UDC 519.246.5

Estimation of Heavy Tail Dependence Based on Copulas for the Precipitation Analysis

Leonid A. Sevastianov*, Nikita D. Rassakhan†, Eugeny Yu. Shchetinin‡

* *Department of Applied Probability and Informatics,*
Peoples' Friendship University of Russia (RUDN University),
6 Miklukho-Maklaya str., Moscow, 117198, Russia
† *Department of Applied Mathematics*
Moscow State Technology University "STANKIN"
3a Vadkovsky Ln., Moscow, 127055, Russian Federation
‡ *Financial University under the Government of the Russian Federation*
Leningradsky pr. 49, 111123, Moscow, Russian Federation

Email: sevastianov_la@rudn.university, rassahan@gmail.com, riviera-molto@mail.ru

Consideration of tail dependence is a very important part of risk analysis in many applied sciences that is measured in order to estimate the risk of simultaneous extreme events. Usually the tail dependence coefficient is the measurement in question. Pearson correlation coefficient unfortunately is not a suitable measure for estimating dependencies between two quantities in the context of simultaneous occurrence of extreme events when these events are of interest for the researcher because it takes extreme events into account with the same weight as it takes "normal" events although dependence of extreme values may slightly differ.

Present work emphasizes the importance of taking into account tail dependencies in bivariate statistical analysis using copulas. Due to increasing frequency of environmental cataclysms the issue of analyzing risks (e.g. economic losses) and their consequences comes to the fore. Moreover, researchers should take into consideration consequences of their joint occurrence. Three non-parametric estimators of tail dependence coefficients were compared in order to estimate correlation between daily cumulative rainfall totals recorded in central European part of Russia. The majority of existing estimators depends on threshold k and thus there is a trade-off between variance and bias during the calculation of the best value for k. For balancing an algorithm is presented that is based on using moving average filter and then searching the "stable" part of tail dependence coefficient. Estimate of tail dependence coefficient is assumed to be equal to mean value on the "stable" part.

Key words and phrases: extreme value theory, spatial modelling, extreme precipitation, spatial structures of statistical dependence, tail dependence coefficient.

In: K. E. Samouylov, L. A. Sevastianov, D. S. Kulyabov (eds.): Selected Papers of the VIII Conference "Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems", Moscow, Russia, 20-Apr-2018, published at http://ceur-ws.org

1. Introduction

One of the most important parts of multivariate extreme value analysis is the study of extremal dependencies [13]; basically tail dependence coefficient is used for this purpose. For bivariate vector (X_1, X_2) upper tail dependence coefficient has the following form [5]:

$$\lambda_U = \lim P\left(F_1(X_1) > t | F_2(X_2) > t\right), \quad t \to 1^-, \tag{1}$$

where F_1, F_2 are distribution functions of random variables X_1, X_2 respectively, $0 < t \leqslant 1$ is the threshold.

Using the copula function [8] equation (1) can be written in alternative form [15, 18]:

$$\lambda_U = \lim_{t \to 1^-} \frac{1 - 2t + C(t,t)}{1-t}. \tag{2}$$

2. Main section

Tail dependence coefficient estimation methods are essential analysis tools for extremal structures that are studied in this paper on precipitation data. Onward we will describe some of them. Foremost such estimators are non-parametric estimators based on empirical copula $C^{(n)}(u,v)$ concept [4], [16] with $F_{(n)}(\cdot)$ as empirical distribution function.

Let $\left(X_1^{(1)}, X_2^{(1)}\right), \ldots, \left(X_1^{(n)}, X_2^{(n)}\right)$ be independent identically distributed copies of bivariate random vector (X_1, X_2). Using their joint distribution function [12] and equation (2) an estimator for upper tail dependence (1) coefficient can be derived [7]:

$$\hat{\lambda}^{SEC} \equiv \hat{\lambda}^{SEC}(k) = 2 - \frac{1 - \hat{C}\left(1 - \frac{k}{n}, 1 - \frac{k}{n}\right)}{\frac{k}{n}}, \quad 1 \leqslant k < n. \tag{3}$$

Then in respect that $\log(1-t) \sim -t$, $t \approx 0$ next estimator can be obtained:

$$\hat{\lambda}^{LOG} \equiv \hat{\lambda}^{LOG}(k) = 2 - \frac{\log \hat{C}\left(1 - \frac{k}{n}, 1 - \frac{k}{n}\right)}{\log\left(1 - \frac{k}{n}\right)}, \quad 1 \leqslant k < n. \tag{4}$$

where $\hat{C}$ denotes empirical copula.

Note that both estimators depend on choice of threshold k and thereafter k^{th} order statistic [19]. It is very important to choose the right value for k which is not an easy task due to the trade-off between variance and bias.

Another estimator for upper tail dependence coefficient is suggested in works [9, 10]:

$$\hat{\lambda}^{CFG} = 2 - 2\exp\left[\frac{1}{n}\sum_{i=1}^{n} \log\left\{\frac{\sqrt{\log\frac{1}{\hat{F}_1\left(X_1^{(i)}\right)} log\frac{1}{\hat{F}_2\left(X_2^{(i)}\right)}}}{\log\frac{1}{\max\left(\hat{F}_1\left(X_1^{(i)}\right), \hat{F}_2\left(X_2^{(i)}\right)\right)^2}}\right\}\right]. \tag{5}$$

Main advantage of this equation is that $\hat{\lambda}$ doesn't depend on k. However, copula $C(X_1, X_2)$ must be well approximated with extreme-value copulas for correctness of the estimator.

It follows from equations (3), (4) that estimators depend on choice of threshold k which is determined by balancing variance and bias for estimator according to stability theorem for λ_U [8]. Increasing the value of k leads to reduction of bias and increase in variance; it goes the same the other way around. For big enough data sample size n balance between bias and variance is described by the "stable" part of λ_U plot. An algorithm for finding this "stable" part is presented in paper [6]:

1. Empirical estimation is smoothed with moving average filter window size of which is equal to $b = int(0.05n)$. Sequence $\hat{\lambda}_1, \ldots, \hat{\lambda}_{n-2b}$ is obtained as a result.
2. Vector $\left(\hat{\lambda}_k, \ldots, \hat{\lambda}_{k+m-1}\right)$, where $k = 1, \ldots, n-2b+m-1$, $m = int\left(\sqrt{n-2b}\right)$ can be made from the sequence $\hat{\lambda}_1, \ldots, \hat{\lambda}_{n-2b}$ by a sequential search.
3. If the current vector satisfies

$$\sum_{i=k+1}^{k+m-1} |\bar{\lambda}_i - \bar{\lambda}_k| \leqslant 2\sigma,$$

where σ is standard deviation of $\hat{\lambda}_1, \ldots, \hat{\lambda}_{n-2b}$ then the final expression for λ_U takes the form of

$$\lambda_U = \frac{1}{m} \sum_{i=1}^{m} \bar{\lambda}_{k+i-1}.$$

If the condition is not satisfied after sequential searching, then $\lambda_U = 0$.
Example of algorithm realisation using R language is presented below:

```
lambda_sec <-
  (1/k * rank_sum(msk_rank, spb_rank, length(df$Msk), k))
lambda_sec2 <- 2 - (1/k * rank_sum2(msk_rank, mhzsk_rank,
                          length(df$Msk), k))
rank_sum <- function (rank1, rank2, lgth, k){ a = 0
  for (i in 1:lgth){
    a = a + ifelse( (rank1[i] > lgth - k)
        & (rank2[i] > lgth-k), 1, 0) } return (a) }
rank_sum2 <- function (rank1, rank2, lgth, k){
  a = 0;  for (i in 1:lgth){
    a = a + ifelse( (rank1[i] > lgth - k)
        | (rank2[i] > lgth-k), 1, 0)
  } return (a) }
b = trunc (0.05 * length(df$Msk))
m = trunc (sqrt(length(df$Msk-2*b)))
wow <- 0; wow2 <-0; sssuum <- 0
for (i in 1:(length(ls_ma_na)-2*b+m-1)){
    rw <- ls_ma_na[i:(i+m)]
    for (l in 1:m) sssuum <- sssuum + abs(rw[l]-rw[1])
    if (sd(rw)>= sssuum/m) {wow <- i
                            wow2 <- mean(rw)
                            break} sssuum <- 0}
```

In this study the precipitation data of the All-Russian Research Institute of Hydrometeorological Information — the World Data Center of the Russian Federation is used, which contains data on daily precipitation in 11 cities of the European part of Russia [11]. The data is freely available on the website `http://aisori.meteo.ru/ClimateR` and is represented by a set of tables (a separate table for each city); each table contains daily rainfall value for the period 1966–2016 years.

Implementation of this algorithm is presented in Fig. 1 [2]. Both plots are using monthly maximum of precipitation in Moscow and Kostroma to evaluate upper tail dependence coefficient using estimators $\hat{\lambda}^{SEC}$ (left) and $\hat{\lambda}^{LOG}$ (right). Black line

corresponds to $\hat{\lambda}(k)$; blue smooth line is $\hat{\lambda}(k)$ after applying moving average filter to it. Pink transparent plateau is the resulting value for $\hat{\lambda}_U$ where placement of plateu corresponds to indexes of vector $\hat{\lambda}_k, \ldots, \hat{\lambda}_{k+m-1}$ from the algorithm above.

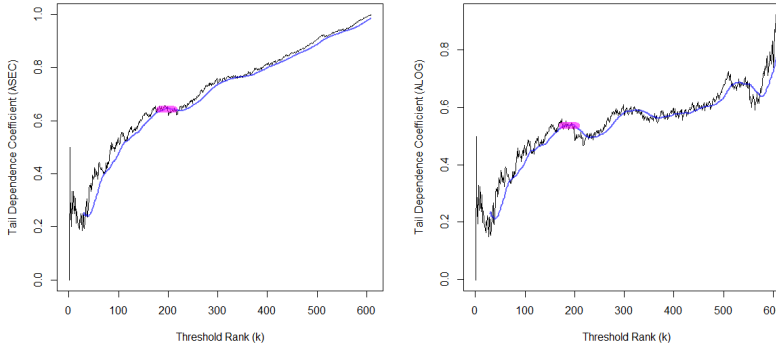

Figure 1. Implementation of the algorithm for finding "stable" part of TDC for $\hat{\lambda}^{SEC}$ (left) and $\hat{\lambda}^{LOG}$ (right). Moscow and Kostorma were used as a pair of cities from the area under study

All three estimators (3), (4), (5) for upper tail dependence coefficient were calculated for 55 pairs of 11 cities under study [1,3]. Furthermore, Pearson's correlation coefficient (PCC) was also calculated in order to compare it with estimators. Results for some of the pairs are represented by Table 1. As we can see, PCC is quite different from all other estimators in some cases (the Sp. Petersburg — N. Novgorod pair as an example) but it actually is close enough to at least one of the estimators for the most pairs.

Table 1
Values for estimators of λ_U and Pearson's correlation coefficient calculated for some pairs of cities in the European part of Russia

Pair of cities	$\hat{\lambda}^{SEC}$	$\hat{\lambda}^{LOG}$	$\hat{\lambda}^{CFG}$	PCC
Moscow – Kolomna	0.6417793	0.5358919	0.5210492	0.6021248
Kolomna – Ryazan	0.5561837	0.4623016	0.5022938	0.5509277
Pskov – Smolensk	0.54967	0.437089	0.3722651	0.4314
Kostroma – N.Novgorod	0.6346736	0.415878	0.3856968	0.4766541
Bryansk – Mozhaisk	0.6846937	0.1764084	0.433269	0.6021248
St. Petersburg – N.Novgorod	0.1470922	0.0770384	0.2608262	0.3353886
Smolensk – Moscow	0.488586	0.3740484	0.3968146	0.3952423
St. Petersburg – Pskov	0.5338348	0.4111191	0.4124836	0.4867774
N.Novgorod – Tambov	0.2271567	0.1740808	0.3475754	0.3898402
Pskov – Kostroma	0.5107715	0.4133031	0.3863028	0.5022300

Finally the attempt to find the correlation between estimators for upper tail dependence and the distance between cities under study was made [14]. Scatterplots for four values compared in Table 1 were plotted as a result; they are presented in Fig. 2.

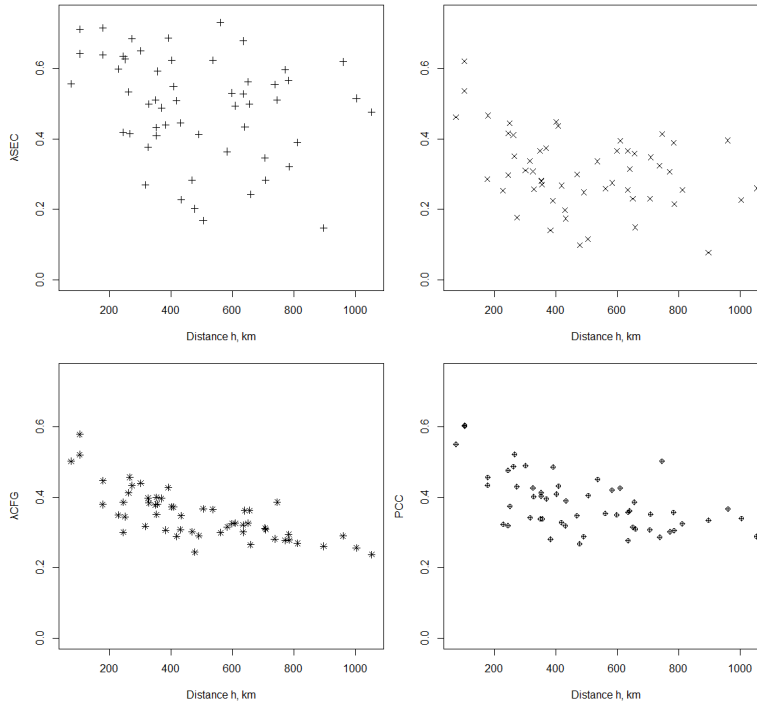

Figure 2. Comparison of 4 estimators and their dependence on the distance h between observation points: $\hat{\lambda}^{SEC}$ — upper left, $\hat{\lambda}^{LOG}$ — upper right, $\hat{\lambda}^{CFG}$ — lower left, PCC — lower right

It is obvious that there must be an inverse relation between the distance h and dependence estimators. Therefore $\hat{\lambda}^{SEC}$ is a bad estimator for the problem under study. Three other estimators ($\hat{\lambda}^{LOG}, \hat{\lambda}^{CFG}$ and PCC) show roughly the same with some correction, which is why they are considered to be more trustworthy. So it is proposed to take the average of $\hat{\lambda}^{LOG}$ and $\hat{\lambda}^{CFG}$ or just $\hat{\lambda}^{CFG}$ as the resulting estimator for λ_U.

3. Conclusions

This paper highlights the importance of taking into account the tail dependence coefficient in the context of multivariate frequency analysis using copulas. The three following nonparametric estimators($\hat{\lambda}^{SEC}, \hat{\lambda}^{LOG}, \hat{\lambda}^{CFG}$) have been compared. The aim of this comparison was to choose the best estimator in the context of our application [20]. No estimator works in every case yet some of them show poor performance thus they

need to be excluded. It is therefore important to pursue research in this field to get the right estimation for λ_U based on values of $\hat{\lambda}^{SEC}$, $\hat{\lambda}^{LOG}$ and $\hat{\lambda}^{CFG}$.

Most non-parametric estimators have to deal with the choice of the number k of order statistics to be considered in the production of an estimate. This is not an easy task since it requires a trade-off between variance and bias (small values of k cause large variance and large values of k increase the bias).

Frahm et al. [6] introduced a simple algorithm to find the optimal threshold k in order to estimate λ_U. Since this very simple algorithm revealed some potential, we intend to develop this idea further. $\hat{\lambda}^{CFG}$ is considered to be the best estimator out of $\hat{\lambda}^{SEC}$, $\hat{\lambda}^{LOG}$ and $\hat{\lambda}^{CFG}$, it looks like PCC with some corrections.

References

1. M. Ferreira, Nonparametric estimation of the Tail-dependence coefficient, Revstat **11** (1) (2013) 1–16.
2. A. Poulin, D. Huard, A.-C. Favre, S. Pugin, Importance of Tail Dependence in Bivariate Frequency Analysis, Journal of hydrologic engineering, 2007.
3. M. Ferreira, S. Silva, An Analysis of a Heuristic Procedure to Evaluate Tail (in)dependence, Journal of Probability and Statistics 2014 (2014), Article ID 913621.
4. M. Sibuya, Bivariate extreme statistics, I. Annals of the Institute of Statistical Mathematics **11** (2) (1959) 195–210.
5. G. Draisma, H. Drees, A. Ferreira, L. De Haan, Bivariate tail estimation: dependence in asymptotic independence. Bernoulli **10** (2) (2004) 251–280.
6. G. Frahm, M. Junker, R. Schmidt, Estimating the tail-dependence coefficient: Properties and pitfalls. Insurance: Mathematics and Economics **37** (1) (2005) 80–100.
7. H. Joe, R.L. Smith, I. Weissman, Bivariate Threshold Methods for Extremes, Journal of the Royal Statistical Society. Series B **54** (1) (1992) 171–183.
8. S. Coles, J. Heffernan, J. Tawn, Dependence Measures for Extreme Value Analyses, Extremes **2** (4) (1999) 339–365.
9. P. Caperaa, A.-L. Fougeres, C. Genest, A nonparametric estimation procedure for bivariate extreme value copulas, Biometrika **84** (5) (1997) 567–577.
10. R. Schmidt, U. Stadtmuller, Non-parametric Estimation of Tail Dependence, Scandinavian Journal of Statistics **33** (2) (2006) 307–335.
11. E. Yu. Shchetinin, N. D. Rassakhan. Modeling of Extreme Precipitation Fields on the Territory of the European Part of Russia, RUDN Journal of Mathematics, Information Sciences and Physics **26** (1) (2018) 74–83. doi:10.22363/2312-9735-2018-26-1-74-83
12. A. Juri, M. V. Wüthrich, Tail dependence from a distributional point of view, Extremes **6** (3) (2004) 213—246.
13. J. Beirlant, Y. Goegebeur, J. Segers, J.L. Teugels, Statistics of Extremes: Theory and Applications, Wiley, 2004.
14. M. de Carvalho, A. Ramos, Bivariate Extreme Statistics, II, Revstat **10** (1) (2012) 83–107.
15. S. M. Berman, Convergence to bivariate limiting extreme value distributions, Annals of the Institute of Statistical Mathematics **13** (1) (1961) 217–223.
16. P. Hall, N. Tajvidi, Distribution and dependence-function estimation for bivariate extreme-value distributions. Bernoulli **6** (5) (2000) 835-–844.
17. J. Dobric, F. Schmid, Nonparametric estimation of the lower tail dependence in Bivariate Couplas, Journal of Applied Statistics **32** (4) (2005) 387–407.
18. M. Falk, R. Reiss, Efficient Estimation of the Canonical Dependence Function, Extremes **6** (1) (2003) 61—82.
19. J. Galambos, Order Statistics of Samples from Multivariate Distributions, Journal of the American Statistical Association **70** (351) (1975) 674–680.
20. R.-D. Reiss, M. Thomas, Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields, Birkhauser, Basel, 2007.