Available online at www.sciencedirect.com
ScienceDirect
Insurance: Mathematics and Economics 42 (2008) 127–146
INSURANCE MATHEMATICS & ECONOMICS
www.elsevier.com/locate/ime

Modelling dependence

Wilbert C.M. Kallenberg

Department of Applied Mathematics, Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

Received August 2006; received in revised form January 2007; accepted 15 January 2007

Abstract

A new way of choosing a suitable copula to model dependence is introduced. Instead of relying on a given parametric family of copulas or applying the other extreme of modelling dependence in a nonparametric way, an intermediate approach is proposed, based on a sequence of parametric models containing more and more dependency aspects. In contrast to a similar way of thinking in testing theory, the method here, intended for estimating the copula, often requires a somewhat larger number of steps. One approach is based on exponential families, another on contamination families. An extensive numerical investigation is supplied on a large number of well-known copulas. The method based on contamination families is recommended. A Gaussian start in this approximation looks very promising.

MSC: 62H12; 62H20; 62P05

Keywords: Copula; Legendre polynomials; Exponential family; Contamination family; Nonlinear correlation

1. Introduction

The classical way to deal with dependence for a multivariate distribution is to assume multivariate normality and to estimate the correlation coefficients. Also outside the normal model linear correlation is often taken as a tool to measure dependence. However, capturing only linear correlation is far too restricted. Other forms of dependence are important too. In particular in finance and insurance (see e.g. Cherubini et al. (2004), Embrechts et al. (2002, 2003), McNeil et al. (2005)), but also in other areas like for instance hydrology (see e.g. Genest and Favre (2007)), there is last year's increased attention given to going beyond linear dependence.

A second problem with linear correlation is that the marginal distributions are mixed up with the dependence. Sklar's (1959, 1996) theorem shows that for continuous multivariate distribution functions, the univariate margins and the multivariate dependence structure can be separated, and the dependence structure can be represented by a so called copula. This copula is the multivariate distribution function of the random vector obtained by applying on each of the components its probability integral transformation, thus giving them uniform marginals. Embrechts et al. (2003) remark: "since linear correlation is not a copula-based measure of dependence, it can often be quite misleading and should not be taken as the canonical dependence measure". For a lot of results on copulas see also Joe (1997), Nelsen (1999), Cherubini et al. (2004), McNeil et al. (2005).

In view of Sklar's theorem, the study of multivariate dependence can be performed in two distinct steps: estimating the marginal distributions and estimating the "intrinsic" dependence structure. The first step is very well-known. Here we investigate the second step. What should be done, is choosing an appropriate (family of) copula(s).

There are many families of copulas proposed in the literature, each with its own merits. One may rely on a parametric family of copulas, like the Frank copulas or the Gumbel copulas, etc. Having chosen the family one (only) needs to estimate the parameter(s) of this family. However, the choice of the parametric family is not that clear. A possible way-out is to check whether a certain copula or family of copulas is suitable. Goodness-of-fit tests for the simple null hypothesis of a given copula, or the composite hypothesis of a parametric family of copulas are developed, see e.g. Fermanian (2005), Panchenko (2005) and references therein. But in case

E-mail address: w.c.m.kallenberg@math.utwente.nl.

0167-6687/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.insmatheco.2007.01.008

of rejection, it is not clear what to do. In Biau and Wegkamp (2005) the problem of finding a particular copula, given a (parametric) class of candidate copulas is attacked. They restrict attention to copulas with a bounded density. In their oracle inequality the upper bound consists of a model error term, expressing the distance between the true density and the parametric family of candidate copulas, and a second part giving the stochastic error or estimation error. We will also split up the total error in the model error and the stochastic error, see (2.7), (2.8) and (2.11) and (2.12).

The advantage of a parametric family is that only one (or a few) parameters have to be estimated, thus obtaining a relatively small stochastic error. The disadvantage might be the restriction to one family and a possible gap between the true density and the chosen family. The latter can be avoided by the other extreme of a nonparametric approach. But in that case the estimation step will lead to large errors, unless we have an enormous amount of observations. Here we propose an intermediate approach. This intermediate approach consists of two steps: a modelling step and an estimation step. In the modelling step a sequence of parametric copula models is introduced, approximating the true copula more and more. In the estimation step out of this sequence of parametric models a suitable one is selected (using a model selection rule) and subsequently the parameters within the selected model are estimated. To avoid too many technicalities we concentrate in this paper on bivariate distributions. Moreover, we concentrate in this paper on the modelling step. The estimation step will be treated in a forthcoming paper. Obviously, firstly it should be made clear that the modelling step has good approximation properties. Therefore, the aim of the present paper is to investigate the approximation error in the modelling step.

To model (and afterwards estimate) the true copula density, a sequence of parametric models is introduced, containing more and more dependency aspects. The method has a parametric flavour, but considering higher and higher dimensions we get in the limit the true density and in this way the method is "nonparametric". A somewhat similar approach is applied successfully in testing theory, see e.g. Kallenberg and Ledwina (1999) and Janic-Wróbleska et al. (2004). However, as a rule, in testing theory heavy forms of dependence are detected easily and therefore main attention is on copulas not too far from independence. In estimation theory the whole scope of dependent copulas should be considered carefully. This makes the modelling step more difficult.

Starting point is a given (family of) copula(s). For instance, one may simply start with the uniform density on the unit square (corresponding to independence). Another prominent starting point is the family of Gaussian copulas. Other favorite starting points of families of copulas can be used as well. Dependency aspects not covered by the starting point are added by subsequent parametric steps. In this way the method automatically improves an a priori chosen parametric family. In particular, when the starting point is not too far away from the true copula, only a few steps are needed to get a sufficiently small model error.

Well-known families of parametric models for the subsequent parametric steps are so called exponential families. For properties of exponential families we refer to Barndorff-Nielsen (1978). Approximation of (univariate) densities by exponential families has been done e.g. by Barron and Sheu (1991), Yang and Barron (1998) and Castellan (2003). Contamination families are candidates as well. The advantage of exponential families over contamination families is that the density is automatically positive and integrates to 1. However, the estimation step is more complicated. Moreover, the marginal distributions are no longer uniform distributions, implying that fitting covariances is not equivalent to fitting correlations.

In general, two random variables X and Y are independent if and only if $\mathrm{cov}(f_1(X), f_2(Y)) = 0$ for all f_1 and f_2 ranging over a separating class of functions (see e.g. Breiman (1968), p. 165 ff.). Eubank et al. (1987) have considered a measure of association, called ϕ^2 (see also Lancaster (1969), p. 91 ff.). Let $U = F_X(X)$, $V = F_Y(Y)$, where F_X and F_Y are the marginal distribution functions of X and Y, respectively, and let b_j be the jth Legendre polynomial on $(0, 1)$. If $\phi^2 < \infty$, then the condition $\mathrm{cov}(b_r(U), b_s(V)) = 0$ for all $r, s \geq 1$ implies that X and Y are independent. So, under this mild condition, the Legendre polynomials form a separating class. Therefore, both the exponential families and the contamination families are based on suitable Legendre polynomials.

The parametric steps are designed to fit $Eb_r(U)b_s(V)$. For the contamination family this is equivalent to fitting $\mathrm{cov}(b_r(U), b_s(V))$ or the correlation coefficient $\rho(b_r(U), b_s(V))$. For instance, when $r = s = 1$ this concerns the linear correlation of the copula. Within the exponential family the maximum likelihood estimator produces the required fit of $Eb_r(U)b_s(V)$. At the same time this member of the exponential family is closest in terms of Kullback Leibler information to the true density. For the contamination family moment estimators are invoked. They are linked up with the L_2-distance. Both for the exponential families and for the contamination families it holds that the higher the dimension of the family, the better the fit and hence the smaller the model error. On the other hand, the higher the dimension, the more parameters have to be estimated and the larger the stochastic error due to the estimation part.

It turns out that finding those parameters in the exponential family that fit $Eb_r(U)b_s(V)$ is much more difficult than the corresponding step in the contamination family. In the k-parameter exponential family k (rather complicated) equations should be solved, while the contamination family gives explicit expressions for the parameters involved. Moreover, fitting $Eb_r(U)b_s(V)$ in the contamination family gives automatically a fit of the covariance and correlation coefficient. Also the reduction in model error due to taking a larger dimension has a far more easy form for the contamination model.

The paper is organized as follows. In Section 2 the exponential families and the contamination families are introduced. Properties of these families in terms of Kullback Leibler information for the exponential families and L_2-distance for the contamination families are derived. Moreover, the decomposition of the total error in model error and stochastic error is discussed. After introduction of the