Estimating Risk with Sarmanov Copula and Nonparametric Marginal Distributions

Zuhair Bahraoui, Catalina Bolancé, and Ramon Alemany

Dept. Econometrics, Riskcenter-IREA
University of Barcelona Av. Diagonal,
690, 08034 Barcelona, Spain
{zuhair,bolance,ralemany}@ub.edu

Abstract. We show that Sarmanov copula and kernel estimation can be mixed to estimate the risk of an economic loss. We use a bivariate sample from a real data base. We show that the estimation of the dependence parameter of the copula using double transformed kernel estimation to estimate marginal cumulative distribution functions provides balanced risk estimates.

Keywords: Copula, kernel estimation, value at risk, conditional tail expectation.

1 Introduction

Estimating the risk of loss is a major challenge in finance and in insurance, which has been extensively studied in the literature (see, for instance, the books by [1], [2] and [3] or articles such as [4], [5] and [6], among many others). In this work, we propose to use the Sarmanov copula (see [7]) with non-parametric marginals to estimate the risk of a loss that is obtained as the aggregation of two dependent losses. Value-at-Risk (VaR) and Tail Value-at-Risk (TVaR) are the selected risk measures. We show that estimating marginals using double transformed kernel estimation (DTKE) as proposed by [6] is the method that best fits our purpose. We apply our proposal to a real insurance database corresponding to a random bivariate sample of the cost of claims.

The aim of this work is to show as the transformed kernel estimator of cumulative distribution function allows us to obtain a good fit of the Sarmanov copula. Unlike the rest, in this copula the dependence structure is not separated strictly of the marginal distributions, i.e. the marginals are incorporated into the dependence structure; therefore, the estimation of these marginal distributions is essential to estimate the parameter of the copula.

2 Sarmanov Copula

Let (X,Y) be a bivariate random vector with marginal probability distribution functions (pdfs) f_X and f_Y. Also, let ϕ_1 and ϕ_2 be two bounded non constant function such that:

$$\int_{-\infty}^{+\infty} f_X(t)\phi_1(t)dt = 0, \quad \int_{-\infty}^{+\infty} f_y(t)\phi_2(t)dt = 0 .$$

M.Á. Fernández-Izquierdo, M.J. Muñoz-Torres, R. León (Eds.): MS 2013, LNBIP 145, pp. 91–98, 2013.

Then the bivariate pdf introduced by [7] is defined as:

$$h(x,y)=f_X(x)f_y(y)(1+\omega\phi_1(x)\phi_2(y)).$$

From Sklar's theorem, we deduce that the associated copula can be expressed as:

$$C(u,v)=uv+\omega\int_0^u\int_0^v\phi_1(F_X^{-1}(t))\phi_2(F_y^{-1}(s))dtds. \quad (1)$$

and its density is:

$$c(u,v)=1+\omega\phi_1(F_X^{-1}(u))\phi_2(F_y^{-1}(v)). \quad (2)$$

where F_X and F_Y are cumulative distribution functions (cdfs) of X and Y, respectively. Parameter ω is a real number that satisfies the condition $1+\omega\phi_1(x)\phi_2(y)\geq 0$ for all x and y. This parameter is related to the correlation between X and Y (if it exists), ω is called the dependence parameter. As [8] shows, the dependence between X and Y is:

$$\rho=\frac{v_1v_2}{\sigma_1\sigma_2},$$

where $v_1=E(X\phi_1(X))$, $v_2=E(X\phi_2(X))$, $\sigma_1^2=Var(X)$ and $\sigma_2^2=Var(Y)$ When we take $\phi_1(x)=1-2F_X(x)$ and $\phi_2(y)=1-2F_Y(y)$, we have the classical Farlie-Gumbel–Morgenstern (FGM) copula. In this case the dependence parameter has the range $-1/3\leq\omega\leq 1/3$.

Another special case is when we consider functions of the type:

$$\phi_1(x)=x-\mu_X \text{ and } \phi_2(y)=y-\mu_Y \quad (3)$$

where $\mu_X=E(X)$ and $\mu_Y=E(Y)$. The author in [8] shows that, if the support of f_X and f_Y is contained in $[0,1]$, then the range of the dependence parameter is:

$$\max\left(\frac{-1}{\mu_X\mu_Y},\frac{-1}{(1-\mu_X)(1-\mu_Y)}\right)\leq\omega\leq\min\left(\frac{1}{\mu_X(1-\mu_Y)},\frac{1}{(1-\mu_X)\mu_Y}\right).$$

If the support of f_X is contained in $[a,b]$ and f_Y is contained in $[c,d]$, we can easy prove that:

$$\max\left(\frac{-1}{(b-\mu_X)(d-\mu_Y)},\frac{-1}{(\mu_X-a)(\mu_Y-c)}\right)\leq\omega\leq\min\left(\frac{1}{(b-\mu_X)(\mu_Y-c)},\frac{1}{(\mu_X-a)(d-\mu_Y)}\right).$$

.

2.1 Simulating from the Sarmanov Copula

To generate a bivariate random variable from Sarmanov's copula in (1), we use the procedure described by [9] which is based on the conditional distribution of a random vector (U,V)

$$P(V \le v \mid U = u) = C_u(v),$$

where:

$$C_u(v) = \lim_{\delta \to 0^{\leftarrow}} \frac{C(u+\delta, v) - C(u,v)}{\delta} = \frac{\partial C(u,v)}{\partial u}.$$

The algorithm is implemented as follows:

1. Generate two independent random variables u and t from an Uniform$(0,1)$ distribution.
2. Set $v = C_u^{-1}(t)$, where $C_u^{-1}(t)$ denotes a quasi-inverse of C_u.
3. The desired pair is (u,v).

For our case, when we consider functions ϕ_1 and ϕ_2 as defined in (3), we have:

$$C_u(v) = v + \omega\left(F_X^{-1}(u) - \mu_X\right)\int_0^v \left(F_Y^{-1}(s) - \mu_Y\right)ds. \tag{4}$$

This result can easily be shown if the derivative of (1) is calculated or, alternatively, using the following relationship (see Lee, 1996):

$$P(Y \le y \mid X = x) = F_Y(y) - \omega\int_y^{+\infty} f_Y(t)\phi_2(t)dt, \tag{4}$$

Taking $v = F_Y(y)$ and $u = F_X(x)$. Expression (4) can be calculated using the change of variable $t = F_Y^{-1}(s)$ in the integral.

3 Nonparametric Approximation of cdf

In this work, we propose to use different ways of obtaining a kernel estimation of the cdf in order to estimate marginal cdfs F_X and F_Y, respectively. We consider classical kernel estimation, transformed kernel estimation and double transformed kernel estimation. We describe these three methods to the specific case when marginals F_X and F_Y are the same.

Classical kernel estimation (CKE) of cdf F_X is obtained by integration of the classical kernel estimation of its pdf f_X. By means of a simple change of variable it follows that:

$$\hat{F}_X(x) = \int_{-\infty}^{x} \hat{f}_X(t)dt = \int_{-\infty}^{x} \frac{1}{nb}\sum_{i=1}^{n} k\left(\frac{t-X_i}{b}\right)dt$$
$$\frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\frac{x-X_i}{b}} k(t)dt = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{b}\right) \quad , \tag{5}$$

where $k(\cdot)$ is a pdf, which is known as kernel function. Very common kernels are Gaussian or Epanenchnikov kernels (see [10]). Function $K(\cdot)$ is the cdf of $k(\cdot)$. Parameter b is the bandwidth; it controls the smoothness of the function estimate. Silverman in [11] analyzes the statistical properties of (5).

Classical kernel estimation is not a good alternative when data are right skewed (see [6]). An alternative is transformed kernel estimation, that consists of transforming the data so that the transformed observations are symmetric. Authors in [12] propose to use a shifted power transformation family:

$$T_{(\lambda_1,\lambda_2)}(x) = \begin{cases} (x+\lambda_1)^{\lambda_2} sign(\lambda_2) if\ \lambda_2 \neq 0 \\ \ln(x+\lambda_1)^{\lambda_2} \qquad if\ \lambda_2 = 0 \end{cases}, \tag{5}$$

where $\lambda_1 \geq -\min(X_1,\ldots,X_n)$ and $\lambda_2 \leq 1$ for right skewed data. Transformed kernel estimation (TKE) of a cdf is:

$$\hat{F}_X(x) = \sum_{i=1}^{n}\frac{1}{n}\sum_{i=1}^{n} K\left(\frac{T_{(\lambda_1,\lambda_2)}(x) - T_{(\lambda_1,\lambda_2)}(X_i)}{b}\right). \tag{6}$$

Then, the transformed kernel estimation is a classical kernel estimation with transformed data ([12] describe a method to choose transformation parameters λ_1 and λ_2).

Double transformed kernel estimation (DTKE) needs two steps. First, a transformation of the data $T(X_i) = Z_i, i = 1,\ldots,n$, is chosen, where the transformed data have a distribution that is close to the Uniform $(0,1)$ distribution. Second, data are transformed again using the inverse of the Beta $(3,3)$ distribution with pdf and cdf:

$$m(x) = \frac{15}{16}\left(1-x^2\right)^2, -1 \leq x \leq 1$$

and

$$M(x) = \frac{3}{16}x^5 - \frac{5}{8}x^3 + \frac{15}{16}x + \frac{1}{2}.,$$

The resulting transformed data have a distribution that is close to the Beta$(3,3)$ distribution. This distribution can be estimated optimally using classical kernel estimation (see the discussion in [6]).

The double transformation kernel estimation (DTKE) is:

$$\hat{F}_X(x) = \sum_{i=1}^{n} \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{M^{-1}(T(x)) - M^{-1}(T(X_i))}{b} \right). \tag{6}$$

where T(x) is the generalized Champernowne cdf:

$$T(x) = \frac{(x+c)^{\gamma} - c^{\gamma}}{(x+c)^{\gamma} + (M+c)^{\gamma} - 2c^{\gamma}}$$

with parameters $\gamma, M > 0$ and $c \geq 0$ (authors in [13] describe a maximum pseudo-likelihood method to estimate parameters of the Champernowne distribution).

4 Value-at-Risk and Tail Value-at-Risk

Let $S = X + Y$ be the sum of two possibly dependent random variables X and Y. The Value-at-Risk of S with a confidence level α is:

$$VaR_{\alpha}(S) = \inf\{s, F_S(s) \geq \alpha\}.$$

and the Tail Value-at-Risk of S with a confidence level α is:

$$TVaR_{\alpha}(S) = E(S \mid S \geq VaR_{\alpha}(s)) = \frac{1}{1-\alpha} \int_{\alpha}^{1} VaR_{\alpha}(u) du \, .$$

Our goal in this section is to calculate the VaR and TVaR using Sarmanov copula to model dependency and the nonparametric approach to estimate marginal cdfs using the Monte Carlo method. The procedure is described here:

1. Estimate with non-parametric method the cdfs $\hat{F}_X$ and $\hat{F}_Y$.
2. Replace the cdf estimates in (2) in order to obtain the parameters of the copula that maximize likelihood.
3. Generate pairs (u_i, v_i) for $i = 1 \ldots r$ using (4), where r is the number of simulated pairs.
4. Solve $\hat{F}_X(X_i) = u_i$ and $\hat{F}_Y(Y_i) = v_i$ and obtain the simulated losses (X_i, Y_i).
5. Calculate $S_i = X_i + Y_i$ and estimate $VaR_{\alpha}(S)$ and $TVaR_{\alpha}(S)$ empirically once r repetitions are available.

5 Results

The data corresponds to a random sample of claims that were obtained from motor insurance accidents. An insurance company kindly gave us access these data. We have two costs: Cost1, contains the amount paid to the insured person to compensate for own damages to the vehicle and all other losses to third-parties damages, and Cost2, corresponds to the expenses related to medical treatments and hospitalization as a result of the accident (see [14] for more information on these data).

To estimate marginal cdfs with classical kernel estimation (CKE), transformed kernel estimation (TKE) and double transformed kernel estimation (DTKE) we use the Epanechnikov kernel and the bandwidth based on the asymptotic minimization of weighted integrated mean squared error (WISE, see [6]).

In Table 1 we summarize the values of the dependence parameters of the Sarmanov copula for the different kernel estimations for the marginal cdf. We note that there are significant differences between the estimated dependence parameters when we use CKE, TKE or DTKE to estimate the marginal cdfs. The log-likelihood column refers only to the dependence parameter estimation, that is not a full likelihood, which is why positive values are obtained. The log-likelihood shows that the best fit is obtained using DTKE to estimate marginal cdfs.

Table 1. Estimated parameters of Sarmanov copula for bivariate claims cost data

	w	Log-Likelihood*
CKE	9.53×10^{-9}	4.96227
TKE	109.05	29.68099
DTKE	0.99	87.86507

* This is not full likelihood, so values can be larger than zero

Tables 2 and 3 summarize the $VaR_\alpha(S)$ and $TVaR_\alpha(S)$ calculated for different confidence levels α, using a Monte Carlo simulation method. We use $r = 2000$ simulated samples. In Fig. 1 we plot the estimated $VaR_\alpha(S)$ together with the confidence interval at 95% for the empirical estimation.

To estimate the confidence intervals, we use the Bootstrap method. We generate samples with replacement with the same size of the original sample. This methodology allows us to obtain the intervals with different confidence levels.

Table 2. Estimated VaR with Sarmanov copula for bivariate claims cost data at tolerance levels 95%, 99% and 99.5%

VaR	0.95	0.99	0.995
CKE	7639.09	20582.40	24781.87
TKE	8622.01	48179.00	100784.01
DTKE	9610.28	24638.11	26338.82

Table 3. Estimate TVaR with Sarmanov copula for bivariate claims cost data at tolerance levels 95%, 99% and 99.5%

$TVaR$	0.95	0.99	0.995
CKE	47745.92	61114.87	81381.10
TKE	246990.90	326447.20	465581.20
DTKE	36907.52	46006.60	56690.85

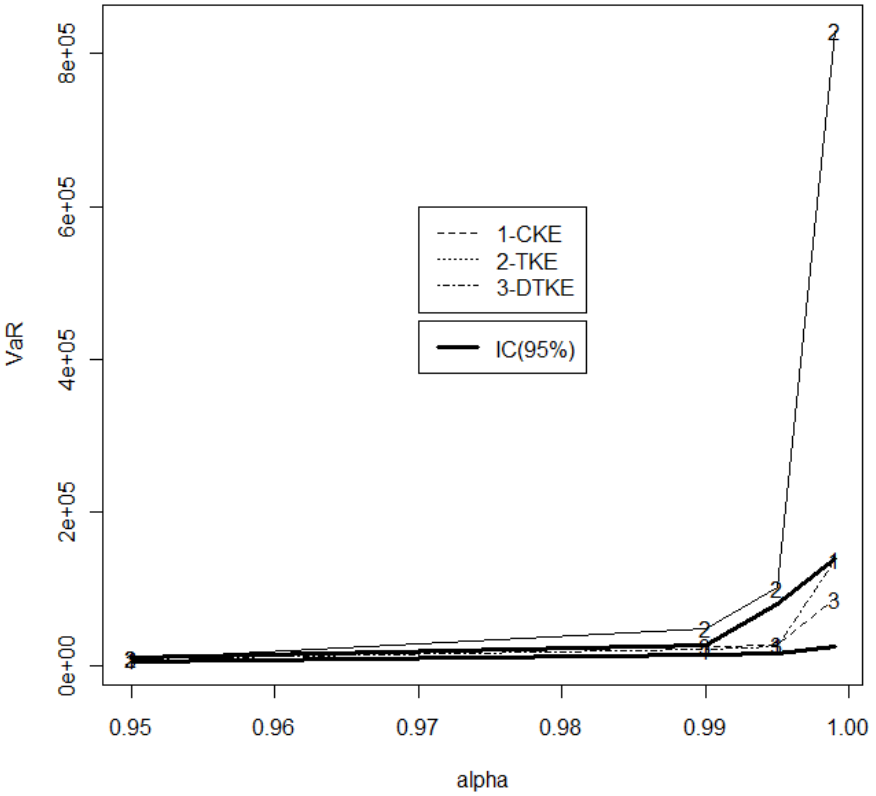

Fig. 1. Estimated VaR

The results show that using the DTKE we obtain the best results, inside confidence levels and ensuring balanced risk estimation, neither overestimated nor underestimated.

6 Conclusions

In this paper we present an example, using a random sample of claims from a real database, where, to estimate the risk of an aggregated loss, we mix copulas with kernel estimations. We show that double transformed kernel estimation (DTKE) of a

cdf can be a useful tool combined with copulas, which allows to estimate the VaR and the TVaR using Monte Carlo simulation method.

We propose to use the Sarmanov copula, which so far has not been used in the context of the quantifying risk. As we say in the introduction, in this copula the dependence structure is not separated strictly of the marginal distributions. We show how this fact can affect the risk estimate since the estimation of the marginal distribution affects significantly the goodness of fit of this copula. As principal future lines for research we want to analyze how kernel estimation can improve the goodness of fit of alternative and well known copulas.

References

[1] McNeil, A.J., Frey, R., Embrechts, P.: Quantitative Risk Management: Concept, Techniques and Tools. Princeton University Press (2005)
[2] Jorion, P.: Value at risk. The new Benchmark for managing Financial Risk. The McGraw-Hill Companies (2007)
[3] Bolancé, C., Guillén, M., Nielsen, J.P., Gustafsson, J.: Quantitative Operational Risk Models. CRC finance series. Chapman and Hall, New York (2012)
[4] Dhaene, J., Denuit, M., Goovarts, M.J., Kaas, R., Vyncke, D.: Risk measure and comonotonicity: A Review. Stochastic Models 22, 573–606 (2006)
[5] Dowd, K., Blake, D.: After VaR: The theory, estimation, and insurance applications of quantile-based risk measures. The Journal of Risk and Insurance 73, 193–229 (2006)
[6] Alemany, R., Bolancé, C., Guillén, M.: Non-parametric estimation of Value-at-Risk. Insurance: Mathematics and Economics 52, 255–262 (2013)
[7] Sarmanov, O.V.: Generalized normal correlation and two-dimensional Fréchet. Soviet Mathematics. Doklady. 25, 1207–1222 (1996)
[8] Lee, M.L.T.: Properties and aplications of the Sarmanov family of bivariate distributions. J. Theory and Methods 25, 1207–1222 (1996)
[9] Nelson, R.B.: An Introduction to Copulas, 2nd edn. Springer, Portland (2006)
[10] Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Chapman and Hall, London (1986)
[11] Azzalini, A.: A note on the estimation of a distribution function and quantiles by a kernel method. Biometrika 68, 326–328 (1981)
[12] Bolancé, C., Guillén, M., Nielsen, J.P.: Kernel density of actuarial loss functions. Insurance: Mathematics and Economics 32, 19–36 (2003)
[13] Buch-Larsen, T., Guillén, M., Nielsen, J., Bolancé, C.: Kernel density estimation for heavy-tailed distributions using the Champernowne transformation. Statistics 39, 503–518 (2005)
[14] Bolancé, C., Guillén, M., Pelican, E., Vernic, R.: Skewed bivariate models and nonparametric estimation for CTE risk measure. Insurance: Mathematics and Economics 43, 386–393 (2008)