Journal of Multivariate Analysis 120 (2013) 85–101

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

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

Factor copula models for multivariate data

Pavel Krupskii *, Harry Joe

Department of Statistics, University of British Columbia, Vancouver, BC, V6T 1Z4 Canada

ARTICLE INFO

Article history:
Received 19 August 2012
Available online 27 May 2013

AMS 2000 subject classifications:
62H25
60H99

Keywords:
Conditional independence
Factor analysis
Pair-copula construction
Partial correlation
Tail dependence
Tail asymmetry
Truncated vine

ABSTRACT

General conditional independence models for d observed variables, in terms of p latent variables, are presented in terms of bivariate copulas that link observed data to latent variables. The representation is called a factor copula model and the classical multivariate normal model with a correlation matrix having a factor structure is a special case. Dependence and tail properties of the model are obtained. The factor copula model can handle multivariate data with tail dependence and tail asymmetry, properties that the multivariate normal copula does not possess. It is a good choice for modeling high-dimensional data as a parametric form can be specified to have $O(d)$ dependence parameters instead of $O(d^2)$ parameters. Data examples show that, based on the Akaike information criterion, the factor copula model provides a good fit to financial return data, in comparison with related truncated vine copula models.

1. Introduction

The multivariate normality assumption is widely used to model the joint distribution of high-dimensional data. The univariate margins are transformed to normality and then the multivariate normal distribution is fitted to the transformed data. In this case, the dependence structure is completely defined by the correlation matrix and different models on the correlation structure can be used to reduce the number of dependence parameters from $O(d^2)$ to $O(d)$, where d is the multivariate dimension or number of variables. When dependence in the observed variables is thought to be explained by a few latent variables, the Gaussian or normal factor model assumes a linear relation on a few unobserved normally distributed factors.

The main contribution of this paper is to propose and study the copula version of the multivariate normal distribution with a correlation matrix that has a factor structure. We name the extension as the *factor copula model*. The classical factor model is a special case but within our framework, the parameterization is different as it involves partial correlations. The factor copula model is useful when the dependence in observed variables is based on a few unobserved variables, and there exists tail asymmetry or tail dependence in the data, so that the multivariate normality assumption is not valid.

The copula is a function linking univariate margins into the joint distribution. Let $\mathbf{X} = (X_1, \ldots, X_d)$ be a random d-dimensional vector with the joint cumulative distribution function (cdf) $F_{\mathbf{X}}$. Let F_{X_j} be the marginal cdf of X_j for $j = 1, \ldots, d$. The copula $C_{\mathbf{X}}$, corresponding to $F_{\mathbf{X}}$, is a multivariate uniform cdf such that $F_{\mathbf{X}}(x_1, \ldots, x_d) = C_{\mathbf{X}}(F_{X_1}(x_1), \ldots, F_{X_d}(x_d))$. By Sklar [26], there exists a unique copula $C_{\mathbf{X}}$ if $F_{\mathbf{X}}$ is continuous. Copulas are suitable for modeling non-normal data such as financial asset returns or insurance data; see [24,21] and others. The superiority of non-normal copulas over the normal copula in modeling financial and insurance data has been discussed in [7].

* Corresponding author.
E-mail address: KrupskiiPV@yandex.ru (P. Krupskii).

0047-259X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmva.2013.05.001

The vine copula or the pair-copula construction has been popular in recent years; see, for example [20,4]. The number of bivariate (conditional or marginal) copulas used in the vine construction is $d(d-1)/2$ for d variables, so typically vine copulas involve $O(d^2)$ number of parameters. Dißmann et al. [6] propose an algorithm which allows the fit of regular vines to data. If d is large (e.g., many asset returns), the conditional independence can be assumed at higher levels of the vine to reduce the number of parameters in a truncated vine model to $O(d)$; see also [5]. The factor copula model is an alternative copula modeling approach to truncated vines that has the order of $O(d)$ dependence parameters.

In many multivariate applications, the dependence in observed variables can be explained through latent variables; in multivariate item response in psychology applications, latent variables are related to the abstract variable being measured through items, and in finance applications, latent variables are related to economic factors. Classical factor analysis assumes (after transforms) that all observed and latent random variables are jointly multivariate normal. Books on multivariate analysis (see for example [16]) often have examples with factor analysis and financial returns. We show for some financial return data that, in terms of the Akaike or Bayesian information criteria, the factor copula model can be a better fit than truncated vines (because of a simpler dependence structure) and the classical factor model (because of tail dependence).

An important advantage of factor models is that they can be nicely interpreted. In case of stocks in a common sector, the current state of this sector can affect all of their change of prices, but the sector index, if measured, might not contain all of the latent information that explains the dependence. Similarly for market data, the state of the economy as a whole can determine the latent dependence structure. The "state variables" are aggregated from many exogenous variables (such as interest rate, refinancing rate, political instabilities, etc.) and cannot be easily measured, therefore factor copula models based on latent variables might be a good choice.

The rest of this paper is organized as follows. In Section 2 we define the factor copula model and give more details for the one-factor and two-factor models. Some dependence and tail properties of bivariate margins of the factor copula models are given in Section 3. The results imply that different types of dependence and tail asymmetry can be modeled with appropriate choices of bivariate linking copulas. Computational details for maximum likelihood estimation of the factor copula model are given in Section 4. Section 5 discusses diagnostics for choices of bivariate linking copulas, reports on some simulation results, and shows applications of the factor copula model to US stock returns. Section 6 concludes with a discussion of future research.

2. Factor copula model

In multivariate models with copulas, a copula or multivariate uniform distribution is combined with a set of univariate margins. This is equivalent to assuming that variables $X_1, \ldots, X_d$ have been transformed to uniform random variables. So we assume that $\mathbf{U} = (U_1, \ldots, U_d)$ is a random vector with $U_i \sim U(0, 1)$. The joint cdf of the vector $\mathbf{U}$ is then given by $C(u_1, \ldots, u_d)$ where C is a d-dimensional copula. In the p-factor copula model, $U_1, \ldots, U_d$ are assumed to be conditionally independent given p latent variables $V_1, \ldots, V_p$. Without loss of generality, we can assume V_i are independent and identically distributed (i.i.d.) $U(0, 1)$. Let the conditional cdf of U_j given $V_1, \ldots, V_p$ be denoted by $F_{j|V_1,\ldots,V_p}$. Then,

$$C(u_1, \ldots, u_d) = \int_{[0,1]^p} \prod_{j=1}^{d} F_{j|V_1,\ldots,V_p}(u_j|v_1, \ldots, v_p)\, dv_1 \cdots dv_p. \tag{1}$$

Any conditional independence model, based on p independent latent variables, can be put in this form after transforms to $U(0, 1)$ random variables. Hence, the dependence structure of $\mathbf{U}$ is then defined through conditional distributions $F_{1|V_1,\ldots,V_p}, \ldots, F_{d|V_1,\ldots,V_p}$. We will call (1) a *factor copula model*, with $F_{j|V_1,\ldots,V_p}$ expressed appropriately in terms of a sequence of bivariate copulas that link the observed variables U_j to the latent variables V_k. Some of the bivariate copulas are applied to conditional distributions. Details are given in Sections 2.1 and 2.2.

In the finance literature there are several factor copula models (e.g., Section 9.7.2 of McNeil et al. [21], Hull and White [12] and Oh and Patton [23]); these all have a linear latent structure and are not as general as our model. Oh and Patton [23] have overlapping ideas with our research but our approach was developed independently of their approach. With the conditional independence model with 2 or more latent variables, there could be alternative ways to specify a model for $F_{j|V_1,\ldots,V_p}$ than we have. Throughout the remainder of this paper, we assume that all copulas are absolutely continuous and have densities, so that the log-likelihood for continuous data will involve the density of the factor copula.

2.1. One- and two-factor copula models

We first study the case of $p = 1$ latent variable in (1). For $j = 1, \ldots, d$, denote the joint cdf and density of (U_j, V_1) by C_{j,V_1} and c_{j,V_1} respectively. Since U_1, V_j are $U(0, 1)$ random variables, then $F_{j|V_1}$ is just a partial derivative of the copula C_{j,V_1} with respect to the second argument. That is, $F_{j|V_1}(u_j|v_1) = C_{j|V_1}(u_j|v_1) = \partial C_{j,V_1}(u_j, v_1)/\partial v_1$. With $p = 1$, Eq. (1) becomes:

$$C(u_1, \ldots, u_d) = \int_0^1 \prod_{j=1}^{d} F_{j|V_1}(u_j|v_1)\, dv_1 = \int_0^1 \prod_{j=1}^{d} C_{j|V_1}(u_j|v_1)\, dv_1. \tag{2}$$