Lund Institute of Technology   Centre for Mathematical Sciences   Mathematical Statistics

---

QUANTITATIVE RISK MANAGEMENT USING COPULAS, FMSN65/MASM33

COMPUTER ASSIGNMENT 3

This assignment is a compulsory part of the course and will be carried out by each group consisting of two students. The report (a PDF file) should be uploaded at the home page of the course in *Canvas* by the corresponding deadline. A text file (with txt-extension in the name) containing theR commands or programs used in each part of the assignment has to be also uploaded separately.

Bivariate extreme value analysis

Exercise 1: Some theoretical questions related to bivariate extreme value distributions

Answer the following questions:

1. Present in detail the characterization of bivariate extreme value distributions with unit Fréchet margins.
2. What are the properties of the dependence function?
3. Write the formula for extreme value copula and give at least three examples of parametric bivariate extreme value copulas. How can one check that a given copula is an extreme value copula?
4. Explain in detail how one can construct a bivariate extreme value distribution with arbitrary univariate extreme value margins?
5. Derive the formula for Pickand's estimator of the dependence function.

Getting started with the computational part:

If you plan to carry out the assignment on your own computer, you need to download and install `R` from http://www.r-project.org.

If you want to carry out the assignment in the computer room MH:230, you just need to start one of the PCs first. Then choose the latest version of `R` from the `Start` menu.

In this assignment you need to use `evd`, `SimCop`, `copula` and `mgpd` packages. It is strongly recommended that you download the corresponding PDF files of the manuals from cran.r-project.org to use as a reference in the assignment. Type `library(evd)`, `library(SimCop)`, `library(copula)` and `library(mgpd)` to load the packages to your `R`-session. These packages are installed in all computers in the computer room MH:230. If you use your own laptop you need to install the packages yourself; see the help page for `install.packages` for more information.

Exercise 2: Random number generation from bivariate extreme value distributions

In all exercises below you will plot several datasets. Note that you can use `mfrow` to collect any number of the plots in one single figure.

1. In the package `evd` nine parametric bivariate extreme value models have been implemented. Study each family and specify which parameter(s) stand for strength of dependence in each family. If the model is asymmetric specify also which parameter stands for asymmetry. Note that the parametrization of some families in `evd` package (as you can see in the help page for `bvevd` in the `evd` manual which you can download from cran.r-project.org) is different from the course literature and lecture notes (see e.g. the *logistic* model).

2. Choose one symmetric and one asymmetric family and generate 200 observations from each model using the package `evd`. In order to check your answer to the previous part, in each case simulate for a couple of different values of parameters in the model and make a scatter plot of the data. Does the dependence parameter have the expected effect on the dependence of simulated data?

   For each pair of the simulated sample calculate the empirical values of Pearson's correlation ρ, Kendall's τ and Spearman's ρ_S.

3. In the package `SimCop` nine parametric bivariate and multivariate copulas have been implemented. Download the reference manual `SimCop.pdf` and study each family and specify which parameter(s) stand for strength of dependence in each family. If the model is asymmetric specify also which parameter stands for asymmetry.

4. Generate 200 observations from the bivariate asymmetric logistic extreme value copula for different combinations of the parameters and plot the results. For each plot comment on the effect of changing the copula parameters on the dependence structure of the model. Check the help documents for `GenerateRV.CopApprox` and `NewBEVAsyLogisticCopula` to see examples of how these functions can be used to generate random observations from copulas. You do not need to use `Metropolis-Hastings` algorithm (`MH`) argument) in your simulations. See also the help document for `plot.CopApprox`. Note that by default the `plot` function for the class `CopApprox` creates interactive plots but by setting the `type=öriginal"` you can create the traditional 2-dimensional plots.

Exercise 3: Extreme value analysis of maximum sea level using `evd` package

This part of the assignment is concerned with the annual maximum sea-level in Fremantle and Portpirie in Australia. The datasets "`fremantle.R`" and "`portpirie.R`" can be downloaded from the following locations

- The page `Datasets in R` under the `Pages` in the homepage of the course in *Canvas*,
- http://www.maths.lth.se/matstat/kurser/fmsn15masm23/datasetsR.html.

Note that these datasets should be loaded to `R` using `source` command.

1. Find out for which years the maximum sea level is available in each location.

2. Create a data frame which contains data for those years for which maximum sea level is available in both locations. One way of doing this is to use the function `merge` in `R`. See the help file of the function for details.

3. Create a scatter plot of maximum sea level in both locations.

4. Choose at least two parametric models (one symmetric and one asymmetric) and fit the models to the dataset. In both cases use full maximum likelihood (FML) and inference for margins (IFM) methods to estimates the parameters. Compare your results.

5. Choose at least one non-parametric model and fit it to the dataset. Use both parametric and empirical transformation of margins to estimate the dependence function (see the argument `epmar` to `abvnonpar`). Plot both estimates and compare them.

6. Suppose X and Y stand for the annual maximum sea level in Fremantle and Portpirie, respectively. Estimate the probabilities

   - $P((X, Y) > (1.7, 4.2))$,
   - $P((X, Y) > (1.8, 4.4))$, and
   - $P((X, Y) \not< (1.478, 3.850))$

   based on both parametric and non-parametric fits above. The values in the last probability above are the 30%-quantiles of the sea levels in the dataset. Compare them also with the empirical estimates of the probabilities.

   Remarks: For parametric models you can use `pbvevd` and `pgev` to calculate the distribution function of bivariate and univariate extreme value distributions, respectively. There is no such functions if the dependence function is estimated by a non-parametric method. In this case one needs to use the following relationship as it was shown in "Exercises 2": $G(x, y) = G_*((1+\gamma_1\frac{x-\mu_1}{\sigma_1})_+^{\frac{1}{\gamma_1}}, (1+\gamma_2\frac{y-\mu_2}{\sigma_2})_+^{\frac{1}{\gamma_2}})$ where $G_*(x, y) = e^{-(\frac{1}{x}+\frac{1}{y})A(\frac{x}{x+y})}$. The function `abvnonpar` can be used to find the value of dependence function in a specific point.

7. With the same notation as in previous question estimate

   $$P((X, Y) < (1.95, 4.8) | (X, Y) \not< (1.478, 3.850)).$$

8. Plot both parametric and non-parametric estimates for upper quantile curves of bivariate extreme value distributions fitted above for $p = 0.75, 0.90, 0.95$ (see the help page for the function `plot.bvevd`).

9. The Fremantle dataset was also analyzed in the "Computer Assignment 3" of the course on univariate extreme values. It was noted that there seems to be a trend in Fremantle sealevel data which might be explained by including "Year" and "SOI" (South Oscillation Index) as linear trend in the location parameter. Fit the same parametric models as above but include "Year" and "SOI" as covariates for Fremantle sealevel in the model (see `nsloc1` and `nsloc2` arguments in the help file for `fbvevd`). Recall also that in order to avoid numerical difficulties in optimization you need to normalize "Year" to $[-1, 1]$. Use log-likelihood ratio test to verify if any of these have significant effect on the fit. The functions `deviance, logLik` and `anova` can also be used to compare nested models in `R`.

Exercise 4: Extreme value analysis of maximum sea level using copulas

As the package `copula` provides some extra functions such as goodness of fit tests for fitting different copulas there are some advantages in using this package even for exrtreme value analysis. On the other hand there are only five families of extreme value copulas included in this package (see the help file for `evCopula`)

1. Fit parametric models “huslerReiss” , ”gumbel” and “galambos” implemented in the `copula` package to the dataset using FML method. Compare your results with the corresponding models in the `evd` package (“hr”, ”log” and “neglog” models) to make sure that the estimated parameters are the same. What is the difference in parametrization of ”logistic” model in these two packages?

2. Use the Inference For Margins (IFM) method based on `GEV` margins and Canonical Maximul Likelihood (CML) method without specifying the marginal distributions and fit the models to the dataset.

3. Use the function `gofEVCopula` for goodness-of-fit tests for parametric bivariate extreme-value copulas which you have fitted to the dataset above.

4. Use the function `An.biv` and find non-parametric estimators of the dependence function according to `Pickands` and `CFG` method. Plot the estimates in a figure.

5. Suppose X and Y stand for the annual maximum sea level in Fremantle and Portpirie, respectively. Estimate the probabilities

   - $P((X,Y) > (1.7, 4.2))$,
   - $P((X,Y) > (1.8, 4.4))$, and
   - $P((X,Y) \not< (1.478, 3.850))$

   based on both parametric and non-parametric fits above. The conditional values in the last probability above are the 30%-quantiles of the sea levels in the dataset. Compare them also with the empirical estimates of the probabilities. You have done this part in the previous exercise as well using `evd` package. Compare your results.

6. With the same notation as in previous question estimate

$$P((X,Y) < (1.95, 4.8) | (X,Y) \not< (1.478, 3.850)).$$

Exercise 5: Extreme value analysis of maximum sea level using `SimCop` package

One advantage of using `SimCop` package is that we can create a copula model based on non-parametric smoothing splines estimate of the dependence function. This can be done by using `NonparEstDepFct` and `SplineFitDepFct` functions. Start with reading the help files for these functions. In particular check the examples which demonstrate how the functions can be used.

1. Find the Pickand’s estimator of the dependence function for the annual maximum sea level dataset you have analyzed above. Plot the estimates for two cases corresponding to setting `convex.hull` to `TRUE` or `FALSE`.

2. Find the estimate of dependence function using cubic smoothing splines. As input use the non-parametric estimator of dependence function with the argument `convex.hull = FALSE`. Plot the resulting estimate.

3. Create a copula for the smoothing spline fit using the function `NewBEVSplineCopula`. Find an approximation to the copula by using the function `GetApprox` and plot the corresponding copula.

4. Suppose we want to estimate the same probabilities as in the previous exercises by using simulated observations from the smoothing splines copulas. Generate 1000 observations from the smoothing splines copula and estimate the following probabilities:

(a) $P((X,Y) > (1.7, 4.2))$,

(b) $P((X,Y) > (1.8, 4.4))$,

(c) $P((X,Y) \not< (1.478, 3.850))$, and

(d) $P((X,Y) < (1.95, 4.8) | (X,Y) \not< (1.478, 3.850))$.

Compare the results with your answers to the corresponding probabilities in exercises 3 and 4.

Exercise 6: Peaks over threshold analysis of sea level

In this exercise the same dataset on sea level will be analyzed by using peaks over threshold method. As discussed in the course there are at least two ways of defining exceedances in higher dimensions. In the first definition a distribution is fitted to the observations $\{(x,y)|(x,y) > (u_x, u_y)\}$ where u_x and u_y are suitable thresholds for each margin. Second definition aims to fit a distribution to $\{(x,y)|(x,y) \not< (u_x, u_y)\}$ where (u_x, u_y) is defined as before. These distributions will be called Type I and Type II bivariate generalized Pareto distributions (BGPD), respectively.

Type I BGPD

The function `fbvpot` in package `evd` fits a Type I model by maximizing the censored likelihood as given in e.g. Section 8.3.1 of Coles (2001) which is attached.

1. Calculate the 30%-quantiles of the sea levels in both locations. Use these values as thresholds for the margins in the analysis below.

2. Choose at least two parametric models (one symmetric and one asymmetric) and fit a Type I BGPD to the dataset above the thresholds.

3. Plot upper quantile curves of the distributions fitted above for $p = 0.75, 0.9, 0.95$. (see the help page for the function `plot.bvpot`). Compare these quantile curves with what you have plotted in part 8 of exercise 2. Do they agree?

4. Calculate
$$P((X,Y) < (1.95, 4.8) | (X,Y) \not< (1.478, 3.850))$$
based on your parametric models.

   Remarks: Note that you need to transform (X,Y) to $(\widetilde{X}, \widetilde{Y})$ according to the theory which has been discussed in the course (see also attached pages from *Coles* book). Read also **Details** section in the help page for *bvevd* on page 14 of *evd* manual to check which parametrization has been used for each model and how they can be transformed to an arbitrary bivariate extreme value distribution with `GEV` margins.

5. Note that if the shape parameter in GPD is negative then the distribution has finite right end point $\sigma/|\gamma|$. It is important to take this into consideration when you transform the margins to unit Frechét distribution. As an example calculate
$$P((X,Y) < (2,5) | (X,Y) \not< (1.478, 3.850)).$$

Type II BGPD

Type II bivariate generalized Pareto distributions can be fitted by using the package `mgpd`. The main function is `fbgpd`. As an extra help on how to use the functions in this package you can download and use the file ”`bgpdExampleRun.R`” from the following locations

- The page `Datasets in R` under the `Pages` in the homepage of the course in *Canvas*,
- http://www.maths.lth.se/matstat/kurser/fmsn15masm23/datasetsR.html.

1. Choose at least two parametric models (one symmetric and one asymmetric) and fit a Type II BGPD to the exceedances over the same thresholds as for the Type I BGPD above.

   Note that the estimation is based on a complicated likelihood function and therefore it is strongly dependent on good start values. One way of finding suitable start values for a model is to use estimates from a simpler model (e.g. Logistic model) as a starting point. Further, if you get an error in `R` such as "invalid NA contour values"it means that the density for some points has become not available". In this case you should try to change `x` and `y` vectors in the `bgpdExampleRun.R`.

2. Calculate
   $$P((X,Y) < (1.95, 4.8) | (X,Y) \not< (1.478, 3.850)).$$
   Compare the result with your answer to question 4d in exercise 3 and question 4 in type I BGPD.

3. Plot estimates of prediction regions for $p = 0.75, 0.9, 0.95$ based on parametric models you have fitted to the dataset. Compare your results with your answer to the same question for the Type I BGPD above and comment on any differences you see.

Finishing Off:
When you’ve finished, close down R by typing `q()`. Choose ‘Save’ when prompted as to whether you want to retain your workspace.

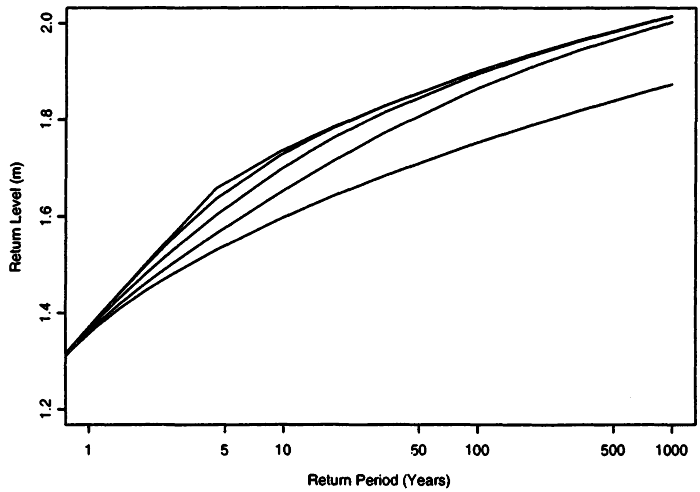

FIGURE 8.5. Comparison of return level plot for $Z = \min\{M_x, (M_y - 2.5)\}$ in logistic model analysis of Fremantle and Port Pirie annual maximum sea-level series with $\alpha = 0$, 0.25, 0.5, 0.75 and 1, respectively. Lowest curve corresponds to $\alpha = 1$; highest to $\alpha = 0$.

excess model and point process model can be obtained. In this section we give a brief description of both techniques.

8.3.1 Bivariate Threshold Excess Model

In Chapter 4 we derived as a class of approximations to the tail of an arbitrary distribution function F the family

$$G(x) = 1 - \zeta \left\{ 1 + \frac{\xi(x-u)}{\sigma} \right\}^{-1/\xi}, \quad x > u. \tag{8.18}$$

This means there are parameters ζ, σ and ξ such that, for a large enough threshold u, $F(x) \approx G(x)$ on $x > u$. Our aim now is to obtain a bivariate version of (8.18), i.e. a family with which to approximate an arbitrary joint distribution $F(x, y)$ on regions of the form $x > u_x, y > u_y$, for large enough u_x and u_y.

Suppose $(x_1, y_1), \ldots, (x_n, y_n)$ are independent realizations of a random variable (X, Y) with joint distribution function F. For suitable thresholds u_x and u_y, the marginal distributions of F each have an approximation of the form (8.18), with respective parameter sets $(\zeta_x, \sigma_x, \xi_x)$ and $(\zeta_y, \sigma_y, \xi_y)$.

The transformations

$$\tilde{X} = -\left(\log\left\{1 - \zeta_x\left[1 + \frac{\xi_x(X - u_x)}{\sigma_x}\right]^{-1/\xi_x}\right\}\right)^{-1} \tag{8.19}$$

and

$$\tilde{Y} = -\left(\log\left\{1 - \zeta_y\left[1 + \frac{\xi_y(Y - u_y)}{\sigma_y}\right]^{-1/\xi_y}\right\}\right)^{-1} \tag{8.20}$$

induce a variable $(\tilde{X}, \tilde{Y})$ whose distribution function $\tilde{F}$ has margins that are approximately standard Fréchet for $X > u_x$ and $Y > u_y$. By (8.5), for large n,

$$\begin{aligned}\tilde{F}(\tilde{x}, \tilde{y}) &= \left\{\tilde{F}^n(\tilde{x}, \tilde{y})\right\}^{1/n} \\ &\approx \left[\exp\left\{-V(\tilde{x}/n, \tilde{y}/n)\right\}\right]^{1/n} \\ &= \exp\left\{-V(\tilde{x}, \tilde{y})\right\},\end{aligned}$$

because of the homogeneity property of V. Finally, since $F(x, y) = \tilde{F}(\tilde{x}, \tilde{y})$, it follows that

$$F(x, y) \approx G(x, y) = \exp\{-V(\tilde{x}, \tilde{y})\}, \quad x > u_x, y > u_y, \tag{8.21}$$

with $\tilde{x}$ and $\tilde{y}$ defined in terms of x and y by (8.19) and (8.20). This assumes that the thresholds u_x and u_y are large enough to justify the limit (8.5) as an approximation. We discuss this point further in Section 8.4.

Inference for this model is complicated by the fact that a bivariate pair may exceed a specified threshold in just one of its components. Let

$$R_{0,0} = (-\infty, u_x) \times (-\infty, u_y), R_{1,0} = [u_x, \infty) \times (-\infty, u_y),$$
$$R_{0,1} = (-\infty, u_x) \times [u_y, \infty), R_{1,1} = [u_x, \infty) \times [u_y, \infty),$$

so that, for example, a point $(x, y) \in R_{1,0}$ if the x-component exceeds the threshold u_x, but the y-component is below u_y. For points in $R_{1,1}$, model (8.21) applies, and the density of $\tilde{F}$ gives the appropriate likelihood component. On the other regions, since $\tilde{F}$ is not applicable, it is necessary to censor the likelihood component. For example, suppose that $(x, y) \in R_{1,0}$. Then since $x > u_x$, but $y < u_y$, there is information in the data concerning the marginal x-component, but not the y-component. Hence, the likelihood contribution for such a point is

$$\Pr\{X = x, Y \le u_y\} = \left.\frac{\partial F}{\partial x}\right|_{(x, u_y)}$$

as this is the only information in the datum concerning F. Applying similar considerations in the other regions, we obtain the likelihood function

$$L(\theta; (x_1, y_1), \ldots, (x_n, y_n)) = \prod_{i=1}^{n} \psi(\theta; (x_i, y_i)), \tag{8.22}$$

where θ denotes the parameters of F and

$$\psi(\theta;(x,y)) = \begin{cases} \left.\frac{\partial^2 F}{\partial x \partial y}\right|_{(x,y)} & \text{if } (x,y) \in R_{1,1}, \\ \left.\frac{\partial F}{\partial x}\right|_{(x,u_y)} & \text{if } (x,y) \in R_{1,0}, \\ \left.\frac{\partial F}{\partial y}\right|_{(u_x,y)} & \text{if } (x,y) \in R_{0,1}, \\ F(u_x,u_y) & \text{if } (x,y) \in R_{0,0}, \end{cases}$$

with each term being derived from the joint tail approximation (8.21). Maximizing the log-likelihood leads to estimates and standard errors for the parameters of F in the usual way. As with the componentwise block maxima model, the inference can be simplified by carrying out the marginal estimation, followed by transformations (8.19) and (8.20), as a preliminary step. In this case, likelihood (8.22) is a function only of the dependence parameters contained in the model for V.

An alternative method, if there is a natural structure variable $Z = \phi(X,Y)$, is to apply univariate threshold techniques to the series $z_i = \phi(x_i,y_i)$. This approach now makes more sense, since the z_i are functions of concurrent events. In terms of statistical efficiency, however, there are still good reasons to prefer the multivariate model.

8.3.2 Point Process Model

The point process characterization, summarized by the following theorem, includes an interpretation of the function H in (8.6).

Theorem 8.2 Let $(x_1,y_1),(x_2,y_2)\ldots$ be a sequence of independent bivariate observations from a distribution with standard Fréchet margins that satisfies the convergence for componentwise maxima

$$\Pr\{M^*_{x,n} \le x, M^*_{y,n} \le y\} \to G(x,y).$$

Let $\{N_n\}$ be a sequence of point processes defined by

$$N_n = \{(n^{-1}x_1, n^{-1}y_1),\ldots,(n^{-1}x_n, n^{-1}y_n)\}. \tag{8.23}$$

Then,

$$N_n \xrightarrow{d} N$$

on regions bounded from the origin $(0,0)$, where N is a non-homogeneous Poisson process on $(0,\infty) \times (0,\infty)$. Moreover, letting

$$r = x + y \quad \text{and} \quad w = \frac{x}{x+y}, \tag{8.24}$$

the intensity function of N is

$$\lambda(r,w) = 2\frac{dH(w)}{r^2}, \tag{8.25}$$

where H is related to G through (8.5) and (8.6). □