TP2 Statistical learning with extreme values

Anne Sabourin, Antoine Doizé

October, 10, 2024

1 Introduction

In the course you studied theoretical properties of multivariate extension of the EVT. In particular you characterized the shape of limiting distribution of maxima or excesses of multivariate samples.
The goal of this TP is to give an idea of how to sample from those limiting distributions.
There are a huge variety of approaches to sample from such distributions. We will only look at a specific method to sample from the logistic model, with the methods suggested in Stephenson 2003.
For more methods, you can look first at the very short review of the `mev` package which is one of the main R packages for Extreme Value Modelling : click here to see mev documentation for simulation of multivariate extreme.

2 Multivariate logistic model

2.1 Brief introduction to multivariate logistic model

We will use the method suggested in [2]. The idea is to use a parametric distribution family to model the dependence structure of the multivariate extreme value distribution fonction. Let's denote by $d \in \mathbb{N}$ the dimension of our space. Denote by $G(x), \quad x \in \mathbb{R}_+^d$ the cdf of a multivariate distribution function. Let's define, as in the course, $V(\cdot) = -\log G(\cdot)$ which is the exponent measure of the distribution. Then, multivariate logistic model is characterized in [1] by having random variables having exponent function $V(x) = \mu([0,x]^c)$ defined by

$$V_L(x) = \left(\sum_{j=1}^{d} x_j^{-1/\alpha} \right)^{\alpha} .$$

$0 \leq \alpha \leq 1$ is called the dependence parameter, it measures the dependence between $X_j, j = 1 \ldots d$ the extremes $\alpha \to 1$,$\alpha \to 0$ correspond respectively to independence and complete dependence.

We will see in this TP how to sample X multivariate random variables having exponent function V_L.

2.2 Algorithm 1 for simulation

Consider the transformations

$$Z = \left(\sum_{j=1}^{d} X_j^{-1/\alpha} \right)^{\alpha} \tag{3}$$

$$T_i = (X_i Z)^{-1/\alpha} = \frac{X_i^{-1/\alpha}}{\sum_{j=1}^{d} X_j^{-1/\alpha}} \tag{4}$$

of [1], so that $\sum_{i=1}^{d} T_i = 1$. Let $\mathscr{S}_d = \left\{ (\omega_1, \ldots, \omega_d) \in \mathbb{R}_+^d : \sum_j \omega_j = 1 \right\}$ denote the $d-1$ dimensional unit simplex. Shi shows in [1] that $T_1, \ldots, T_d$ are independent of Z, $(T_1, \ldots, T_d)$ is distributed uniformly on $\mathscr{S}_d$, and that Z is a mixture of gamma distributions with density

$$f_d(z) = \sum_{j=1}^{d} p_{d,j} \Gamma(z, j)$$

where

$$\Gamma(z, k) = \frac{1}{\Gamma(k)} z^{k-1} \mathrm{e}^{-z}, \quad z > 0$$

is the density function of a gamma distribution with unit scale and shape parameter k. The mixture probabilities can be calculated using the recurrence relations

$$\begin{aligned} p_{d,1} &= \frac{\Gamma(d-\alpha)}{\Gamma(d)\Gamma(1-\alpha)} \\ (d-1)p_{d,j} &= (d-1-\alpha j)p_{d-1,j} + \alpha(j-1)p_{d-1,j-1}, \quad j = 2, \ldots, d-1, \\ p_{d,d} &= \alpha^{d-1} \end{aligned}$$

given in [1] Caution there is a typo in the [2] paper for value of $p_{d,d}$ which is α^{d-1} and not α^d

This provides a simple numerical method for the calculation of the d mixture probabilities $p_{d,1}, \ldots, p_{d,d}$ of $f_d(z)$ at any given dimension d and any $\alpha \in (0,1]$. Let $W_1, \ldots, W_d$ be independent standard exponential random variables. Denoting the cumulative probabilities by $P_m = \sum_{i=1}^m p_{d,i}$ for $m = 1, \ldots, d$ and setting $P_0 = 0$ yields the following algorithm, which is (using slightly different transformations) reproduced from [1].

Algorithm 1: Multivariate logistic

1. Set $(T_1, \ldots, T_d) = \Big(W_1 / \sum_{j=1}^d W_j, \ldots, W_d / \sum_{j=1}^d W_j\Big)$.

2. Generate U uniformly over $(0,1)$ and find $k \in \{1, \ldots, d\}$ such that $P_{k-1} \leq U < P_k$.

3. Generate Z from a gamma distribution with shape parameter k and unit scale.

4. Set $X = (X_1, \ldots, X_d) = (1/ZT_1^\alpha, \ldots, 1/ZT_d^\alpha)$.

2.3 Brief overview of the algorithm

1. Question on the model: Check that the exponent measure V_L satisfies the homogeneity property (Hint: you can compute $V_L(tx), \quad t > 0, \quad x \in \mathbb{R}_+^d$, and $\lim_{x_k \to \infty, k \neq j} V_L((x_1 \ldots x_d))$ with $j = 1 \ldots d$ and $x_j = 1$)

2. Questions on **Step 1**

   (a) What is the support of the random vector $(T_1, \ldots, T_d)$?

   (b) What is its law on this support ? (Give a brief argument)

3. Questions on **Step 2**

   (a) Calculate $p_{1,1}$, $p_{2,1}$, $p_{2,2}$, $p_{3,1}$, $p_{3,2}$, $p_{3,3}$.

   (b) The k defined in step 2 is random (its value will depend on U value). What is its law ? (Just give the result)

4. Question on **Step 3**: Can you deduce the law of Z ?

5. Question on **Step 4**: Can you deduce the law of X ?

2.4 Sampling from multivariate logistic model

You can use the `tp2.ipynb` file provided.

1. Try to sample T random variables from the `step 1` function implemented in the`tp2.ipynb` file. Plot on the same graph histograms of the marginals of T. How do you interpret this graph ?

2. Complete the `step 2` function in the `tp2.ipynb` file (Caution there is a typo in [2]: use the recurrence relationship in [1] or the corrected version of this TP.

3. Implement the algorithm to generate a sample of n variables from the exponent measure, in dimension $d \in \mathbb{N}$, with $\alpha \in (0, 1]$. Use $d = 4$, $\alpha = 1/4$.

4. Sample $(X_i)_{i=1\dots n}$ for a large n. Plot boxplots of the angle of X for a growing value of its norm. (You can make boxplots of the angles of the variables $X_i \mathbb{I}_{r \leq ||X|| < r+1}, r = 1 \dots 100$). How do you interpret this graph ?

5. Set $d = 2$ and compare the angle distribution (marginal and bivariate distribution) for $\alpha \to 0$ and for $\alpha \to 1$.

2.5 To go further

To go further, you can

- Dive into the paper [2] and try to reproduce Algorithm 2.1 which is another way of sampling from the logistic model, and compare both methods.

- Dive into the paper [2] and try to reproduce Algorithm 2.2 or 1.2 to sample from asymmetric logistic model. (this part can be time consuming, because of high computing times in high dimension.

References

[1] Daoji Shi. Multivariate extreme value distribution and its fisher information matrix. *Acta Mathematicae Applicatae Sinica*, 11, 1995.

[2] Alec Stephenson. Simulating Multivariate Extreme Value Distributions of Logistic Type. *Extremes*, 6, 2003.