures $\mathbf{R}_k$ are proved for Gaussian copula case. As (by formulas (1) and (2)) the general forms of conditional expectation for Gaussian copula and Student's t-copula are similar, analogous result holds for Student's t-copula and we can formulate the following lemma.

Lemma 2. *Let $(T_1, \ldots, T_{k+1})$ be a random vector with Student's t-distributed marginals $T_j,\ j = 1, \ldots, k+1$ (with degrees of freedom ν), and correlation matrix $\mathbf{R}_{k+1}$. Then the following result holds:*

$$E(T_{k+1}|T_1 = t_1, \ldots, T_k = t_k) = \begin{cases} \frac{\rho}{1+(k-1)\rho} \sum_{i=1}^{k} t_i, & \text{if } \mathbf{R}_{k+1} \text{ has } CS \text{ structure,} \\ \rho t_k, & \text{if } \mathbf{R}_{k+1} \text{ has } AR \text{ structure,} \\ \frac{1}{|\mathbf{R}_k|} \sum_{j=1}^{k} (-1)^{k-j} |\mathbf{R}_{j-1}| \rho^{k-j+1} t_j, & \text{if } \mathbf{R}_{k+1} \text{ has } BT \text{ structure,} \end{cases}$$

where $|\mathbf{R}_j|,\ j = 1, \ldots, k$, is the determinant of correlation matrix of observed data and $|\mathbf{R}_0| = 1$.

5. Choice of copulas

Several copulas with varying shapes are available providing flexibility in modelling. The choice of different copulas naturally raises the question which copula model suits best to our problem. To identify the appropriate copula we can apply certain goodness of fit tests (besides the graphical methods). It is suggested in recent literature (see, e.g. Berg, 2009; Genest et al, 2009) that the suitability of copula is measured in the sense of the loss-function $S_n = \sum_{i=1}^{n} (C_{emp}(\hat{\mathbf{u}}_i) - C_\theta(\hat{\mathbf{u}}_i))^2$, where C_{emp} is the empirical copula based on data, C_θ is the proposed theoretical copula and $\hat{\mathbf{u}}_i$ are so-called pseudo-observations, $\hat{\mathbf{u}}_i = (\hat{u}_{i1}, \ldots \hat{u}_{ik})$, where $\hat{u}_{ij} = \frac{n\hat{F}_j(y_{ij})}{n+1}$, and $\hat{F}_j$ is the empirical distribution based on observations y_{ij}. The p-values corresponding to test statistic S_n can be found using certain parametric bootstrap method (for more details, see Genest and Rémillard, 2008; Kojadinovic and Yan, 2010). Additionally, a computationally more efficient approach based on central limit theorems is proposed by Kojadinovic and Yan (2010).

6. Final remarks

The prediction model using copulas and conditional expectation presented in this paper can be applied to several empirical prediction or imputation tasks, copulas are a widely used technique for modelling dependency in finance and insurance risk problems. Extension of the Gaussian or t-copula by adding a parameter to regulate skewness results in skew-normal or skew t-copula, which give the possibility to solve a broad problem of modelling skewed multivariate data in areas such as insurance (insurance risks or losses), environmental (air pollution and rainfall data) and biomedical (microarray data) science (see, e.g., Frees and Wang, 2005; Mendes and Arslan, 2006; Owzar et al, 2007).

As we can see from Lemma 2 for simple correlation structure the expectation algorithm is quite easy to use. This algorithm can also be extended to other correlation structures and copulas using similar approach.

Acknowledgments

This work is supported by Estonian Science Foundation grants No 7313 and No 8294.

REFERENCES

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* **12**, 171–178.

Azzalini, A., Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, 83 (4), 715–726.

Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *J. R. Statist. Soc., B*, **61**, 579–602.

Berg, D. (2009). Copula Goodness-of-Fit Testing: An Overview and Power Comparison. *The European Journal of Finance*, 15, 675–701.

Clemen, R.T., Reilly, T.(1999). Correlations and Copulas for Decision and Risk Analysis. *Management Science*, 45 (2), 208–224.

Demarta, S., McNeil, A.J. (2005). The t Copula and Related Copulas. *International Statistical Review*, 73 (1), 111–129.

Frees, E.W., Wang, P. (2005). Credibility using copulas. *North Americal Actuarial Journal*, 9 (2), 31–48.

Genest, C., Rémillard B. (2008). Validity of the Parametric Bootstrap for Goodness-of-Fit Testing in Semiparametric Models. Annales de l'Institut Henri Poincaré: Probabilités et Statistiques, 44, 1096–1127.

Genest, C., Rémillard, B., Beaudoin, D. (2009). Goodness-of-Fit Tests for Copulas: A Review and a Power Study. *Insurance: Mathematics and Economics*, 44, 199–213.

Käärik, M., Käärik, E. (2010). Imputation by Gaussian Copula Model with an Application to Incomplete Customer Satisfaction Data. *In: Proceedings of Compstat'2010*, Springer, Berlin/Heidelberg, 485–492.

Käärik, E., Käärik, M. (2009). Modelling Dropouts by Conditional Distribution, a Copula-Based Approach. *Journal of Statistical Planning and Inference*, 139 (11), 3830–3835.

Kotz, S., Nadarajah, S. (2004). Multivariate t distributions and their applications. Cambridge University Press, Cambridge/New York.

Leong, Y.K., Valdez, E.A. (2005). Claims prediction with Dependence using Copula Models. *Working manuscript*. Available at: http://www.math.uconn.edu/˜valdez/other-pubs.html

Mendes, B.V.M, Arslan, O. (2006). Multivariate Skew Distributions Based on the GT-Copula. *Brazilian Review of Econometrics*, 26 (2), 235–255.

Nelsen R.B. (1998). An Introduction to Copulas. Springer, New York.

Owzar, K., Jung, S.-H., Sen, P.K. (2007). A Copula Approach for Detecting Prognostic Genes Associated With Survival Outcome in Microarray Studies. *Biometrics*, 63, 1089–1098.

Vernic, R. (2005). On the multivariate Skew-Normal distribution and its scale mixtures. *An. St. Univ. Ovidius Constanta*, 13(2), 83-96.

Yan, J., Kojadinovic, I. (2010). Modeling Multivariate Distributions with Continuous Margins Using the copula R Package. *Journal of Statistical Software*, 34 (9), 1–20.