5th International Workshop on Climate Informatics
September 24-25, 2015
Hosted by the National Center for Atmospheric Research in Boulder, CO

BAYESIAN CLUSTERING AND DIMENSION REDUCTION IN MULTIVARIATE EXTREMES

Sabrina Vettori[1], Raphaël Huser[1], Marc G. Genton[1]

***Abstract*—The spatial dependence structure of climate extremes may be represented by the class of max-stable distributions. When the domain is very large, describing the spatial dependence between and within subdomains is particularly challenging and requires very flexible, yet interpretable, models. In this work, we use the inherent hierarchical dependence structure of the (max-stable) nested logistic distribution for clustering and dimension reduction in multivariate extremes, taking into account the occurrence times of extreme events. Methods are tested both through a simulation study and by analysing extreme air temperatures at different stations in Switzerland.**

I. INTRODUCTION

Advances in modelling environmental extremes are necessary to investigate increasingly abundant, diverse, heterogeneous and high-dimensional climatic datasets.

Max-stable distributions and processes have garnered widespread interest in extreme-value studies since they are asymptotically justified in describing the (multivariate or spatial) dependence structure of rare events. Although the set of possible extremal dependence structures is non-parametric by nature, inference typically relies on simple parametric but flexible and intuitive max-stable models. Here we focus on the limiting max-stable models, although there is a need to develop sub-asymptotic models as well [4]. The simplest and most tractable max-stable model is the logistic model [5, 6]. However, it is often overly simplistic in applications, as the dependence structure is summarised by only one parameter and its components are exchangeable, a feature that is rarely supported by complex (high-dimensional) datasets. Currently, active research is devoted on developing useful models and methods to handle high-dimensional extreme data. Clustering of extremes and dimension reduction are two possible (related) approaches to tackle this important problem. Indeed, depending on the topographical characteristic of the domain, e.g., altitude, or other factors, we expect different subdomains, e.g., plains, foothills, mountains, etc. to behave differently while being spatially correlated.

To address this issue, we focus on a flexible max-stable distribution, namely the nested logistic model [3, 8, 11]. This model is able to describe the dependence within and between distinct groups of variables residing in different subdomains with logistic distributions. This hierarchical dependence structure may therefore be summarised by a tree structure; see Figure 1. This model is similar to using nested factor copula models [7]. The idea of our work is to estimate model parameters, as well as to infer the hidden tree, by embedding the model in a Bayesian framework. This allows us to cluster variables that behave similarly and to achieve dimension reduction if some groups are found to be exactly independent. Moreover, through the inclusion of timestamp data, simultaneous occurrence of extreme events is considered in the maximum likelihood procedures [10], leading to greater efficiency both statistically and computationally.

II. MODELLING

Suppose that $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \sim F$ are independent and identically distributed D-dimensional random vectors, with common unit Fréchet marginal distributions, representing a variable of interest (e.g., rainfall or temperatures) observed at several monitoring stations. It can be shown that, as $n \to \infty$, the only limit $\mathbf{Z}$ of the renormalized componentwise maximum vector $n^{-1}\mathbf{M}_n$, where

$$\mathbf{M}_n = (M_{1,n}, \ldots, M_{D,n})^T = \left(\max_{i=1}^{n} Y_{1,i}, \ldots, \max_{i=1}^{n} Y_{D,i}\right)^T,$$

is max-stable. The joint distribution of the random vector $\mathbf{Z}$, containing the extreme observations, may be written as

$$\mathrm{P}(\mathbf{Z} \leq \mathbf{z}) = \exp\{-V(\mathbf{z})\}, \quad \mathbf{z} \in \mathbb{R}_+^D,$$

where $V(\mathbf{z})$ is the exponent measure. For example, the logistic model is defined by setting the exponent measure as

$$V_{\log}(\mathbf{z} \mid \alpha) = \left(z_1^{-1/\alpha} + \cdots + z_D^{-1/\alpha}\right)^{\alpha}, \qquad (1)$$

Corresponding author: S Vettori, King Abdullah University of Science and Technology, Saudi Arabia sabrina.vettori@kaust.edu.sa
[1] King Abdullah University of Science and Technology, Saudi Arabia

5th International Workshop on Climate Informatics
September 24-25, 2015
Hosted by the National Center for Atmospheric Research in Boulder, CO

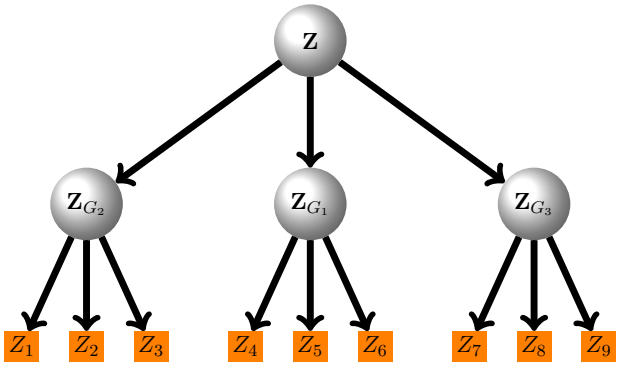

Figure 1. Tree with two layers summarising extremal dependence of a vector $\mathbf{Z} = (\mathbf{Z}_{G_1}^T, \mathbf{Z}_{G_2}^T, \mathbf{Z}_{G_3}^T)^T = (Z_1, \ldots, Z_9)^T$.

with $\alpha \in (0, 1]$. Here, the parameter α summarises the dependence strength between the components of $\mathbf{Z}$: $\alpha \to 0$ and $\alpha = 1$ correspond to the cases of perfect dependence and complete independence, respectively.

Suppose now that model (1) is not flexible enough to capture the dependence structure of $\mathbf{Z}$, and that one may form correlated groups (or clusters) of variables $\mathbf{Z}_{G_1}, \ldots, \mathbf{Z}_{G_K}$ that individually follow (1). In this case, a nested logistic structure may be appropriate. The exponent measure may be expressed as

$$V_{\text{nlog}}(\mathbf{z} \mid \alpha) = \left\{ V_{\log}(\mathbf{z}_{G_1}^{\alpha} \mid \alpha_1) + \cdots + V_{\log}(\mathbf{z}_{G_K}^{\alpha} \mid \alpha_K) \right\}^{\alpha},$$

where $\alpha, \alpha_1, \ldots, \alpha_K \in (0, 1]$ are between-groups and within-groups dependence parameters. The model is illustrated in Figure 1 with three groups; in this case, only four parameters are needed to specify the whole dependence structure of a nine-dimensional vector. This structure may easily be generalised to more complicated, potentially deeper, trees. As a result, a large family of max-stable distributions may be obtained since any given tree and vector of parameters yield a different model.

III. Inference

Inference for max-stable distributions is known to be challenging and computationally demanding [1]. To perform Bayesian inference, we rely on the Stephenson–Tawn likelihood [10], which involves the additional information about simultaneous occurrence of maxima, dramatically reducing the computational burden. To infer the unknown tree structure from the data, we propose a reversible jump Metropolis–Hastings algorithm, based on vague priors. To allow groups of variables to be totally independent (i.e., to perform dimension reduction), the prior and proposal distribution for the main dependence parameter α are defined in terms of a mixture of a continuous distribution on the interval $[0, 1]$ and a point mass at 1 (similarly to, e.g., Coles and Pauli [2]).

IV. Numerical experiments

The methods are tested on simulated and real air temperature data recorded at several sites in Switzerland over the period 1998–2007. Each site consists of two time series collected at two nearby weather stations, one in an open field and the other under forest canopy. We illustrate the flexibility of the nested logistic model in this context, and discuss possible model extensions to more complicated frameworks.

Acknowledgments

We thank Miguel de Carvalho from Pontificia Universidad Cathólica de Chile for kindly sharing the air temperature data.

References

[1] Castruccio, S., Huser, R. and Genton, M. G. (2016) High-order composite likelihood inference for max-stable distributions and processes. *Journal of Computational and Graphical Statistics, to appear* .

[2] Coles, S. G. and Pauli, F. (2002) Models and inference for uncertainty in extremal dependence. *Biometrika* **89**(1), 183–196.

[3] Coles, S. G. and Tawn, J. A. (1991) Modelling extreme multivariate events. *Journal of the Royal Statistical Society. Series B* **53**(2), 377–392.

[4] Goix, N., Sabourine, A. and Clémençon, S. (2015) Learning the dependence structure of rare events: a non-asymptotic study. *Proceedings of The 28th Conference on Learning Theory* **40**, 1–18.

[5] Gumbel, E. J. (1960a) Distributions des valeurs extrêmes en plusieurs dimensions. *Publications de l'Institute de Statistique de l'Université de Paris* **9**(294), 171–173.

[6] Gumbel, E. J. (1960b) Bivariate exponential distributions. *Journal of the American Statistical Association* **55**, 698–707.

[7] Krupskii, P. and Joe, H. (2015) Structured factor copula models: Theory, inference and computation. *Journal of Multivariate Analysis* **138**, 53–73.

[8] Stephenson, A. (2003) Simulating Multivariate Extreme Value Distributions of Logistic Type. *Extremes* **6**(1), 49–59.

[9] Stephenson, A. (2009) High-Dimensional Parametric Modelling Of Multivariate Extreme Events. *Australian & New Zealand Journal of Statistics* **51**(1), 77–88.

[10] Stephenson, B. A. and Tawn, J. (2005) Exploiting occurence times in likelihood inference for componetwise maxima pp. 213–227.

[11] Tawn, J. A. (1990) Modelling multivariate extreme value distributions. *Biometrika* **77**(2), 245–253.