Journal of Statistical Planning and Inference 139 (2009) 3942 – 3946

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

On multivariate Gaussian copulas

Ivan Žežula*

P.J. Šafárik University, Košice, Slovakia

ARTICLE INFO

Available online 29 May 2009

MSC:
62H20
62E17

Keywords:
Correlation
Copulas
Kendall
Spearman

ABSTRACT

Gaussian copulas are handy tool in many applications. However, when dimension of data is large, there are too many parameters to estimate. Use of special variance structure can facilitate the task. In many cases, especially when different data types are used, Pearson correlation is not a suitable measure of dependence. We study the properties of Kendall and Spearman correlation coefficients—which have better properties and are invariant under monotone transformations—used at the place of Pearson coefficients. Spearman correlation coefficient appears to be more suitable for use in such complex applications.

1. Introduction

Let us denote standard (univariate) normal density by $\varphi(x)$, standard normal cumulative distribution function by $\Phi(x)$, and corresponding quantile function by $\Phi^{-1}(x)$. Then, general normal distribution has a density of the form $f_{(1)}(x)=(1/\sigma)\varphi(u)$ and cdf $F_{(1)}(x)=\Phi(u)$, where $u=(x-\mu)/\sigma$. General p-variate normal density can be expressed as

$$f_{(p)}(x)=\frac{1}{(2\pi)^{p/2}\prod_{i=1}^{p}\sigma_i|R|^{1/2}}\,\exp\left\{-\frac{1}{2}u'R^{-1}u\right\}$$

where $u=(u_1,\dots,u_p)'$, $u_i=(x_i-\mu_i)/\sigma_i$, and R is a correlation matrix.

For any multivariate absolutely continuous distribution, with cdf F and marginal cdf's F_i, *copula* C is such distribution function on $\langle 0;1\rangle^p$ (with uniform one-dimensional marginals) that it holds

$$F(x_1,\dots,x_p)=C(F_1(x_1),\dots,F_p(x_p)).$$

Let f be the corresponding joint density and f_i, $i=1,\dots,p$, the marginal densities. *Copula density* c is defined by

$$c=\frac{\partial^p C}{\partial F_1\dots\partial F_p};$$

then, joint density can be expressed as

$$f(x)=c(F_1(x_1),\dots,F_p(x_p))\prod_{i=1}^{p}f_i(x_i).$$

* Tel.: +421 55 2346174; fax: +421 55 6227034.
E-mail address: ivan.zezula@upjs.sk.

0378-3758/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2009.05.039

Since multivariate normal density can be re-written in the form

$$f_{(p)}(x) = \frac{1}{|R|^{1/2}} \exp\left\{-\frac{1}{2}u'(R^{-1}-I)u\right\}\prod_{i=1}^{p}\frac{1}{\sigma_i}\varphi(u_i),$$

where $u_i = \Phi^{-1}(F_i(x_i))$, we can easily see that the density of a Gaussian copula is

$$c(x) = \frac{1}{|R|^{1/2}} \exp\left\{-\frac{1}{2}u'(R^{-1}-I)u\right\},$$

where F_i in $u_i = \Phi^{-1}(F_i(x_i))$ is an arbitrary continuous distribution function.

Gaussian copulas are handy tools in many situations, since they allow any marginal distribution and any p.d. correlation matrix. Some possible ways of application are described e.g. in Clemen and Reilly (1999) and Adermann and Pihlak (2005).

However, Gaussian copulas consider only pairwise dependence between individual components of a random variable, which does not encompass whole depth of possible dependence structures.

2. Special structures

There are other problems, which are connected with the practical use of Gaussian copulas:

- when p is large, R can be difficult to estimate, there are too many parameters;
- Gaussian densities are parameterized using Pearson correlation coefficients which are not invariant under monotone transformations of original variables, but such transformations are often required;
- Pearson ρ is not appropriate measure of dependence in many situations (heavy tails, non-elliptical distributions, marginals from different families), but usually these are the situations where we want to use copula modeling.

The problem of too many parameters can sometimes be avoided. In fact, many times we meet simpler correlation structures, which can be used for the construction of simple multivariate Gaussian copulas with few parameters to estimate. Most important examples of these are

1. uniform correlation structure (intra-class correlation coefficient model);
2. serial (1st order autoregressive) correlation structure.

2.1. Uniform correlation structure

This model is characterized by correlation matrix

$$R = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} = (1-\rho)I_p + \rho\mathbf{11}',$$

where $\rho \in \left\langle \frac{-1}{p-1}; 1\right\rangle$.

It is easy to see that if $\rho \neq 1$ and $\rho \neq -1/(p-1)$ then R^{-1} exists and

$$R^{-1} = \frac{1}{1-\rho}\left(I_p - \frac{\rho}{1+(p-1)\rho}\mathbf{11}'\right).$$

Since $|R| = [1+(p-1)\rho](1-\rho)^{p-1}$, we get

$$c(x) = \frac{1}{[1+(p-1)\rho]^{1/2}(1-\rho)^{(p-1)/2}} \exp\left\{\frac{-\rho}{2(1-\rho)}u'\left(I_p - \frac{1}{1+(p-1)\rho}\mathbf{11}'\right)u\right\}.$$

This can be written componentwise

$$c(x) = \frac{1}{\sqrt{[1+(p-1)\rho](1-\rho)^{p-1}}} \exp\left\{\frac{-\rho}{2(1-\rho)[1+(p-1)\rho]}\left((p-1)\rho\sum_{i=1}^{p}u_i^2 - 2\sum\sum_{i<j}u_iu_j\right)\right\}.$$

It is in accordance with Kelly and Krzysztofowicz (1997) for $p=2$.

2.2. Serial correlation structure

In this model we have

$$R = \begin{pmatrix} 1 & \rho & \cdots & \rho^{p-1} \\ \rho & 1 & \cdots & \rho^{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \cdots & 1 \end{pmatrix} = I_p + \sum_{i=1}^{p-1} \rho^i (C_1^i + C_1^{i\,\prime}),$$

where $C_1 = \begin{pmatrix} 0_{p-1} & I_{p-1} \\ 0 & 0'_{p-1} \end{pmatrix}$ and $\rho \in \langle -1; 1 \rangle$.

If $\rho \neq \pm 1$ then R^{-1} exists and it holds

$$R^{-1} = \frac{1}{1-\rho^2}[(1+\rho^2)I_p - \rho^2(e_1 e_1' + e_p e_p') - \rho(C_1 + C_1')],$$

where e_i contains 1 at i-th place and 0 at all other places.

Since $|R| = (1-\rho^2)^{p-1}$, we get

$$c(x) = \frac{1}{(1-\rho^2)^{(p-1)/2}} \exp\left\{ \frac{-\rho}{2(1-\rho^2)} u'[2\rho I_p - \rho(e_1 e_1' + e_p e_p') - (C_1 + C_1')]u \right\}.$$

Componentwise it is

$$c(x) = \frac{1}{\sqrt{(1-\rho^2)^{p-1}}} \exp\left\{ \frac{-\rho}{2(1-\rho^2)} \left(2\rho \sum_{i=1}^{p} u_i^2 - \rho(u_1^2 + u_p^2) - 2\sum_{i=1}^{p-1} u_i u_{i+1} \right) \right\}.$$

This is again in accordance with Kelly and Krzysztofowicz (1997) for $p = 2$.

3. Choice of dependence measure

When working with non-elliptical distributions, it is better not to use Pearson ρ. Usual alternatives are Kendall τ and Spearman ρ_S:

- they are both invariant on monotone transformations;
- they are both measures of concordance.

Under normality, there is one-to-one relationship between these coefficients (Kruskal, 1958):

- $\tau = (2/\pi)\arcsin\rho \Leftrightarrow \rho = \sin(\pi\tau/2)$,
- $\rho_S = (6/\pi)\arcsin\frac{\rho}{2} \Leftrightarrow \rho = 2\sin(\pi\rho_S/6)$.

Thus, we can estimate τ or ρ_S and convert it to ρ, which is needed for Gaussian copula. Uniform correlation structure is not affected by this conversion. What can we expect in the case of serial structure? The main problem is that we convert one correlation coefficient to another, but their powers can behave differently.

At first, we investigated the differences between powers of τ and ρ under normality for various p. It can be proved that the maximum and minimum of $\tau^p - \rho^p$ is the same in absolute value, and is equal to

$$\sqrt{1-\frac{4}{\pi^2}} - \frac{2}{\pi}\cdot\arcsin\sqrt{1-\frac{4}{\pi^2}} \approx 0.211$$

for any p. The bigger the p, the closer are the extrema to ± 1. We can see it in Fig. 1.

Since the value 0.211 is relatively very big, and the differences follow various patterns for different p's, using powers of τ instead of ρ can substantially distort the dependence structure. That is why we cannot recommend using τ for construction of serial copula in non-elliptical distributions.

The situation with ρ_S is better. Maximum and minimum of $\rho_S^p - \rho^p$ is again the same in absolute value, shifting to ± 1 with increasing p, but its value is only

$$2\sqrt{1-\frac{9}{\pi^2}} - \frac{6}{\pi}\cdot\arcsin\sqrt{1-\frac{9}{\pi^2}} \approx 0.018.$$

This is illustrated in Fig. 2.

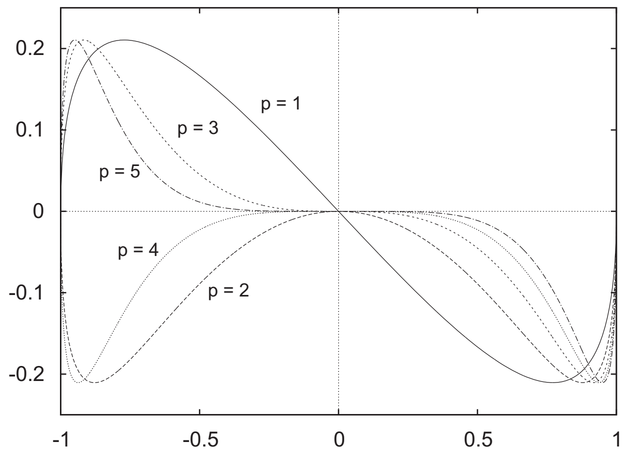

Fig. 1. Differences between powers of τ and ρ $(p = 1, \ldots, 5)$.

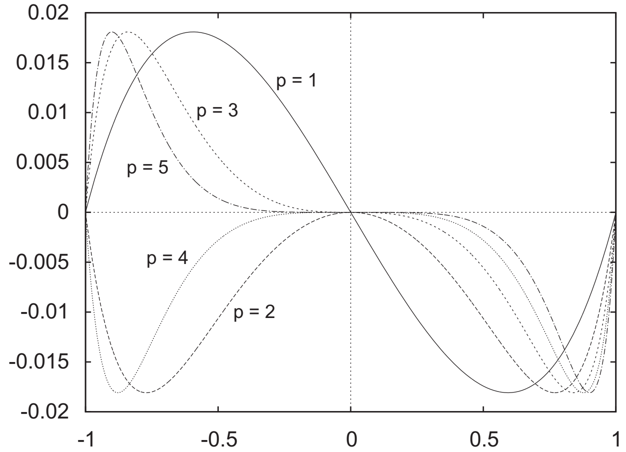

Fig. 2. Differences between powers of ρ_S and ρ $(p = 1, \ldots, 5)$.

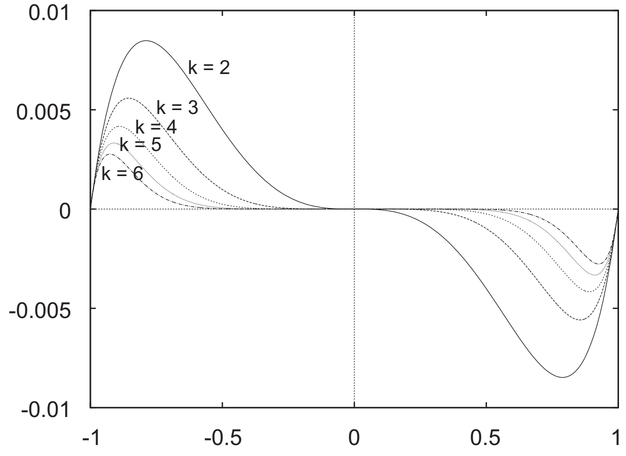

Fig. 3. Differences $(\rho_S^k/\rho_S^{k-1}) - \rho$ $(k = 2, \ldots, 6)$.

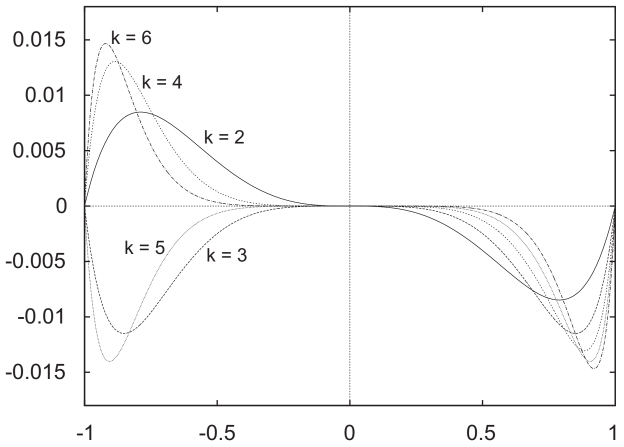

Fig. 4. Differences $(\rho_S^k/\rho_S) - \rho^{k-1}$ $(k = 2, \ldots, 6)$.

In this case, we investigated also the differences $(\rho_S^k/\rho_S^{k-1}) - \rho$ and $(\rho_S^k/\rho_S) - \rho^{k-1}$ under normality. Advantageous properties of ρ_S were also confirmed here, since the extrema converge to 0 with increasing k in the first case, and converge to $2\sqrt{1-(9/\pi^2)} - (6/\pi) \cdot \arcsin\sqrt{1-(9/\pi^2)}$ in the second case. This can be seen in Figs. 3 and 4.

As a result, we can recommend using ρ_S for the construction of serial Gaussian copula.

4. Conclusions

Multivariate Gaussian copulas with uniform and serial correlation structures seem to be a simple tool for modeling dependence in complex situations.

Spearman ρ_S is a good dependence measure for use in these situations, since the differences between powers of ρ_S and ρ are small.

References

Adermann, V., Pihlak, M., 2005. Using copulas for modeling the dependence between tree height and diameter at breast height. Acta et Commentationes Universitatis Tartuensis de Mathematica 9, 77–85.

Clemen, R.T., Reilly, T., 1999. Correlations and copulas for decision and risk analysis. Management Science 45 (2), 208–224.

Kelly, K.S., Krzysztofowicz, R., 1997. A bivariate meta-Gaussian density for use in hydrology. Stochastic Hydrology and Hydraulics 11 (1), 17–31.

Kruskal, W., 1958. Ordinal measures of association. JASA 53, 814–861.