Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2006) 809–822
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Bayesian copula selection

David Huard, Guillaume Évin, Anne-Catherine Favre*

Institut National de la Recherche Scientifique, Centre Eau, Terre & Environnement, Qué., Canada G1K 9A9

Received 1 March 2005; received in revised form 23 June 2005; accepted 30 August 2005
Available online 6 October 2005

Abstract

In recent years, the use of copulas has grown extremely fast and with it, the need for a simple and reliable method to choose the right copula family. Existing methods pose numerous difficulties and none is entirely satisfactory. We propose a Bayesian method to select the most probable copula family among a given set. The copula parameters are treated as nuisance variables, and hence do not have to be estimated. Furthermore, by a parameterization of the copula density in terms of Kendall's τ, the prior on the parameter is replaced by a prior on τ, conceptually more meaningful. The prior on τ, common to all families in the set of tested copulas, serves as a basis for their comparison. Using simulated data sets, we study the reliability of the method and observe the following: (1) the frequency of successful identification approaches 100% as the sample size increases, (2) for weakly correlated variables, larger samples are necessary for reliable identification.

Keywords: Copulas; Model selection; Bayes' theorem; Goodness-of-fit test; Kendall's tau; Pseudo-likelihood

0. Introduction

In order to extrapolate extreme quantiles from data sets, or to generate random variables, it is usually necessary to select a distribution function matching the available data. The choice of the best distribution is not an exact science and relies on guesswork and testing of multiple hypotheses. Since each hypothesis comes with its particular test, the whole procedure is too complicated for end-users and generally left to experts, along with the interpretation of the results. Furthermore, existing methods cannot compare distributions without specifying an optimal parameter set for each one of them. The selection of the best distribution is thus intertwined with the estimation of parameters, a non-trivial problem itself.

The situation is even worse in the case of two-dimensional distributions, for which even more parameters need to be estimated. Fortunately, the elegant concept of copulas greatly simplifies matters. Copulas are multivariate distributions modeling the dependence structure between variables, irrespective of their marginal distribution. They allow to choose completely different margins, the dependence structure given by the copula, and merge the margins into a genuine multivariate distribution. The choice of the best bivariate distribution can then be done in two steps: choose the optimal margins, and then choose the optimal copula. In this paper, we introduce a simple Bayesian method to choose the "best" copula, given some bivariate data expressed by quantiles.

* Corresponding author. Tel.: +1 418 6543789; fax: +1 418 6542600.
E-mail addresses: david.huard@ete.inrs.ca (D. Huard), guillaume.evin@ete.inrs.ca (G. Évin), anne-catherine_favre@ete.inrs.ca (A.-C. Favre).

0167-9473/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2005.08.010

The structure of the paper is as follows Section 1 introduces the main ideas of copula theory. Section 2 reviews existing approaches to select copulas and highlights salient features. Section 3 describes the proposed method and its derivation from Bayes' theorem. Results from numerical simulations are shown in Section 4, along with their analysis. Finally, we draw conclusions on the overall performance of the method and propose ideas for future work.

1. Copula theory

The concept of copula has been introduced by Sklar (1959) in the following way

Copula Definition. A copula is a joint distribution function of standard uniform random variables. That is,

$$C\left(u_1,\ldots,u_p\right)=\Pr\left\{U_1\leqslant u_1,\ldots,U_p\leqslant u_p\right\},$$

where $U_i \sim U(0,1)$ for $i=1,\ldots,p$.

For a more formal definition of copulas, the reader is referred to Nelsen (1999). Using the probability integral transformation, it is straightforward to see that a copula computed at $F_1(x_1), F_2(x_2),\ldots,F_p\left(x_p\right)$ is identical to the multivariate distribution function F evaluated at $\left(x_1,\ldots,x_p\right)$, i.e.,

$$C\left(F_1(x_1),F_2(x_2),\ldots,F_p\left(x_p\right)\right)=F\left(x_1,x_2,\ldots,x_p\right).$$

This last equality gives a first insight of the link between distribution functions and copulas, which is the content of Sklar's theorem.

Sklar's Theorem. *Let F be a p-dimensional distribution function with margins $F_1, F_2\ldots,F_p$, then there exists a p-copula C such that for all x in $\bar{\mathbb{R}}^p$,*

$$F\left(x_1,x_2,\ldots,x_p\right)=C\left(F_1(x_1),F_2(x_2),\ldots,F_p\left(x_p\right)\right),$$

where $\bar{\mathbb{R}}$ denotes the extended real line $[-\infty,\infty]$. If $F_1,\ldots,F_p$ are all continuous, then C is unique. Otherwise, C is uniquely determined on $Ran(F_1)\times Ran(F_2)\times\cdots\times Ran\left(F_p\right)$, where Ran stands for the range.

According to Sklar's theorem, copulas separate marginal behavior, as represented by the F_i's, from the dependence structure. This constitutes one great advantage of copulas. In the usual representation of joint probabilities via multivariate distribution functions, the two cannot be separated. The general theory about copulas is summarized in Joe (1997), Nelsen (1999) or more recently in Cherubini et al. (2004). Copulas have been widely used in financial mathematics to determine the Value at Risk (see for example, Embrechts et al., 2002, 2003; Bouyé et al., 2000). Other fields of applications involve lifetime data analysis (Bagdonavicius et al., 1999), actuarial science (Frees and Valdez, 1998), and more recently, hydrology (De Michele and Salvadori, 2003; Favre et al., 2004).

Most copula applications are concerned with bivariate data. One reason for this is that relatively few copula families have practical p-dimensional generalization. The popular Archimedean 2-copulas (Genest and MacKay, 1986) for instance, have two known generalizations, both of them afflicted by serious shortcomings. Archimedean 2-copulas are defined as

$$C(u_1,u_2)=\begin{cases}\varphi^{-1}\left(\varphi(u_1)+\varphi(u_2)\right) & \text{if } \sum_{i=1}^{2}\varphi(u_i)\leqslant\varphi(0),\\ 0 & \text{otherwise}\end{cases}$$

with $\varphi(u)$ a $\mathscr{C}^2$ function satisfying $\varphi(1)=0$, $\varphi'(u)<0$ (φ is decreasing) and $\varphi''(u)>0$ (φ is convex) for all $0\leqslant u\leqslant 1$. $\varphi(u)$ is called the generator of the copula. The first generalization, termed symmetric (Joe, 1997), uses the same generator, thus the same dependence, for all variables

$$C\left(u_1,\ldots,u_p\right)=\varphi^{-1}\left(\varphi(u_1)+\cdots+\varphi\left(u_p\right)\right).$$

Since all variables are described by the same dependence, this generalization is too simplistic for most real life applications. The second generalization, termed asymmetric (Whelan, 2004), uses $(p-1)$ generators. For $p=3$, the