Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 104 (2017) 572 – 577

Procedia Computer Science

ICTE 2016, December 2016, Riga, Latvia

Algorithms of the Copula Fit to the Nonlinear Processes in the Utility Industry

Andrejs Matvejevs[a,]*, Jegors Fjodorovs[a], Anatoliy Malyarenko[b]

[a]*Riga Technical University,Kalku 1, Riga, LV1658, Latvia*
[b]*Mälardalen University, Västerås, Sweden*

Abstract

Our research studies the construction and estimation of copula-based semi parametric Markov model for the processes, which involved in water flows in the hydro plants. As a rule analyzing the dependence structure of stationary time series regressive models defined by invariant marginal distributions and copula functions that capture the temporal dependence of the processes is considered. This permits to separate out the temporal dependence (such as tail dependence) from the marginal behavior (such as fat tails) of a time series. Dealing with utility company data we have found the best copula describing data - Gumbel copula. As a result constructed algorithm was used for an imitation of low probability events (in a hydro power industry) and predictions.

Peer-review under responsibility of organizing committee of the scientific committee of the international conference; ICTE 2016

Keywords: Copula; Diffusion processes; Time series; Semi parametric regressions

1. Introduction

Our research studies the construction and estimation of copula-based semi parametric Markov model for the processes, which involved in water flows in the hydro plants.
Copulas became popular in the finance and insurance community in the past years, where modeling and estimating the dependence structure between several univariate times series are of great interest; see Frees and Valdez[1]and Embrechts et al.[2] for reviews.

* Corresponding author. Tel.: +371 26015121.
E-mail address: Andrejs.Matvejevs@rtu.lv

Peer-review under responsibility of organizing committee of the scientific committee of the international conference; ICTE 2016
doi:10.1016/j.procs.2017.01.174

A copula function is a multivariate distribution function with standard uniform marginals. By Sklar's[3] theorem, one can always model any multivariate distribution by modeling its marginal distributions and its copula function separately, where the copula captures all the scale-free dependence in the multivariate distribution.

The central result of this theorem, which states that any continuous N-dimensional cumulative distribution function F, evaluated at point $x = (x_1, \dots, x_n)$ can be represented as

$$F(x) = C(F_1(x_1), \dots, F_n(x_n)), \tag{1}$$

where C is called a copula function and $F_i(x_i)$, $i = 1, \dots, n$ are the marginal distributions. The use of copulas therefore splits a complicated problem (finding a multivariate distribution) into two simpler tasks. The first task is to model the univariate marginal distributions and the second task is finding a copula that summarizes the dependence structure between them.

The possibility of identifying nonlinear time series using nonparametric estimates of the conditional mean and conditional variance were studied in many papers[4]. As a rule analyzing the dependence structure of stationary time series $\{x_t, t \in Z\}$ regressive models defined by invariant marginal distributions and copula functions that capture the temporal dependence of the processes. As it indicated[4] this permits to separate out the temporal dependence (such as tail dependence) from the marginal behavior (such as fat tails) of a time series. One more advantage of this type regressive approach is a possibility to apply probabilistic limit theorems for transition from deference equations to continuous time stochastic differential equations[5,6]. In our paper, we also study a class of copula-based semi parametric stationary Markov models in a form of scalar difference equation

$$t \in Z : X_t = f(X_{t-1}) + g(X_{t-1})\xi_t, \tag{1a}$$

where $\{\xi_t, t \in Z\}$ is i.i.d., N(0; 1). Regressions (1a) are high-usage equations for simulation and parameter estimation of stochastic volatility models ([2]). But, unfortunately defined by (1a) Markov chain has incompact phase space that complicates an application of probabilistic limit theorem. Copula approach helps to simplify asymptotic analysis of (1a). Let us remember that to construct a copula *C(u; v)* for pair $\{X_{t-1}, X_t\}$ from (1a) one should find a marginal invariant distribution *F(x)* for X_t and to substitute this in joint distribution function $H(x, y) = P(X_{t-1} \le x, X_t \le y)$, that is, $C(u, v) = H(F^{-1}(u), F^{-1}(v))$ and $H(x, y) = C(F(x), F(y))$. After a substitution $U_t = F(X_t)$ in equation (1) for a further diffusion approximation one can write a difference equation in a same form like (1a):

$$t \in Z : U_t = \varphi(U_{t-1}) + \psi(U_{t-1})\xi_t. \tag{2}$$

But now this equation defines Markov chain on the compact [0, 1]. This makes easier formulate construction for transition probability and further estimators of functions $\hat{f}(u)$ and $\hat{g}(u)$. After diffusion approximation of (2) one can make inverse substitution and derive stochastic differential equation as diffusion approximation for (1a).

We found that the best copula describing data is Gumbel copula. As a result constructed equation (1a) was used for low probability events imitation (hydro power industry) and predictions.

The paper is structured as follows. Section 2 describes our approach. In Section 3 we report our results for the data Section 4 concludes and discusses several possible avenues of future research.

2. Evaluation of parameters for the semi parametric regression model

Copula based semi parametric models are characterized by conditional heteroscedasticity and have been often used in modeling the variability of statistical data. The basic idea was to apply a local linear regression to the squared residuals for finding the unknown functions f and g[5,7].

Our methodology builds on the finding conditional expectation of the first and second order.

Let $\{Y_t\}$ be a stationary Markov process of order 1 with continuous state space. Then its probabilistic properties are completely determined by the joint distribution function of $\{Y_{t-1}\}$ and $\{Y_t\}$. For the determination of the copula based model we should use Markov model in the scalar difference equation in the form (1) with a small parameter ε. And our goal reduced to the estimation of conditional moments, which will be our base regression model parameters:

$$g(X_{t-1},\varepsilon) \text{ and } f(X_{t-1},\varepsilon). \tag{3}$$

As was mentioned above it is not easy task, especially this representation complicates an application of probabilistic limit theorem. That is why; if we have stationary distribution our suggestion is to find parameters through Markov chain using copula approach.

And due to persistence of the small parameter ε, we can rewrite our expression:

$$t \in Z : U_t = U_{t-1} + \varepsilon f(U_{t-1},\varepsilon) + \varepsilon g(U_{t-1},\varepsilon)\xi_t$$

$$f(U_{t-1},\varepsilon) = E(U_t \mid U_{t-1} = u) \tag{3a}$$

$$g(U_{t-1},\varepsilon) = E((U_{t-1} - f(U_{t-1},\varepsilon))^2 \mid U_{t-1} = u). \tag{4}$$

After conditional expectations of (3a) and (4) evaluation one can make inverse substitution and derive stochastic differential equation as diffusion approximation for the base semi parametric model (1a). Of course, our algorithm works only if inverse function exists. For example, Gamble copula, which don't have standard inverse function.

Now we derived a tool for model (1a) parameters evaluation. For describing our idea briefly, let's take a look in the next section how works our algorithm with the true market data.

3. Practical approach of the proposed algorithm

We'll analyze a historical observations of the equipment parameter (sample Y, please see Fig. 1). We have daily data from 31.12.2000 till 31.12.2015. As a result of successful operations of the equipment we are interested in a stable, low volatility process, but as it is in a real life, depending on weather conditions parameter values may vary significantly. That is why our idea to get predictions for significant deviations of the observed values in the future. Our main idea to set limit for allowed deviation and find an algorithm for finding distribution of the process which reach this level. It is clear, we are dealing with heteroskedastic process and using first lag of the observation, i.e. we can skip another factors which can involve this equipment stability and use just time series observations, we can use copula densities and build semi parametric model.

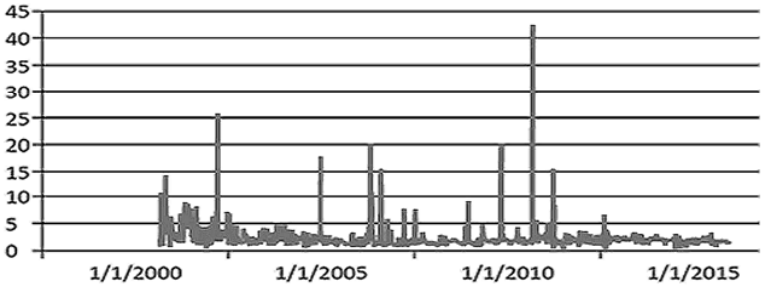

Fig. 1. Historical parameter values (Y) of the utility company equipment.

An easiest way of parameters estimating of the semi regressive model for the time series would be to hold the algorithm:

- Find marginal distributions for the observations of the equipment parameter
- Using marginal distribution, calculate U_t points which is R[0,1] (uniform)
- Build scatter plot for (U_{t-1}, U_t)
- Make several statistical tests to find the suited distribution of data
- Taking into account scatter plot and distribution of data try to choose copula from existing class or build your own copula, if you know marginal distributions
- Test copula consistency to data (for example, AIC and BIC, Kolmogorov distance etc.)
- Find regression parameters

Using Matlab program we have built scatter plots for Y transformed into uniform distribution (R[0,1]) and non transformed data.

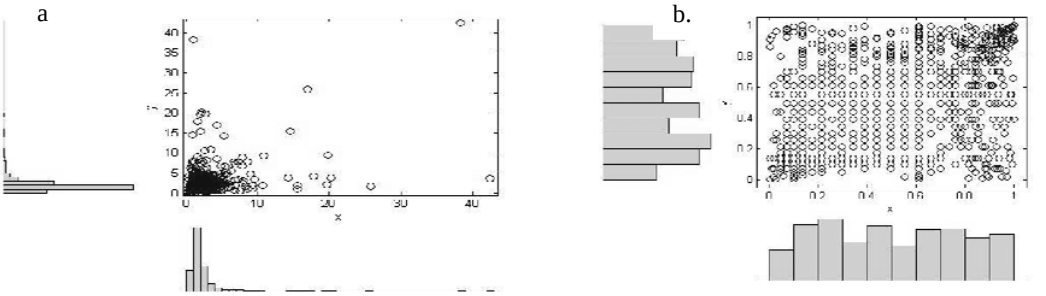

Fig. 2. (a) Scatter plot for non-transformed sample data; (b) Scatter plot for transformed into R[0,1] Y data.

As we see in the Fig. 2a time series Y has outliers. This make difficult to construct marginal distributions. Based on the Kolmogorov – Smirnov test we tried different assumptions about marginal distributions and the best fit was mixed of the exponential and the uniform distributions:

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, \; x < T \\ H(x - T) + e^{-\lambda T}, \; T_1 < x < T \end{cases},$$

$$H = \frac{e^{-\lambda T_1}}{T - T_1},$$

where T is a size of sample and T_1 is size of a sample without outliers.

Basically, taking into account margins we transformed into uniform distribution (R[0,1]) our observations. An important issue faced by an applied researcher interested in using the class of semi parametric copula-based time series models is the choice of an appropriate parametric copula. In different papers Chen et al.[8] propose two simple tests for the correct specification of a parametric copula in the context of modeling the contemporaneous dependence between several univariate time series and of the innovations of univariate GARCH models used to filter each univariate time series (2) Chen and Fan[9] establish pseudo-likelihood ratio tests for selection of parametric copula models for multivariate i.i.d. observations under copula misspecification[4]. But our suggestion is simpler – we can choose the best copula fit using AIC and BIC criteria or using Kolmogorov –Smirnov test for data distribution. We take for copula comparisons - Kolmogorov-Smirnov (KS) test (see Table 1).

$$D_{KS} = \max_{i,j} \left| C_n(U_{1,i}, U_{2,j}) - C_\theta(U_{1,i}, U_{2,j}) \right|$$

Table 1. Kolmogorov – Smirnov test (distance) for Y data.

Copula	KS value
Gumbel copula	0.67
Frank copula	0.65
Normal copula	0.18
T	0.7

Taking into account KS test results we should choose Normal copula for further model estimation. But Normal copula leads us to linear dependence between random variable. But in our research we are interested in rare jumps of the equipment parameter values. For this purpose we can take more tail dependence case – Gumbel copula and based on this copula density derive semi parametric regression parameters:

$$C(U_{t+1}, U_t) = \exp\left\{\left((-\ln U_{t+1})^{\theta} + (-\ln U_t)^{\theta}\right)^{1/\theta}\right\}, \quad \theta = 1{,}2847 \tag{5}$$

And insert expression (5) into conditional expectation, we get our parameters:

$$E(U_{t+1} \mid U_t = u) = \int_0^1 u_{t+1}\, dF_{u_{t+1}|u_t}(u) = \int_0^1 u_{t+1}\, p(u_{t+1} \mid u_t)\, du_{t+1} = \int_0^1 u_{t+1} \frac{\partial C(u_{t+1}, u_t)}{\partial u_{t+1} \partial u_t} du_{t+1} = \int_0^1 u_{t+1}\, c(u_{t+1}, u_t)\, du_{t+1} \tag{6}$$

$$g(U_{t+1} \mid U_t = u) = E\left((U_{t+1} - f(U_t))^2 \mid U_t = u\right) = \int_0^1 (U_{t+1} - f(U_t))^2\, c(U_{t+1}, U_t)\, dU_{t+1} \tag{7}$$

It is impossible to solve analytically (6) and (7) expressions. But numerically it is doable for example in the Matab. For the Gumbel copula we can use inverse transformation with the aim to return to our base equation (1). Of course, if we want use this model in practice, it is crucial to compare different class models, which could be suitable for this data. Finally, we have depicted possible algorithm for constructing semi parametric copula based regressions and find solutions for modelling processes with heteroskedastic nature. Proposed algorithm allows us to make imitations of the process and find distribution of time when process reaches certain border. This is very critical in a utility industry for making special preparations before equipment may go out of order. In the next sections, there is example of process Y imitation based on Gumbel copula:

- Construct marginal distributions for data Y_t
- Find copula and it parameters
- Estimate semi parametric regression model via copula
- Construct iteration procedure in points t_n with small parameter $h = 0.01$
- Find distribution of time $\tau(x, \Gamma)$ to reach $X_n = \Gamma$ via Monte Carlo imitation
- Make iterations of the 4-th step N times until $y(t_n) \geq F(\Gamma)$ and remember number $n^{(k)}$ (after every iteration)
- Construct histogram of the $\{n^{(k)}, k = 1, ..., N\}$ and find distribution

4. Conclusions and discussions of the proposed algorithm

Having built algorithm for the constructions copula based regressions and taking into account process imitation procedure (steps 1-4) we have modeled process Y via Gumbel copula based semi parametric regression (see in the Fig. 3). Our model imitation results graphically closed to time series Y values. Basically, our imitations react on volatility fluctuations. This gives possibility to use this model for evaluation border distributions.

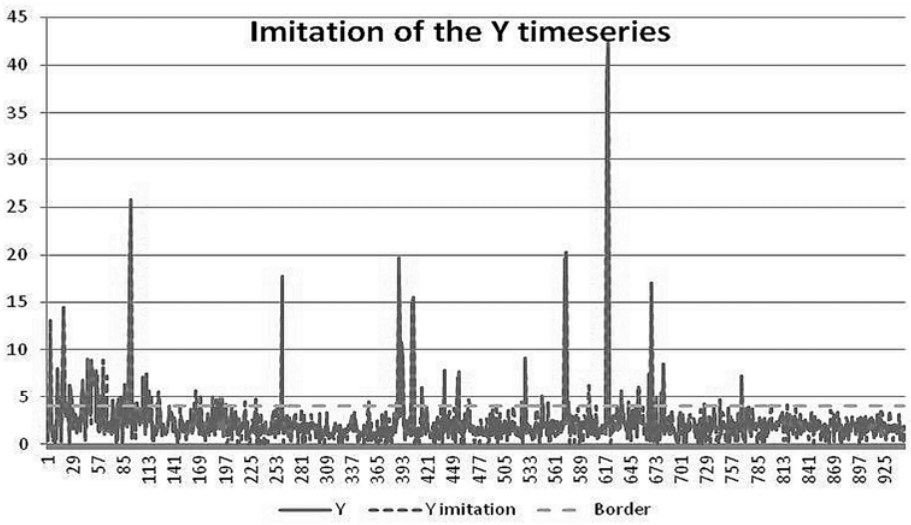

Fig. 3. Historical and modeled values (Y) of the utility company equipment.

But if we deal with copulas we should not skip some facts. For example, it is not easy to say which parametric copula best fits a given dataset, since some copulas may fit better near the center and other near the tails and many copulas do not have moments that are directly related to the Pearson correlation, it is difficult to compare financial models based on correlation.

References

1. Darsow W, Nguyen B, Olsen E. Copulas and Markov processes. *Illinois Journal of Mathematics*. 36; 1992. p. 600–642.
2. Joe H. Multivariate Models and Dependence Concepts. Chapman & Hall/CRC; 1997.
3. Frees EW, Valdez EA. Understanding relationships using copulas. *North American Actuarial Journal*. 2; 1998. p. 1–25.
4. Nelson DB. ARCH models as diffusion approximations. *Journal of Econometrics*. 7 (38); 1990. 441.
5. Chen X, Fan Y. Estimation of copula-based semiparametric time series models. *Journal of Econometrics*; 2006.
6. Ait-Sahalia Y, Kimmel R. Maximum likelihood estimation of stochastic volatility models. *Journal of Financial Economics*; 2007.
7. Fjodorovs J, Matvejevs A. Copula Based Semiparametric Regressive Models. *Journal of Applied Mathematics*. Vol. V; 2012. p.241-248.
8. Chen X, Hansen LP, Carrasco M. Nonlinearity and temporal dependence. Working Paper, University of Chicago; 1998.
9. Chen X, Fan Y. Pseudo-likelihood ratio tests for model selection in semiparametric multivariate copula models. *Canadian Journal of Statistics*; 2004.

Andrejs Matvejevs has graduated from Riga Technical University, Faculty of Computer Science and Information Technology. He received his Doctoral Degree in 1989 and became an Associate Professor at Riga Technical University in 2000 and a Full Professor in 2005. He has made the most significant contribution to the field of actuarial mathematics. For more than 30 years he has taught at Riga Technical University and Riga International College of Business Administration, Latvia. His current professional research interests include applications of Markov chains to actuarial technologies: mathematics of finance and security portfolio. He is the author of about 80 scientific publications. Contact him at andrejs.matvejevs@rtu.lv.