Stochastic Environmental Research and Risk Assessment (2023) 37:2619–2629
https://doi.org/10.1007/s00477-023-02408-1

ORIGINAL PAPER

Changing correlations: a flexible definition of non-Gaussian multivariate dependence

András Bárdossy[1]

Accepted: 13 February 2023 / Published online: 2 March 2023

Abstract
Dependencies between variables are often very complex, and may for high values, be different from that of the low values. As the normal distribution and the corresponding copula behave symmetrically for low and high values the frequent application of the normal copula for the description of the dependence may be inappropriate. In this contribution a new way of defining high dimensional multivariate distributions with changing correlations is presented. The method can also be used for a flexible definition of tail dependence. Examples of copulas with linear changing correlations illustrate the methodology. Parameter estimation methods and simulation procedures are discussed. A five dimensional example using groundwater quality data and another four dimensional one using air pollution data, are used to illustrate the methodology.

Keywords Multivariate copulas · Tail dependence · Asymmetry

1 Introduction

The dependencies between two or more variables, can be very complex. In the case of environmental variables physical, chemical and biological processes have a major influence on the variables of interest. These processes are often not explicitly understood and are non-linear. Statistical investigations of dependence are frequently based on correlations between the variables of interest. This however may lead to a sub-optimal recognition and description of dependencies.

Multivariate distributions relating variables are often considered to be normal or normal after data transformations. Popular transformations like the log transformation or Box–Cox transformations, transform the one dimensional marginals to normal, which does not imply that the bi- or multivariate marginal distributions also become normal. Relationships between variables are often more complex, even non-monotonic and frequently deviate from normal.

✉ András Bárdossy
bardossy@iws.uni-stuttgart.de

[1] Institute for Modelling Hydraulic and Environmental Systems, University of Stuttgart, Pfaffenwaldring 61, 70550 Stuttgart, Germany

An elegant possibility to describe complex relationships between continuous variables is possible by using copulas (Sklar 1959). Copulas enable the investigation of the dependence independently of the one dimensional marginal distributions. A large number of different theoretical copulas are described in different publications such as (Joe 1997) or (Nelsen 1999).

Copulas are frequently used in many different disciplines such as finance, economy, environmental studies or hydrology. In hydrology they are used for describing multivariate flood frequencies (Gräler et al. 2013) relationships between flood characteristics (Chen and Guo 2019) precipitation (Favre et al. 2018) or drought (Won et al. 2020) just to mention a few. In Bárdossy (2006) copulas were used for the spatial statistics for groundwater quality parameters. In Brunner et al. (2019) the Fisher copula was used to investigate the complex dependencies of flood occurrences.

Some of the well known copulas are derived from known multivariate distributions such as the Gaussian or the t-distribution. Another way to construct complex dependencies is to use vine-copulas (Czado and Nagler 2022). While the vine copulas offer a very flexible way to describe dependence, their construction and application for high dimensional cases is relatively complicated.

The purpose of this paper is to introduce a very flexible family of distributions with value dependent (asymmetrical) dependence and varying tail dependence. The paper is divided into 8 sections. After the introduction the definition of copulas is presented, and a new family of multivariate distributions with non-Gaussian copulas is introduced. Parameter estimation issues and simulation procedures are presented. Theoretical examples illustrate the flexibility of the methodology. Two data sets, one five dimensional on groundwater, the other four dimensional on air pollution are used as real-life examples to demonstrate the methodology. A short discussion and conclusions section completes the paper.

2 Methodology

2.1 Copulas

A copula C is defined as a multivariate distribution on the n dimensional unit cube:

$$C : [0,1]^n \rightarrow [0,1] \tag{1}$$

which has to have uniform marginals:

$$C(\mathbf{u}^{(\mathbf{i})}) = u_i \quad \text{if} \quad \mathbf{u}^{(\mathbf{i})} = (1,\ldots,1,u_i,1,\ldots,1)$$

Copulas are related to multivariate distributions through Sklar's theorem (Sklar 1959): Each multivariate distribution $F(t_1,\ldots,t_n)$ can be represented with the help of a copula:

$$F(t_1,\ldots,t_n) = C(F_{t_1}(t_1),\ldots,F_{t_n}(t_n)) \tag{2}$$

where $F_{t_i}(t)$ represents the i-th one dimensional marginal distribution of the multivariate distribution. If the marginal distributions are continuous then the copula C in (2) is unique. Hence, copulas can be regarded as the *pure expression* of the dependence without the influence of the marginal distributions.

Copulas can be defined explicitly using their density functions or another possibility is to define copulas using multivariate distributions by *inverting* (2). This means the copula is defined as:

$$C(u_1,\ldots,u_n) = F(F_{t_1}^{-1}(u_1),\ldots,F_{t_n}^{-1}(u_n)) \tag{3}$$

Many well known copulas such as the normal, the t-copula and the skew-normal copula are constructed using this method.

2.2 Distributions with value dependent correlations

In this paper a very flexible way of defining the multivariate distributions which define copulas with interesting non-Gaussian properties is presented. The basic idea behind the construction of the distribution is to gradually change the dependence structure depending on the values of the variables. As described in Guthke and Bárdossy (2012) one can obtain very similar spatial fields if one uses the same set of random numbers to generate them. A similar methodology was used in Bardossy and Pegram (2012) to exchange correlation structures of simulated precipitation. This idea combined with *continuity* can be used to define random variables with changing dependence structure. Formally:

Definition A matrix valued function

$$F : \mathbb{R} \rightarrow \mathbb{R}^{k\times k}$$

is called continuous if each element i, j of the matrix $F(\tau)_{i,j}$ is a continuous function

Definition Let $\Sigma(\tau)$ be a in τ continuous function of correlation matrices of dimension $k \times k$. Let X_τ be a k-dimensional normal random variable with the correlation matrix $\Sigma(\tau)$ and standard normal marginal distributions. In this case $X(\tau)$ can be coupled with the help of independent standard normal variables $U = (U_1,\ldots,U_k)$ in the form:

$$X(\tau) = \Sigma(\tau)^{\frac{1}{2}}U \tag{4}$$

If the correlation matrices $\Sigma(\tau)$ are continuous in τ than U defines a set of interdependent $X(\tau)$ random variables which are continuous in τ. A random vector $Z = (Z_1,\ldots,Z_k)$ is defined as:

$$Z_i = \min\{\Phi^{-1}(\tau)\ ;\ (\Sigma(\tau)^{\frac{1}{2}}U)_i = \Phi^{-1}(\tau)\} \tag{5}$$

Note that as $\Phi^{-1}(0) = -\infty < X(0)_i$ and $\Phi^{-1}(1) = +\infty > X(1)_i$ and $X(\tau)_i$ is by definition a continuous function for any i and U, the above definition leads to well defined Z_i-s.

This k-dimensional vector variable has, by this definition a dependence structure which is different for small and large values, resulting in a non-Gaussian multivariate distribution.

Figure 1 explains the construction. For a given vector of random numbers $(u_1,\ldots,u_k)$ and for two selected pairs of variables (i_1,i_2) and (j_1,j_2) the corresponding $(X_{i_1}(\tau),X_{i_2}(\tau))$ and $(X_{j_1}(\tau),X_{j_2}(\tau))$ are plotted as a function of $\Phi^{-1}(\tau)$. The (Z_{i_1},Z_{i_2}) and (Z_{j_1},Z_{j_2}) values correspond to the first intersection of the diagonal and the $(X_{i_1}(\tau),X_{i_2}(\tau))$

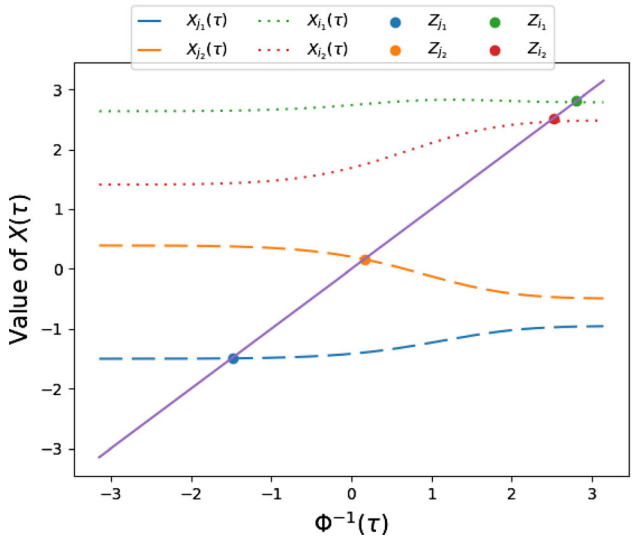

Fig. 1 Definition of the random variable Z: $X_i(\tau)$ functions for two given u_1, u_2 pairs and their intersection with the diagonal defining the values

and $(X_{j_1}(\tau), X_{j_2}(\tau))$ functions. The $(X_*(\tau)$ functions are continuous due to the continuity of the function of correlation matrices. The intersection with the diagonal represents the value assigned to Z_1 and Z_2. As the Figure shows the two different pairs get values from different $X(\tau)$s.

If $\Sigma(\tau) = \Sigma$ is constant for all τ values, then the resulting random variable is multivariate normal with the correlation matrix Σ.

3 Construction of continuous $\Sigma(\tau)$ functions

For the definition of the random variables in equation (5) a continuous correlation matrix valued function is required. The construction of such a function is a non trivial task as these matrices should not have negative eigenvalues to be valid correlation matrices.

3.1 Construction using the square roots of the matrices

In order to define such matrices one can use the fact that Γ is a covariance matrix if and only if it can be written as a product of a matrix Y and its transpose Y^T: $\Gamma = Y^T \cdot Y$. (Y is the square root of the covariance matrix.)

If $Y(\tau)$ is a continuous matrix valued function then $\Gamma(\tau) = Y(\tau)^T \cdot Y(\tau)$ is also a continuous matrix valued function. All these matrices are valid covariance matrices. By defining $\Sigma(\tau) = (\sigma(\tau)_{i,j})$ from $\Gamma(\tau) = (\gamma(\tau)_{i,j})$ as:

$$\sigma_{i,j}(\tau) = \frac{\gamma_{i,j}(\tau)}{\sqrt{\gamma_{i,i}(\tau)\gamma_{j,j}(\tau)}} \tag{6}$$

one obtains a continuous matrix valued function of correlation matrices due to the continuity of the transformation in (6).

The general construction uses infinitely many matrices. For practical use one has to find simplifications. The most simple construct is to use two parameters for each pair of variables - a starting and a final correlation, called two-correlations model in the subsequent text. Formally:

Let $\Sigma(0)$ and $\Sigma(1)$ two valid correlation matrices. If $\Sigma(i) = Y(i)^T \cdot Y(i)$ for $i = 0, 1$ then

$$Y^*(\tau) = \tau Y(1) + (1-\tau)Y(0) \tag{7}$$

can be used to define $\Gamma(\tau) = Y^*(\tau)^T Y^*(\tau)$ covariance matrices. The correlation matrices $\Sigma(\tau)$ corresponding to these covariance matrices form a continuous function of correlation matrices connecting $\Sigma(0)$ and $\Sigma(1)$. This model is called the two correlation linear model.

This construction can be generalized to produce a set of $m+1$ correlation matrices, the $m+1$ *correlation model*:

$$\Sigma(\tau_0), \Sigma(\tau_1), \ldots, \Sigma(\tau_{m-1}), \Sigma(\tau_m)$$

with

$$0 = \tau_0 < \tau_1 < \ldots < \tau_{m-1} < \tau_m = 1$$

.

Then:

$$Y^*(\tau) = \frac{\tau - \tau_j}{\tau_{j+1} - \tau_j} Y(\tau_{j+1}) + \left(1 - \frac{\tau - \tau_j}{\tau_{j+1} - \tau_j}\right) Y(\tau_j) \quad \text{if} \quad \tau_j \le \tau \le \tau_{j+1} \tag{8}$$

defines the sequence of covariance, and a subsequent continuous sequence of correlation matrices.

For any continuous function F of matrices the and continuous function $h : [0,1] \to [0,1]$, $F(h(t))$ is also a continuous function of matrices. This new matrix function defines a different random variable $Z^{(h)}$ with a different copula. The choice of the function $h(t)$ can change the shape of the corresponding copula even if all correlation matrices remain the same. The left and right hand derivatives of the transformation function h at 0 $h'(0^-)$ and at 1 $h'(1^+)$ are responsible for the tail behavior of the corresponding copula.

For those pairs where the starting $\Sigma(0)$ and or the ending $\Sigma(1)$ correlation matrix contain non diagonal values equal to 1, upper and/or lower tail dependence may occur. The value of the tail dependence can be anything between 0 and 1 by adjusting the speed of convergence of τ to 1 or 0 respectively. The corresponding proof is in the Appendix of this paper.

If all correlations are positive then another geometric construction could be used for the definition of the matrix function.

3.2 Construction using spatial statistics

In this case the correlation matrices are constructed using curves in an m dimensional space. All curves are parameterized with τ in [0, 1] and the correlation matrices are defined using a stationary spatial covariance function.

Let $(y_1(\tau), \ldots, (y_k(\tau))$ be such that $y_i(\tau)$ is a point in an $m < k$ dimensional space, and $y_i(\tau)$ is a line continuous in τ. If $C(h)$ is a valid continuous correlation function in space then:

$$\Sigma(\tau) = \begin{pmatrix} C(y_1(\tau) - y_1(\tau)) & \cdots & C(y_1(\tau) - y_k(\tau)) \\ \vdots & \ddots & \vdots \\ C(y_k(\tau) - y_1(\tau)) & \cdots & C(y_k(\tau) - y_k(\tau)) \end{pmatrix} \tag{9}$$

is a valid correlation matrix, and due to the continuity of the $y_i(\tau)$-s the corresponding matrices are continuous in τ.

4 Simulation

The simulation of a realization of the model is quite simple. One only has to know the correlation matrices for each τ and has to simulate k independent standard normal variables.

The exact simulation can be done using the following procedure:

1. Select the starting $\Sigma(0)$ and ending $\Sigma(1)$ correlation matrices.
2. Draw k independent normally distributed random numbers $(u_1, \ldots, u_k)$
3. Solve the linear equation (refeq:regu) for τ_i for each $i = 1, \ldots k$. Assign $x_i = \Phi^{-1}(\tau_i)$. The vector $(x_1, \ldots, x_k)$ is a simulated member.
4. Repeat steps 2–3 N times

This procedure is slightly slower than an approximate simulation using a discrete set of possible τ values. For the simulation of the linear model (7) the following step-by-step procedure can be used.

1. Select the starting $\Sigma(0)$ and ending $\Sigma(1)$ correlation matrices.
2. Calculate the square roots of the starting and end correlation matrices.
3. Select a set of τ values $0 = \tau(0) < \tau_1 < \ldots < \tau_m = 1$.
4. Calculate the in-between correlation matrices for each τ_i using linear interpolation of the square roots and renorming.
5. Draw k independent normally distributed random numbers $(u_1, \ldots, u_k)$
6. For each variable i find the simulated value is the x_i for which:

$$x_i = \Phi^{-1}(\tau_j) \quad \|\Phi^{-1}(\tau_j) - (\Sigma(\tau_j)^{\frac{1}{2}}(u_1, \ldots, u_k)_i\| \text{ minimal}$$

the vector $(x_1, \ldots, x_k)$ is the simulated member.

7. Repeat steps 5–6 N times

Both algorithms are very simple and large samples can be generated with little computational effort.

5 Examples

As a first example consider a bi-variate distribution with correlation matrices

$$\Sigma(0) = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix} \qquad \Sigma(1) = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix} \tag{10}$$

with normal marginals and linearly changing correlation matrices according to (7). The Pearson correlation of the two variables (with normal marginals) is 0.60.

For comparison the multivariate normal distribution with the same Pearson correlation (0.60) is considered. For both bi-variate distributions a sample of $N = 2000$ was generated. Figure 2 shows the corresponding empirical copulas. One can see the effect of changing correlations leading to weak dependence for low values and a strong dependence for high values.

Another example is a tri-variate distribution with two variables being independent and the third having a lower tail dependence with one of the variables and an upper tail dependence with the other one. The value of both tail dependencies can be arbitrary. This distribution can be constructed using the correlation matrices:

$$\Sigma(0) = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \qquad \Sigma(1) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} \tag{11}$$

Figure 3 shows the result of a simulation with $N = 3000$ points. One can see that variables Z_1 and Z_2 are independent, while Z_1 and Z_3 have some kind of upper tail dependence and Z_2 and Z_3 have some kind of lower tail dependence. Such copulas are difficult to construct with other methods. Note that the exact value of the tail

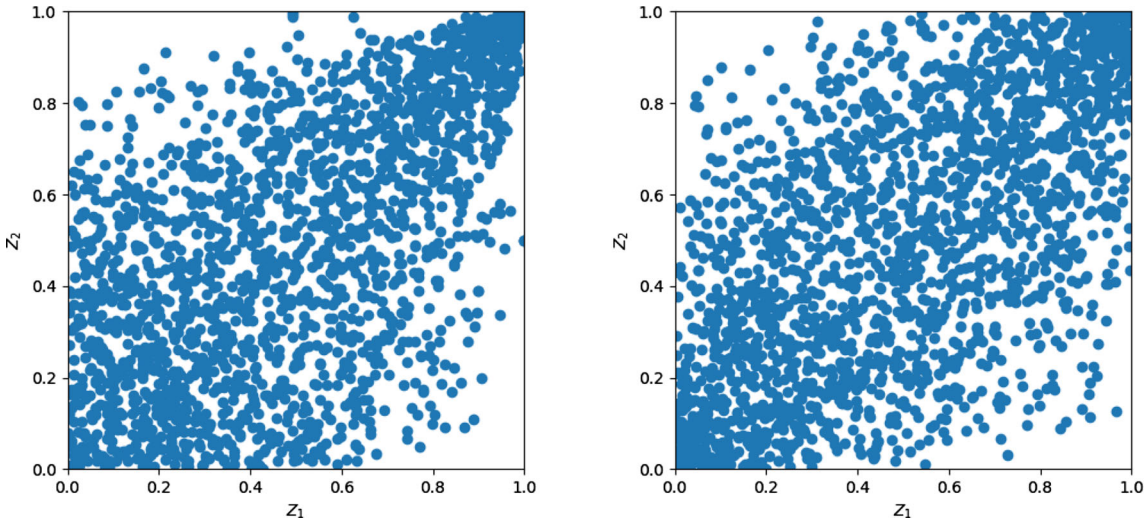

Fig. 2 Simulated empirical copulas corresponding to the two-correlations model ($\rho(0)=0.2 \rightarrow \rho(1)=0.9$) with Pearson correlation = 0.60 left, and corresponding to the normal copula with the same Pearson correlation (0.60)

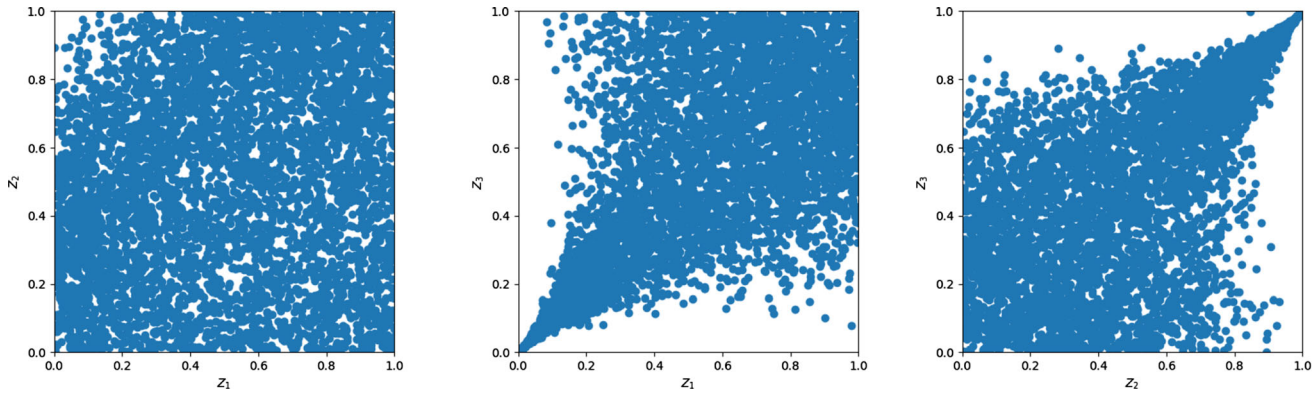

Fig. 3 Simulated empirical copulas corresponding to the 3 dimensional two-correlations model with $\Sigma(0) = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$ $\Sigma(1) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}$

dependence is determined by the speed of convergence to the matrices $\Sigma(0)$ and $\Sigma(1)$.

To show the high flexibility of the copulas obtained by this construction two three-correlations examples in the sense of (8) are shown. They correspond to the three correlation matrices:

$$\begin{aligned} \Sigma(0) &= \begin{pmatrix} 1 & 0.95 \\ 0.95 & 1 \end{pmatrix} \\ \Sigma\left(\frac{1}{2}\right) &= \begin{pmatrix} 1 & 0.65 \\ 0.65 & 1 \end{pmatrix} \\ \Sigma(1) &= \begin{pmatrix} 1 & 0.95 \\ 0.95 & 1 \end{pmatrix} \end{aligned} \tag{12}$$

and

$$\begin{aligned} \Sigma(0) &= \begin{pmatrix} 1 & 0.65 \\ 0.65 & 1 \end{pmatrix} \\ \Sigma\left(\frac{1}{2}\right) &= \begin{pmatrix} 1 & 0.95 \\ 0.95 & 1 \end{pmatrix} \\ \Sigma(1) &= \begin{pmatrix} 1 & 0.65 \\ 0.65 & 1 \end{pmatrix} \end{aligned} \tag{13}$$

The first shows a relationship weaker than a normal dependence for the medium values. For the second the relationship is reversed; the medium values show a stronger dependence than the extremes (Figure 4).

The construction of similar examples in higher dimensions is not difficult, one only has to be careful that the

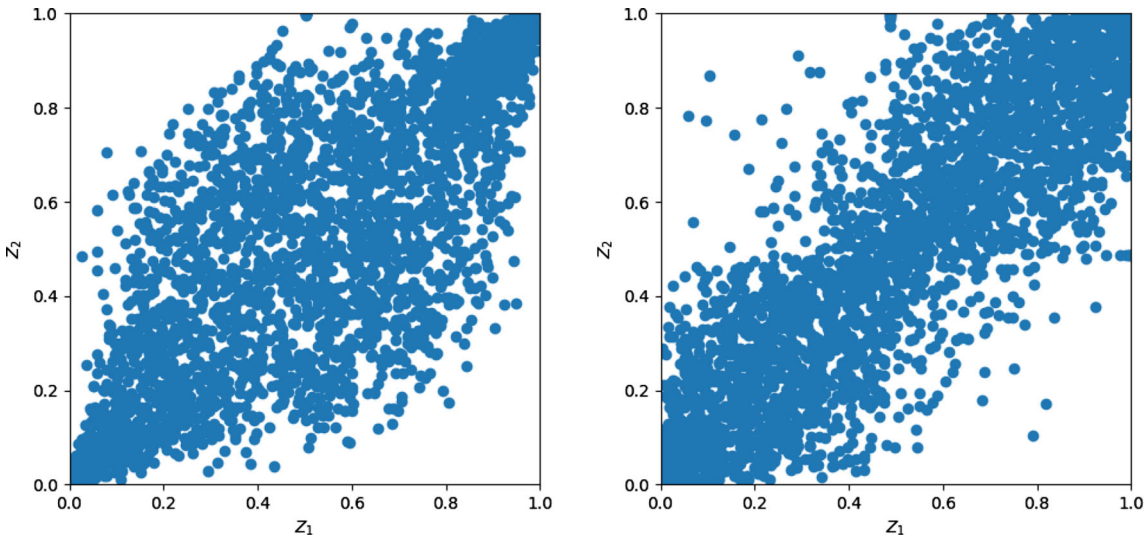

Fig. 4 Simulated empirical copulas corresponding to the three-correlations model with correlation matrices defined in (12) and (13)

correlation matrices corresponding to the *m breakpoints* are chosen so that they are positive semidefinite.

6 Parameter estimation

The parameters of the distributions depend on how the changing correlation structure of the random variable is chosen, as there is no general analytic expression for the density and distribution function of the corresponding random variable, allowing a straightforward maximum likelihood estimation. Instead specific procedures have to be used.

For the simplest two-correlations linear model two matrices $\Sigma(0)$ and $\Sigma(1)$ have to be estimated. This can be done in a pairwise manner as is done in the case of the multivariate normal distribution. The random variable inherits the pairwise definition of the dependence. For each pair of variables (i, j) the two correlation coefficients $\rho_{i,j}(0)$ and $\rho_{i,j}(1)$ have to be estimated. For this purpose two parameters describing properties of the dependence are needed. One possibility is to pick two classical dependence measures from Pearson's correlation of the normal score transformed data, Spearman's rank correlation or Kendall's τ. However these parameters all focus on the strength of dependence and not on value-related differences in dependencies. As an alternative, one may use the measure of dependence asymmetry as introduced for spatial statistics in Bárdossy (2008)

$$\alpha_{i,j} = \sum_{m=0}^{n} \Big((F_i(x_{m,i}) - \frac{1}{2})^2 (F_j(x_{m,j}) - \frac{1}{2}) + (F_i(x_{m,i}) - \frac{1}{2})(F_j(x_{m,j}) - \frac{1}{2})^2 \Big) \tag{14}$$

where n is the number of samples and F_i and F_j are the empirical distribution functions of variables i and j.

This measure describes the difference between the dependence of high values and that of the low values. It is well suited for variables with positive dependence, but for negative dependence it is better to first invert one of the variables.

The bi-variate two correlation linear model is the simplest model. Its parameters can be estimated using the Pearson correlation and the dependence asymmetry. The estimation can be done by using large simulated samples. The algorithm is as follows:

1. The observed data are transformed to normal using the normal score transformation.
2. The Pearson correlation ρ and the asymmetry α of the transformed data are calculated.
3. A pair of correlations $(\sigma(0), \sigma(1))$ is selected
4. A random sample of the size N of the two correlations model is simulated with the correlations $(\sigma(0), \sigma(1))$ denoted as $S_Z = \{(z_1^*(n), z_2^*(n)),\ n = 1, \ldots, N\}$.
5. The correlation and the asymmetry ρ^* and α^* of the simulated data are calculated.
6. The difference function:

   $$\Delta(\rho^*, \alpha^*) = (\rho - \rho^*)^2 + (\alpha - \alpha^*)^2$$

   is then minimized using an appropriate algorithm (for example steepest descent method). In the minimization at each new evaluation an new random sample of the size N is generated. The optimal $(\sigma(0), \sigma(1))$ is taken as parameters of the model.

As the simulation of the two correlations model is very simple and fast very large sample sizes N can be generated to assure that the parameters become stable. The Chebysev

inequality can be used to estimate the appropriate sample size.

As the above procedure is computationally intensive a numerical approximation of the function

$$G(\sigma(0), \sigma(1)) \rightarrow (\rho, \alpha)$$

can be done using a regular grid of $(\sigma(0), \sigma(1))$ values via simulation. The obtained table can then be used to estimate the two parameters of the two correlations model.

Figure 5 shows the Pearson correlation and the asymmetry as a function of the lower $(\sigma(0))$ and the upper $(\sigma(1))$ correlation for the two-correlation linear model. The two measures are kind of orthogonal allowing a simple estimation of $\rho(0)$ and $\rho(1)$ from the observed Pearson correlation and asymmetry.

For the multivariate case the parameters of the two correlations model are estimated in a pairwise manner as in the case of the multivariate normal distribution.

An interesting case is where correlations change their sign - for example low values show a negative correlation and high values are positively correlated. The overall Pearson correlation of such variables is usually close to zero, but the dependence is present and can be detected by other measures such as entropy.

Note that the number of parameters for the simplest two correlations linear model is only twice as much as for the multivariate normal distribution.

For the estimation of the parameters of a more complex dependencies additional measures have to be used. These could correspond to higher moments. As an alternative one may also use a set of indicator correlations. These are defined for different thresholds $0 < \theta < 1$:

$$I_\theta(Z_k) = \begin{cases} 1 & \text{if } F_k(Z_k) > \theta \\ 0 & \text{else} \end{cases} \tag{15}$$

The indicator correlations

$$\rho_I(\theta)_{i,j} = \text{Corr } (I_\theta(Z_i), I_\theta(Z_j)) \tag{16}$$

considered as a function of θ can also be used to to see if a dependence is symmetrical or not.

For copulas like the Gaussian or the t these functions are symmetrical around 0.5, the indicator correlations $\rho_I(\theta)$ and $\rho_I((1-\theta)$ should be equal.

The parameter estimation of the more complex models is also based on large simulated random samples. In this case the difference function of the form:

$$\Delta(\psi) = (\rho - \rho^*)^2 + (\alpha - \alpha^*)^2 + \int_0^1 (\rho_I(\theta) - \rho_I^*(\theta))^2 \, d\theta$$

should be minimized.

For the $m+1$ correlation model for $m > 1$ the breakpoints τ_i also have to be estimated. This changes the parameter estimation problem - as a pairwise estimation is not possible due to the common breakpoints. In order to simplify this problem it is reasonable to select for a given m the $\tau_j = \frac{j}{m}$ for $j = 0, \ldots, m$.

Note that as the third example in Sect. 5 defines random variables symmetrical dependence with respect to the values, thus one cannot use the asymmetry measure defined in (14) for the estimation of the parameters. Instead one could use indicator correlations for the parameter estimation.

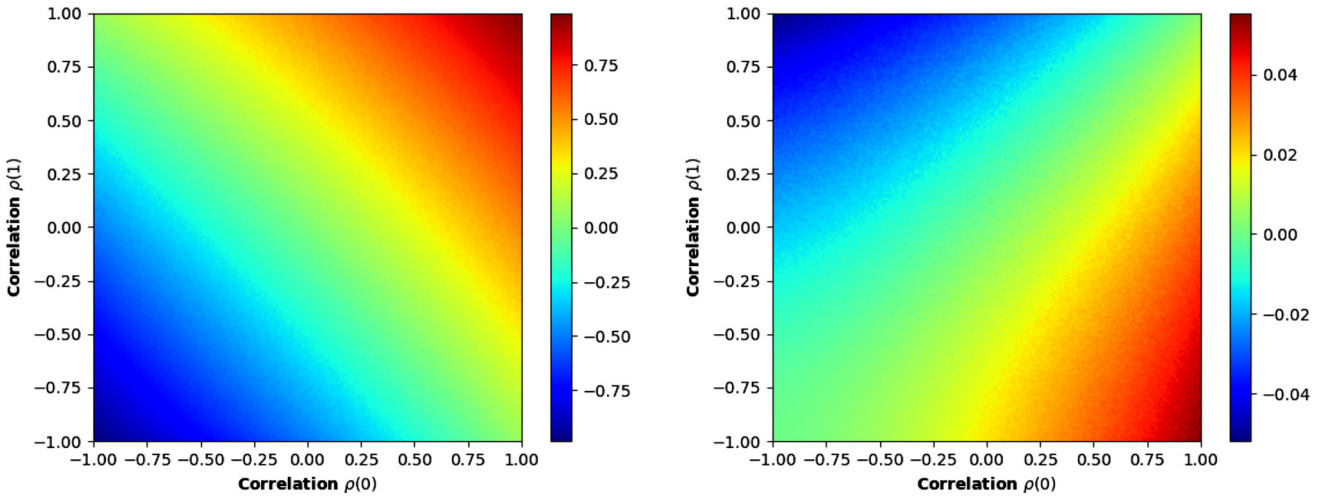

Fig. 5 Pearson correlation (left) and asymmetry (right) as a function of the lower ρ(0) and the upper ρ(1) correlation for the two-correlations model

7 Application

As an example groundwater quality data from the state of Baden-Württemberg in Germany are considered. In the framework of regular groundwater observation samples of more than 2300 wells were collected. These were analysed for their chemical composition. The parameters selected are chloride, nitrate, pH, sulfate and oxygen. The basic statistics of the data are listed in Table 1. The data are highly skewed, with the exception of pH, which has a small negative skew.

Due to the high skew the Pearson correlation of the data is strongly influenced by the large values of the data, and thus it is an imperfect description of the strength of the pairwise dependence. In order to investigate the interdependence between the different parameters the data were transformed to normal using the normal score transformation. As a next step the Pearson correlations (Table 2) and the dependence asymmetry according to equation (14) were calculated for each pair of parameters. In order to investigate the Gaussianity of the dependence, the asymmetry of the Gaussian distribution was calculated for simulated samples with the same size and Pearson correlations $N = 1000$ times. The corresponding asymmetries were calculated and the 95 % confidence interval was identified. For all 10 pairs of parameters the asymmetry was outside the confidence interval of the Gaussian, with the same correlation, indicating that the dependence is not normal.

A two-correlations linear model was fitted to the data using the Pearson correlations and the asymmetries as parameters. The two correlation matrices $\Sigma(0)$ and $\Sigma(1)$ were estimated in a pairwise manner (Table 3).

It is interesting to observe that the Pearson correlation between pH and Sulfate is very low, while the two-correlation model shows a relationship with changing sign. The lower correlation (corresponding to $\tau = 0$) is positive 0.70 while the upper corresponding to $\tau = 1$ is negative -0.52. This changing relationship cannot be captured by the normal copula (and also not by any other commonly used copulas). The indicator correlations shown on Fig. 6 confirms the changing relationship between the two variables, which is reasonably well captured by the two-correlations model, and are not captured by the normal model. This leads to a loss of information when applying the normal copula based model. A three correlation model was also fitted to the data, such that only the relationship between pH and Sulfate was altered. The three correlations were assessed using the indicator correlations. The new relationship is now $\mathbf{0.85} \rightarrow \mathbf{0.00} \rightarrow -0.3$ (corresponding to τ_i =0, 0.5 and 1). Note that the correlation for the low values corresponding to $\tau = 0$ increased, but decreases faster. Figure 6 shows the improvement of the fit to the indicator correlations.

Table 1 Basic statistics of the observed groundwater quality data

	Chloride mg/l	Nitrate mg/l	pH [-]	Oxygen mg/l	Sulfate mg/l
Mean	32.37	31.97	7.34	8.03	67.07
Standard deviation	41.10	27.98	0.41	3.00	120.68
Skewness	7.11	2.30	2.28	− 0.53	6.93
Minimum	0.60	0.20	5.10	0.20	0.70
Maximum	759.00	265.00	13.10	16.70	1628.00
Sample size	2537	2537	2537	2537	2537

Table 2 Pearson correlations of the normal score transformed observed groundwater quality data

	Chloride	Nitrate	pH	Oxygen	Sulfate
Chloride	1.00	0.45	0.04	0.35	0.63
Nitrate	0.45	1.00	0.14	− 0.12	0.37
pH	0.04	0.14	1.00	− 0.16	0.12
Oxygen	0.35	− 0.12	− 0.16	1.00	0.35
Sulfate	0.63	0.37	0.12	0.35	1.00

A comparison of the two dimensional marginals of the simulated data with the observed ones shows that in all cases the Kolmogorov distance between the simulated two-correlations model and the observations, is lower than for the normal copula based simulations.

For the second example, air quality measurements taken near Zurich-Schimmelstrasse in Switzerland were used. Four parameters NO_x, SO_2, NO_2 and PM10 were selected. These data are publicly available for the time period of 2011–2021 on the internet under opendata.swiss. Basic statistics of the data are listed in Table 4. The Pearson correlations of the normal score transformed data are given in Table 5. The correlations are higher than those of the groundwater example.

The matrices of the two-correlations model were fitted to the data by using the Pearson correlation and the asymmetry as measures using the algorithm described in Sect. 6. The fit was done numerically by using large simulated samples for the estimation of the parameters. The pairwise calculated asymmetries showed that out of the 6 pairs, 4 are significantly different from the normal. The theoretical model is in this case contains two pairs for which $\sigma_{i,j}(0) = \sigma_{i,j}(1)$. In order to investigate the appropriateness of the model, $N = 1000$ simulations of the 4 dimensional distributions were considered using the two-correlation linear model with parameters listed in Table 6.

Table 3 Changing correlations of the normal score transformed observed groundwater quality data with the two-correlations model. ($\rho_{ij}(0) \rightarrow \rho_{ij}(1)$)

	Chloride	Nitrate	pH	Oxygen	Sulfate
Chloride	1.00	**0.68 → 0.22**	**0.58 → -0.54**	**0.38 → 0.36**	**0.84 → 0.40**
Nitrate	**0.68 → 0.22**	1.00	**0.52 → -0.26**	**0.14 → -0.38**	**0.56 → 0.16**
pH	**0.58 → -0.54**	**0.52 → -0.26**	1.00	**0.08 → -0.42**	**0.70 → -0.52**
Oxygen	**0.38 → 0.36**	**0.14 → -0.38**	**0.08 → -0.42**	1.00	**0.38 → 0.36**
Sulfate	**0.84 → 0.40**	**0.56 → 0.16**	**0.70 → -0.52**	**0.38 → 0.36**	1.00

Pairs with non-normal dependence at 95 % significancance level are in boldface

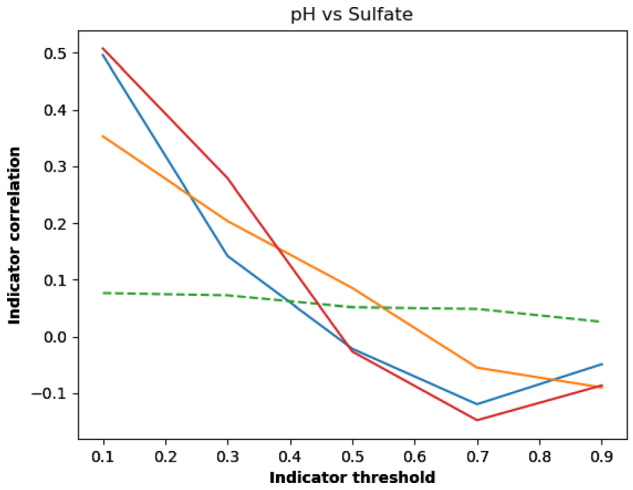

Fig. 6 Indicator correlations for pH and Sulfate - blue = observed, orange = two-correlations model, red = three-correlations model and green dashed = normal model

Table 4 Basic statistics of the observed air quality data

measurement unit	NO_x ppb	SO_2 $\mu g/m^3$	NO_2 $\mu g/m^3$	PM10 $\mu g/m^3$
Mean	39.65	18.61	43.92	28.93
Standard deviation	15.12	10.84	27.44	26.18
Skewness	0.56	1.77	1.99	2.64
Minimum	6.17	1.94	4.87	1.77
Maximum	109.44	109.14	272.87	271.14
Sample size	3903	3903	3903	3903

Table 5 Pearson correlations of the normal score transformed air quality data

	NO_x	SO_2	NO_2	PM10
NO_x	1.00	0.70	0.93	0.85
SO_2	0.70	1.00	0.66	0.59
NO_2	0.93	0.66	1.00	0.98
PM10	0.85	0.59	0.98	1.00

The same number of simulations were carried out for the normal copula case. As an example Fig. 7 shows the empirical copula a simulated normal and a simulated changing correlation copula. The sample sizes are in all cases the same. As one can see, the strong dependence of the low values is not captured by the normal simulation, but well captured by the two-correlations model. A comparison of the two dimensional marginals shows that the two-correlation linear model captures the asymmetrical dependence well. Note that in contrast to the previous example in this case all correlations are positive.

Deviations from the Gaussian dependence are very frequent, and both examples include cases where the high values have stronger dependence than the low ones and the reverse case too.

8 Discussion and conclusions

In this paper a method to construct multivariate non-Gaussian distributions was presented. The construction is very general and special cases with 2 or more parameters for the description of dependence of the pairs can be used.

The copulas obtained via these distributions are not only useful for monotonic dependence but can also represent dependencies with changing character, for example negative dependence for small values and positive dependence for high values.

Copulas were most frequently used for the investigation and description of the dependence of extremes. However, the dependence of the variables might be applied to non-extreme values and also deviations from the normal.

The copulas defined in this paper can describe arbitrary upper and lower tail dependence, and can also be used for the description of asymmetrical dependence even without focusing on the extremes.

A disadvantage of this construction is that the distribution functions do not have a general closed analytical form. This makes the estimation of the parameters difficult. The distributions defined using this construction may have many parameters, but due to the increasing data volumes of

Table 6 Changing correlations of the normal score transformed air quality data with the two-correlations model ($\rho_{ij}(0) \rightarrow \rho_{ij}(1)$)

	NO_x	SO_2	NO_2	PM10
NO_x	1.00	0.76 → 0.63	**0.99 → 0.86**	**0.95 → 0.77**
SO_2	0.76 → 0.63	1.00	0.69 → **0.63**	**0.54 → 0.65**
NO_2	**0.99 → 0.86**	0.69 → **0.63**	1.00	**0.99 → 0.98**
PM10	**0.95 → 0.77**	**0.54 → 0.65**	**0.99 → 0.98**	1.00

Pairs with non-normal dependence at 95 % significancance level are in boldface

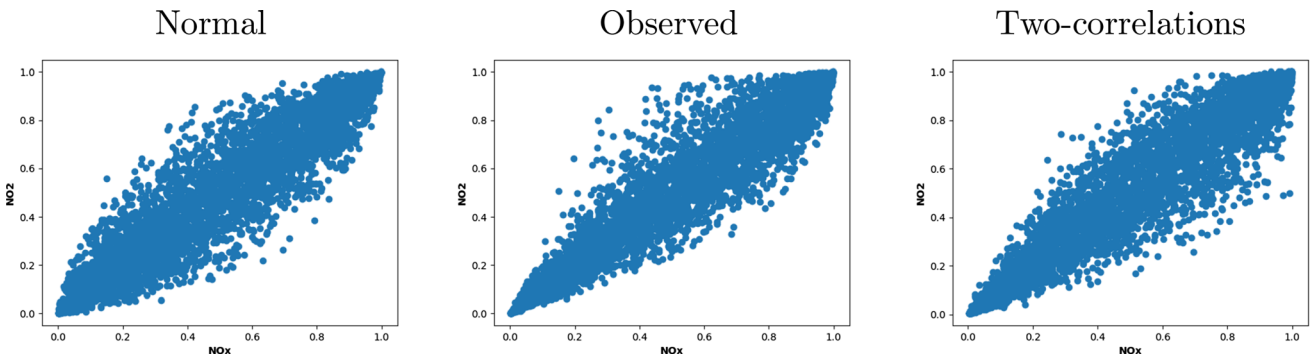

Fig. 7 Empirical and simulated copulas for the air quality parameters NO_x and NO_2 for Zürich: left simulated using normal bivariate copula, middle observed, right simulated using the two-correlations model

environmental variables may be well used for the investigation of big data.

The development of parameter estimation methods for more complex models requires further research.

The presented construction yields bi-variate marginal copulas which are symmetrical with respect the main axis addedof the copula. Non-symmetrical dependence can be achieved by using the same idea with other multivariate distributions, such as the skew-normal distribution. In this case both the parameters of the correlation matrix as the λ parameter defining the skewness can be varied in the same way as done in equation (5).

The use of these copulas as an alternative for multivariate linear regression is also possible, but goes beyond the scope of this contribution.

The methodology can be extended to time series and spatial random fields.

Appendix

Proposition For any $0 \leq a \leq 1$ there is a bi-variate construction such that the upper tail dependence is exactly a

Proof Let $\Phi_2(z_1, z_2, \rho)$ be the bi-variate normal distribution function with standard normal marginals and ρ correlation. For any $0 \leq a < 1$ and τ there is a correlation ρ^* such that the exceedence probability

$$\frac{1 - 2\tau + \Phi_2(\Phi^{-1}(\tau), \Phi^{-1}(\tau), \rho^*)}{1 - \tau} = a$$

As for $\rho \rightarrow 1 > a$ the left hand side converges to 1 and for $\rho = 0$ the left hand side is 0 by the continuity there must be a ρ^* which fulfills the equation. Defining $\rho(\tau) = \rho^*$ leads to the construction. For $a = 1$ a function converging to 1 has to be selected at the right hand side. The above construction can also be used to obtain variables with a lower tail dependence a. This way a random variable with a lower tail dependence a and an upper tail dependence $b \neq a$ can be constructed.

Acknowledgements Research leading to this paper has partly been supported by the German Research Foundation (DFG), by funding of the research group FOR 2416"Space-Time Dynamics of Extreme Floods (SPATE)".The author thanks the anonymous reviewer and G. Pegram for their useful comments, and dedicates this paper to his memory.

Author Contributions AB did everything.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors declare no competing interests

References

Bárdossy A (2006) Copula-based geostatistical models for groundwater quality parameters. Water Resour Res. https://doi.org/10.1029/2005WR004754

Bárdossy A, Li J (2008) Geostatistical interpolation using copulas. Water Resour Res. https://doi.org/10.1029/2007WR006115

Bardossy A, Pegram G (2012) Multiscale spatial recorrelation of rcm precipitation to produce unbiased climate change scenarios over large areas and small. Water Resour Res. https://doi.org/10.1029/2011WR011524

Brunner MI, Furrer R, Favre A-C (2019) Modeling the spatial dependence of floods using the fisher copula. Hydrol Earth Syst Sci 23(1):107–124. https://doi.org/10.5194/hess-23-107-2019

Chen L, Guo S (2019) Copula-based flood frequency analysis. Copulas Appl Hydrol Water Resour. https://doi.org/10.1007/978-981-13-0574-0_3

Czado C, Nagler T (2022) Vine copula based modeling. Annu Rev Stat Appl 9:453–477. https://doi.org/10.1146/annurev-statistics-040220-101153

Favre A-C, Quessy J-F, Toupin M-H (2018) The new family of fisher copulas to model upper tail dependence and radial asymmetry: Properties and application to high-dimensional rainfall data. Environmetrics 29(3):e2494

Gräler B, van den Berg MJ, Vandenberghe S, Petroselli A, Grimaldi S, De Baets B, Verhoest NEC (2013) Multivariate return periods in hydrology: a critical and practical review focusing on synthetic design hydrograph estimation. Hydrol Earth Syst Sci 17(4):1281–1296. https://doi.org/10.5194/hess-17-1281-2013

Guthke P, Bárdossy A (2012) Reducing the number of MC runs with antithetic and common random fields. Adv Water Resour 43:1–13. https://doi.org/10.1016/j.advwatres.2012.03.014

Joe H (1997) Multivariate models and dependence concepts. Chapman Hall, Boca Raton

Nelsen RB (1999) An introduction to copulas. Springer, New York

Sklar A (1959) Fonctions de répartition à n dimensions et leurs marges. Publ Inst Stat Paris 8:229–231

Won J, Choi J, Lee O, Kim S (2020) Copula-based joint drought index using SPI and EDDI and its application to climate change. Sci Total Environ 744:140701