Journal of Multivariate Analysis 101 (2010) 869–881

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Estimation and inference for dependence in multivariate data

Olha Bodnar [a], Taras Bodnar [a], Arjun K. Gupta [b,*]

[a] Department of Statistics, European University Viadrina, PO Box 1786, 15207 Frankfurt (Oder), Germany

[b] Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403, USA

ARTICLE INFO

Article history:
Received 22 September 2008
Available online 22 November 2009

AMS 2000 subject classifications:
62H12
Secondary 62H20

Keywords:
Multivariate non-normal distribution
Multivariate copula
Gaussian copula
Correlation matrix
Estimation and inference procedure
Pseudo-maximum likelihood method
Test of independence

ABSTRACT

In this paper, a new measure of dependence is proposed. Our approach is based on transforming univariate data to the space where the marginal distributions are normally distributed and then, using the inverse transformation to obtain the distribution function in the original space. The pseudo-maximum likelihood method and the two-stage maximum likelihood approach are used to estimate the unknown parameters. It is shown that the estimated parameters are asymptotical normally distributed in both cases. Inference procedures for testing the independence are also studied.

Published by Elsevier Inc.

1. Introduction

Modeling and estimation of the multivariate distribution is an important issue which has a large number of possible applications in different fields of science. The estimation and the inference for dependence in multivariate data is a related problem of equal importance. Although both issues are most crucial when multivariate data are analyzed and a lot of multivariate models are suggested in the literature, the problem is still unsolved and attracts many researchers and practitioners.

The models for multivariate data can be divided into two large groups. In the first one, the conditional moments of the distribution, usually the first two ones, are modeled. Since the seminal paper of Engle [1] who introduced the autoregressive conditional heteroscedastic (ARCH) process to model the conditional variances of univariate data, the univariate and multivariate generalizations of the ARCH process have become very popular in financial and econometrical literature. The first multivariate ARCH process was derived by Bollerslev et al. [2]. Engle and Kroner [3] designed the BEKK version of the multivariate ARCH process. Other approaches considered diagonal and orthogonal versions of the multivariate ARCH process (see, e.g. [4–6]). In order to model conditional correlations the constant conditional correlation (CCC) and the dynamic conditional correlation (DCC) processes were suggested by Bollerslev [7], Engle [8], and Tse and Tsui [9]. A detailed survey of multivariate ARCH processes is given in [10]. Although the multivariate ARCH models play an important role in modeling multivariate data, they do not answer the question how strong is the dependence. The problem is that they all model the conditional covariance (or correlation) matrix which is the measure of dependence only in the case of the multivariate normal distribution. The second drawback of the multivariate ARCH models is the joint distribution function which is not specified in the closed form up until now.

* Corresponding author.
E-mail address: gupta@bgsu.edu (A.K. Gupta).

0047-259X/$ – see front matter. Published by Elsevier Inc.
doi:10.1016/j.jmva.2009.11.005

The other possibility of modeling multivariate data is to model directly the joint distribution function. The most widespread approach is to assume that the data follow a multivariate normal distribution. Note that only in the case of the normal distribution, the dependence structure is fully determined by the correlation matrix. We use this fact later on when a new procedure of modeling the joint distribution will be presented. Despite this nice property, the application of the normal distribution to model multivariate data is heavily criticized. The main points are the heavy tails and asymmetry usually observed in the empirical distributions of data. Generalizations of the multivariate normal distribution have been done in two directions, namely elliptically contoured distributions (see, e.g. [11,12]) and a skew normal distribution (see, e.g. [13–16]). Although the elliptically contoured distributions provide a good fit for heavy tails, they are also symmetric. The main problem with modeling by the skew normal distribution is the estimation and inferences of the model parameters.

The copula based modeling of multivariate distribution has recently increased its popularity (see, e.g. [17–22]). The method is based on Sklar's [23] theorem that relates an arbitrary distribution function F on $\mathbb{R}^k$ to a copula function C via the univariate marginal distributions F_i, $i = 1, \ldots, k$ of F. The relationship between the distribution function and the copula function is given by

$$F(x_1, \ldots, x_k) = C(F_1(x_1), \ldots, F_k(x_k)).$$

Moreover, if the marginal distributions are continuous then the copula function is uniquely specified. The form of the unique copula is not known. From one side it provides a flexibility of copula modeling that results in different forms of the copula functions. The most popular of them are elliptical and Archimedean families. From the other side, we can be sure that the selected form of the copula function is the true one.

In this paper, we suggest a new measure of dependence and study its distributional properties. Our approach is based on a transformation of the data to a space where the dependence structure can be simply modeled and the joint distribution function can be constructed. Then, using the inverse transformation, the distribution function is obtained in the original space of the data. We use the transformation

$$U_i = \Phi^{-1}(F_i(x_i)), \quad i = 1, \ldots, k, \tag{1}$$

that transforms the univariate data to the space where they are normally distributed. Note that the transformation (1) is not new and has previously been considered in the statistical literature. For example, Efron [24] used this transformation for constructing an improved estimator of the bootstrap confidence intervals. The transformation (1) allows us to determine the structure of the joint distribution function in terms of the univariate marginal distributions and a $k \times k$ correlation matrix that fully describes the dependence structure of multivariate data. This can be done if the univariate marginal distributions are continuous and the joint distribution can be expressed as a Gaussian copula.

Our main results are given in the next section, where a new measure of dependence is presented and its distributional properties are studied. In Section 2.2, we discuss how it can be estimated. The pseudo-maximum likelihood method and the two-stage maximum likelihood approach are used. It is shown, that the estimated parameters are asymptotically normally distributed. In Sections 2.3 and 2.4, the inference procedures for testing the independence for multivariate data are given. An application of the suggested approach to the canonical correlation analysis is presented in Section 3.

2. Main results

2.1. Modeling the multivariate dependence

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be independent realizations of the k-dimensional random vector $\mathbf{X} = (X_1, X_2, \ldots, X_k)'$ with the continuous distribution function $F(.)$ and the marginal distributions $X_i \sim F_i(.)$. There are different ways to measure the dependence between the elements of $\mathbf{X}$. The most widely used is the Pearson correlation coefficient, which is the measure of the linear dependence, and it is defined as

$$\rho_{X_i,X_j} = \frac{\sum_{l=1}^{n}(x_{il} - \bar{x}_i)(x_{jl} - \bar{x}_j)}{\sqrt{\sum_{l=1}^{n}(x_{il} - \bar{x}_i)^2}\sqrt{\sum_{l=1}^{n}(x_{jl} - \bar{x}_j)^2}}, \tag{2}$$

where $\mathbf{x}_l = (x_{1l}, x_{2l}, \ldots, x_{kl})'$ and $\bar{x}_i = \frac{1}{n}\sum_{l=1}^{n} x_{il}$. When $\mathbf{X} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then ρ_{X_i,X_j} is also a dependence measure between the elements of the vector $\mathbf{X}$. It holds that X_i and X_j are independent iff $\rho_{X_i,X_j} = 0$.

The Spearman rank-order correlation coefficient is a non-parametric measure of the dependence defined using the ranks of the data values. It is given by

$$\theta_{X_i,X_j} = \frac{\sum_{l=1}^{n}(\mathrm{rank}(x_{il}) - \overline{\mathrm{rank}(x_i)})(\mathrm{rank}(x_{jl}) - \overline{\mathrm{rank}(x_j)})}{\sqrt{\sum_{l=1}^{n}(\mathrm{rank}(x_{il}) - \overline{\mathrm{rank}(x_i)})^2}\sqrt{\sum_{l=1}^{n}(\mathrm{rank}(x_{jl}) - \overline{\mathrm{rank}(x_j)})^2}}, \tag{3}$$