Estimation of multivariate critical layers: Applications to rainfall data

Elena Di Bernardino[1] & Didier Rullière[2]

[1] *CNAM, Paris, Département IMATH, France, elena.di_bernardino@cnam.fr*
[2] *Université de Lyon 1, ISFA, Laboratoire SAF, France, didier.rulliere@univ-lyon1.fr*

Résumé. Dans un environnement multivarié, le calcul de zones critiques et de périodes de retour associées est un problème difficile. Un cadre théorique possible pour le calcul de ces périodes de retour est essentiellement basé sur la notion de Copule et sur les ensembles de niveau d'une distribution de probabilité multivariée. Dans ce travail, nous proposons une méthodologie rapide et paramétrique pour estimer les zones critiques de distributions multivariées et leurs périodes de retour associées. Le modèle est basé sur des transformations des distributions marginales et sur des transformations de la structure de dépendance au sein de la classe des copules Archimédiennes. La méthodologie est illustrée sur des données réelles de précipitation. Sur ce jeu de données, nous développons également un modèle imbriqué transformé.

Mots-clés. Transformations multivariées de probabilité, ensembles de niveau, estimation de copules, les fonctions de conversion hyperboliques, périodes de retour multivariées.

Abstract. Calculating return periods and critical layers (i.e. multivariate quantile curves) in a multivariate environment is a difficult problem. A possible consistent theoretical framework for the calculation of the return period, in a multi-dimensional environment, is essentially based on the notion of copula and level sets of the multivariate probability distribution. In this paper we propose a fast and parametric methodology to estimate the multivariate critical layers of a distribution and its associated return periods. The model is based on transformations of the marginal distributions and transformations of the dependence structure within the class of Archimedean copulas. The methodology is illustrated on rainfall 5-dimensional real data. We also develop a nested model on this rainfall 5-dimensional real data set.

Keywords. Multivariate probability transformations, level sets, estimation copulas, hyperbolic conversion functions, risk assessment, multivariate return periods.

1 Introduction: the Return Periods

The notion of *Return Period* (RP) is frequently used in environmental sciences for the identification of dangerous events, and provides a means for rational decision making and risk assessment. Roughly speaking, the RP can be considered as an analogue of the "Value-at-Risk" in Economics and Finance, since it is used to quantify and assess the risk (see, e.g., Nappo and Spizzichino, 2009). In engineering practice, finance, insurance

and environmental science the choice of the RP depends on the impact/magnitude of the considered event and the consequences of its realisation. Equally important is the related concept of *design quantile*, usually defined as "the value of the variable characterizing the event associated with a given RP". In the univariate case the design quantile is usually identified without ambiguity. Conversely in the multivariate setting different definitions are possible. For this reason, the identification problem of design events in a multivariate context has recently attracted the attention of many researches (see for instance Salvadori et al., 2007). In the following, we will consider a sequence $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots\}$ of independent and identically distributed $d-$dimensional random vectors, with $d > 1$. Thus each $\mathbf{X}_k$, $k \in \mathbb{N}$, has the same multivariate distribution $F_{\mathbf{X}} : \mathbb{R}_+^d \to [0,1]$ as the nonnegative real-valued random vector $\mathbf{X} \sim F_{\mathbf{X}} = C(F_{X_1}, \ldots, F_{X_d})$ describing the hydrological phenomenon under investigation. The function C is the d-dimensional *copula* associated to F. We write $I = \{1, \ldots, d\}$ the set of indexes of the considered random variables and of their associated cumulative distribution functions, i.e., $F_{X_i}(x_i) = P(X_i \leq x_i)$, for $i \in I$.

Definition 1.1 (Critical layer) *The critical layer $\partial L(\alpha)$ associated to the multivariate distribution function $F_{\boldsymbol{X}}$ of level $\alpha \in (0,1)$ is defined as*

$$\partial L(\alpha) = \{\boldsymbol{x} \in \mathbb{R}^d : F_{\boldsymbol{X}}(\boldsymbol{x}) = \alpha\}.$$

Then $\partial L(\alpha)$ is the iso-hyper-surface (with dimension $d-1$) where F equals the constant value α. The critical layer $\partial L(\alpha)$ partitions $\mathbb{R}^d$ into three non-overlapping and exhaustive regions:

$$\begin{cases} L^<(\alpha) &= \{\mathbf{x} \in \mathbb{R}^d : F_{\mathbf{X}}(\mathbf{x}) < \alpha\}, \\ \partial L(\alpha) &= \text{the critical layer itself}, \\ L^>(\alpha) &= \{\mathbf{x} \in \mathbb{R}^d : F_{\mathbf{X}}(\mathbf{x}) > \alpha\}. \end{cases}$$

The Return Period is defined as the average time required for reaching the set $L^>(\alpha)$, that is:

$$\mathrm{RP}^>(\alpha) = \Delta_t \cdot \mathbb{E}[N] = \frac{\Delta_t}{\mathbb{P}[\mathbf{X} \in L^>(\alpha)]}, \tag{1}$$

where $\Delta_t > 0$ is the (deterministic) average time elapsing between $\mathbf{X}_k$ and $\mathbf{X}_{k+1}$, $k \in \mathbb{N}$. The probability that a realization of this vector belongs to $L^<(\alpha)$ is given by the Kendall's function, which only depends on the copula C of this random vector, i.e.,

$$K_C(\alpha) = \mathbb{P}[X \in L^<(\alpha)] = \mathbb{P}[C(U_1, \ldots, U_d) \leq \alpha], \quad \text{for } \alpha \in (0,1). \tag{2}$$

Then, the considered Return Period can be expressed using Kendall's function in (2), $\mathrm{RP}^>(\alpha) = \Delta_t \cdot \frac{1}{1-K_C(\alpha)}$. This paper aims at:

- giving a *parametric representation of the multivariate distribution F* of a random vector $\mathbf{X}$, here representing rain measurements,

- giving direct *estimation procedure* for this representation,
- giving closed parametric expressions, both for *critical layers* in Definition 1.1 and Return Periods in (1),
- adapting this methodology to some asymmetric dependencies (as, for instance, non-exchangeable random vectors) by using *nested model for recorded data*.

In the next section, we introduce the model used to answer the issues introduced above.

2 The transformed model

We consider the following model, which is detailed in Di Bernardino and Rullière (2013a),

$$\widetilde{F}(x_1,\dots,x_d) = T \circ C_0(T_1^{-1} \circ F_1(x_1),\dots,T_d^{-1} \circ F_d(x_d)),$$

where $F_1,\dots,F_d$ are given parametric *initial marginal cumulative distribution functions*, and where C_0 is a given *initial copula*. Hence the distribution $\widetilde{F}(x_1,\dots,x_d)$ is built from transformed marginals $\widetilde{F}_i(x) = T \circ T_i^{-1} \circ F_i(x)$, for $i \in I$ and from a transformed copula $\widetilde{C}(u_1,\dots,u_d) = T \circ C_0(T^{-1}(u_1),\dots,T^{-1}(u_1))$, under regularity conditions. Transformation T permits to transform the initial dependence structure C_0. For a given T, transformations T_i permit to transform marginals, $i \in I$. Furthermore we will assume in the following that C_0 is an Archimedean copula. In our paper we show how to estimate the transformations T and T_i, $i \in I$. Estimation procedure is omitted here for sake of brevity. For further details the interested reader is referred to Di Bernardino and Rullière (2015).

3 Numerical results on rainfall 5-dimensional real data

In the following, we present some estimation results on rainfall 5-dimensional real data provided using our estimation procedure. Data comes from the website *CISL Research Data Archive (RDA)*. Geographical position of 5 stations and the scatter plot of ranks of data are provided in Figure 1. We perform a Goodness-of-Fit test based on the empirical process in order to test the quality of the adjustment of our transformed copula $\widetilde{C}$ on these multivariate data. We obtain a $p-$value equal to 0.38. In the large scale Monte Carlo experiments carried out by Genest et al. (2009), the statistic S_n gave the best results overall. An approximate p-value for S_n can be obtained by means of a parametric bootstrap-based procedure (see results in Table 1). In the following we intend to show the flexibility of the proposed model and associated estimation procedure. In particular, we adapt our methodology in the case of some asymmetric dependencies (as, for instance, non-exchangeable random vectors). The correlation matrix of the considered rainfall data is displayed in Figure 2(left). As we can see, some pairs of stations present a bigger correlation. To illustrate how our model can be adapted to this situation we have decided to create 2 different clusters. An hierarchical cluster analysis on the set of dissimilarities produce by the distance of the X_i is developed. We obtain the result in

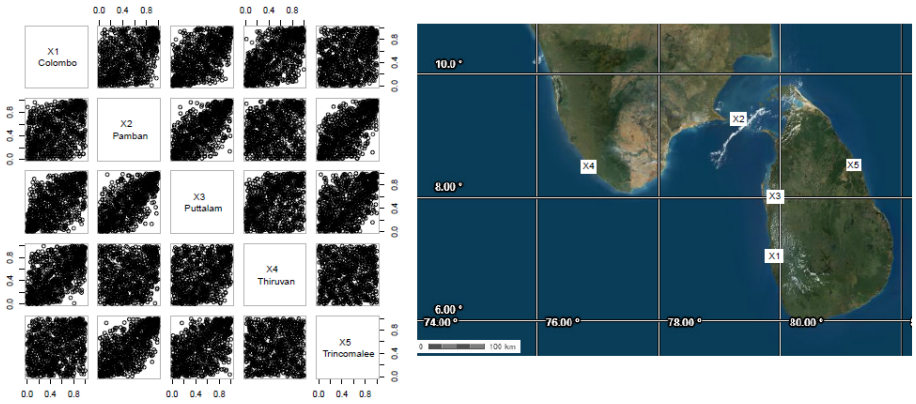

Figure 1: Left: Scatter plot of ranks for the considered 797 monthly rainfall measurements (in decimeter) in 5 stations of Sri-Lanka and India between January 1893 and June 2013. Right: Geographical positions of 5 considered stations.

Copula under H_0	S_n	S_n^B	S_n^C	A_n
Gumbel-Hougaard	0.00331	0.00495	0.00454	0.03465
Clayton	0.00381	0.00980	0.00704	0.00981
Frank	0.00617	0.00941	0.00819	0.08416
t-Student	0.00495	0.00592	0.00498	0.00963
Normal	0.00980	0.00719	0.00454	0.00205
Joe	0.00819	0.00495	0.00454	0.00916

Table 1: The bootstrapped $p-$values for different Goodness-of-Fit tests (see Genest et al., 2009) for competitor copula families on the considered 5-dimensional rainfall data, with $n = 797$. In all cases, the number of Monte Carlo experiments is fixed at $N = 1000$.

Figure 2. As one can see, whatever the distance chosen for dissimilarities, the dendrogram gives a justification to chosen **clusters of station indexes** $\{2, 3, 5\}$ **and** $\{1, 4\}$**.** Then, we firstly fit a $3-$dimensional model for the first group and a $2-$dimensional one for the second one. We generate the pseudo-data coming from these two models and finally we construct the joint (root) copula for these bivariate data-set. In the following we take as initial copula C_0 the independent one, and the initial margins $F_i(x) = 1-\mathrm{e}^{-x}$, $i \in A, B$.

The multivariate distribution for the cluster $A = \{2, 3, 5\}$ is assumed to be written:

$$\begin{aligned} F_A(x_2, x_3, x_5) &= T_A \circ C_0(T_A^{-1} \circ \widetilde{F}_2(x_2),\, T_A^{-1} \circ \widetilde{F}_3(x_3),\, T_A^{-1} \circ \widetilde{F}_5(x_5)), \\ \text{with } \tilde{F}_i &= T_A \circ T_{A_i}^{-1} \circ F_i\,, \quad \text{for } i \in A. \end{aligned} \tag{3}$$

$$\begin{bmatrix} & X_1 & X_2 & X_3 & X_4 & X_5 \\ X_1 & \mathbf{1} & 0.377 & 0.552 & 0.504 & 0.207 \\ X_2 & 0.377 & \mathbf{1} & \mathbf{0.698} & 0.178 & \mathbf{0.660} \\ X_3 & 0.552 & \mathbf{0.698} & \mathbf{1} & 0.276 & 0.523 \\ X_4 & 0.504 & 0.178 & 0.276 & \mathbf{1} & 0.018 \\ X_5 & 0.207 & \mathbf{0.660} & 0.523 & 0.018 & \mathbf{1} \end{bmatrix}$$

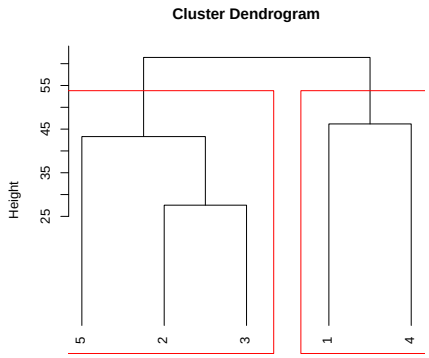

Figure 2: Left: Correlation matrix of the considered rainfall data. Correlations greater than 60% are indicated in bold font. Right: Dendrogram resulting to the hierarchical cluster analysis on the set of dissimilarities produced by the Euclidian distance on the rainfall data. Red boxes show the two considered clusters.

The multivariate distribution for the cluster $B = \{1, 4\}$ is assumed to be written:

$$\begin{aligned} F_B(x_1, x_4) &= T_B \circ C_0(T_B^{-1} \circ \widetilde{F}_1(x_1),\, T_B^{-1} \circ \widetilde{F}_4(x_4)), \\ \text{with } \tilde{F}_i &= T_B \circ T_{B_i}^{-1} \circ F_i\,, \quad \text{for } i \in B. \end{aligned} \tag{4}$$

The whole 5−dimensional distribution is assumed to be written:

$$\widetilde{F}(x_1, x_2, x_3, x_4, x_5) = T \circ C_0\left(T^{-1} \circ F_A(x_2, x_3, x_5),\, T^{-1} \circ F_B(x_1, x_4)\right), \tag{5}$$

where $\tilde{C}(u,\, v) = T \circ C_0\left(T^{-1}(u),\, T^{-1}(v)\right)$ is referred as the *root copula* at point (u, v). It is effectively a proper copula if the transformation T satisfies admissibility conditions that are given in Proposition 2.1 of Di Bernardino and Rullière (2013b). To estimate the external transformation T of model (5) we firstly construct a bivariate pseudo data-set:

$$Z_1 = F_A(X_1, X_4), \quad Z_2 = F_B(X_2, X_3, X_5).$$

Then we fit on this bivariate data-set a model

$$\begin{aligned} \widetilde{F}_{(Z_1,Z_2)}(z_1, z_2) &= T \circ C_0(T^{-1} \circ \tilde{F}_1(z_1),\, T^{-1} \circ \tilde{F}_2(z_2)), \\ \text{with } \tilde{F}_i &= T \circ T_1^{-1} \circ F_i\,, \quad \text{for } i = 1, 2. \end{aligned}$$

Let $\alpha \in (0, 1)$ be a targeted level for a critical layer. Let C_0 be the initial copula to be transformed, and assume that C_0 is the independent copula. Then the analytical critical layers of the distributions F_B and F_A are easy to obtain. For F_B in (4), we have

$$\partial L_B(\alpha) = \left\{(x_1, x_4) : x_1 = \widetilde{F}_1^{-1} \circ T_B\left((T_B^{-1}(\alpha))^p\right), x_4 = \widetilde{F}_4^{-1} \circ T_B\left((T_B^{-1}(\alpha))^{1-p}\right), p \in (0, 1)\right\}.$$

Analytical expressions of the inverse of any transformed margins $\widetilde{F}_i = T \circ T_i^{-1} \circ F_i$, for $i \in B$, are available since inverse transformations are given and since the initial distribution F_i is chosen to be readily invertible. Analogously, we get, for F_A in (3)

$$\partial L_A(\alpha) = \{(x_2, x_3, x_5) : x_2 = \widetilde{F}_1^{-1} \circ T_B\left((T_B^{-1}(\alpha))^{p_1}\right), x_3 = \widetilde{F}_3^{-1} \circ T_B\left((T_B^{-1}(\alpha))^{p_2}\right), \\ x_5 = \widetilde{F}_5^{-1} \circ T_B\left((T_B^{-1}(\alpha))^{1-p_1-p_2}\right),\ p_1, p_2 \in (0,1),\ p_1 + p_2 < 1\}.$$

For the nested distribution $\widetilde{F}$ in (5), one can write,

$$\begin{aligned} \partial L(\alpha) &= \left\{(x_1, \ldots, x_5) : \widetilde{F}(x_1, x_2, x_3, x_4, x_5) = \alpha\right\} \\ &= \left\{(x_1, \ldots, x_5) : T \circ C_0\left(T^{-1} \circ F_A(x_2, x_3, x_5), T^{-1} \circ F_B(x_1, x_4)\right) = \alpha\right\} \\ &= \left\{(x_1, \ldots, x_5) : T^{-1} \circ F_A(x_2, x_3, x_5) \cdot T^{-1} \circ F_B(x_1, x_4) = T^{-1}(\alpha)\right\} \end{aligned}$$

An illustration of critical-layers $\partial L_A(\alpha)$ and $\partial L_B(\alpha)$ derived above is provided in Figure 3.

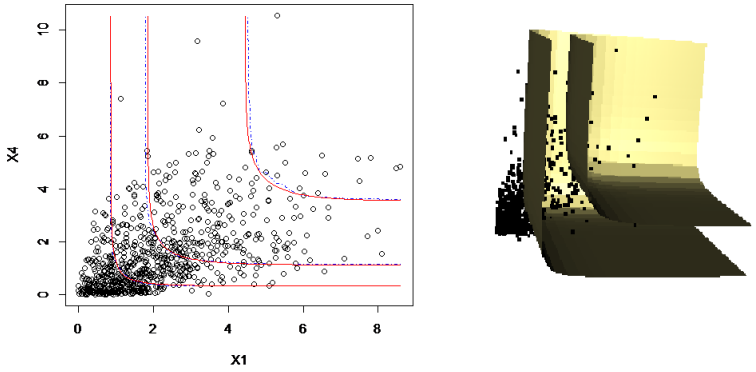

Figure 3: Left: 2-dimensional critical-layers $\partial L_B(\alpha)$ with $\alpha = 0.2, 0.5, 0.9$ with associated empirical critical-layers (blue dashed lines). Right: 3-dimensional critical layers $\partial L_A(\alpha)$ with $\alpha = 0.3, 0.9$. Black dots represent rainfall data (X_1, X_4) (left) and (X_2, X_3, X_5) (right).

Bibliographie

1 Di Bernardino, E. and Rullière, D. (2015). *Estimation of multivariate critical layers: Applications to rainfall data*, Journal de la Société Française de Statistique, à paraître.

2 Di Bernardino, E. and Rullière, D. (2013a). Distortions of multivariate distribution functions and associated level curves : Applications in multivariate risk theory. Insurance : Mathematics and Economics, 53(1) :190 - 205.

3 Di Bernardino, E. and Rullière, D. (2013b). On certain transformations of Archimedean copulas : Application to the non-parametric estimation of their generators. Dependence Modeling, 1 :1-36.

4 Genest, C., Rèmillard, B., and Beaudoin, D. (2009). Goodness-of-fit tests for copulas : A review and a power study. Insurance : Mathematics and Economics, 44(2) :199 - 213.

5 Nappo, G. and Spizzichino, F. (2009). Kendall distributions and level sets in bivariate exchangeable survival models. Information Sciences, 179 :2878 -2890

6 Salvadori, G., De Michele, C., Kottegoda, N., and Rosso, R. (2007). Extremes in Nature : An Approach Using Copulas. Springer-Verlag : Berlin