Accepted Manuscript

Sparse representation of multivariate extremes with applications to anomaly detection

Nicolas Goix, Anne Sabourin, Stephan Clémençon

PII: S0047-259X(17)30406-2
DOI: http://dx.doi.org/10.1016/j.jmva.2017.06.010
Reference: YJMVA 4265

To appear in: *Journal of Multivariate Analysis*

Received date : 28 January 2016

Sparse representation of multivariate extremes with applications to anomaly detection

Nicolas Goix, Anne Sabourin, Stephan Clémençon*

*LTCI, Télécom ParisTech, Université Paris-Saclay
46, rue Barrault, 75013, Paris, France*

Abstract

Capturing the dependence structure of multivariate extreme events is a major concern in many fields involving the management of risks stemming from multiple sources, e.g., portfolio monitoring, insurance, environmental risk management and anomaly detection. One convenient (nonparametric) characterization of extreme dependence in the framework of multivariate Extreme Value Theory (EVT) is the angular measure, which provides direct information about the probable "directions" of extremes, i.e., the relative contribution of each feature/coordinate of the largest observations. Modeling the angular measure in high-dimensional problems is a major challenge for the multivariate analysis of rare events. The present paper proposes a novel methodology aiming at exhibiting a particular kind of sparsity within the dependence structure of extremes. This is achieved by estimating the amount of mass spread by the angular measure on representative sets of directions corresponding to specific sub-cones of $\mathbb{R}^d_+$. This dimension reduction technique paves the way towards scaling up existing multivariate EVT methods. Beyond a non-asymptotic study providing a theoretical validity framework for our method, we propose as a direct application a first anomaly detection algorithm based on multivariate EVT. This algorithm builds a sparse normal profile of extreme behaviors, to be confronted with new (possibly abnormal) extreme observations. Illustrative experimental results provide strong empirical evidence of the relevance of our approach.

Keywords: Anomaly Detection, Dimensionality Reduction, Multivariate Extremes, VC theory

1. Introduction

1.1. Context: multivariate extreme values in large dimension

Extreme Value Theory (EVT in abbreviated form) provides a theoretical basis for modeling the tails of probability distributions. In many applied fields where rare events may have a disastrous impact, such as finance, insurance, climate, environmental risk management, network monitoring [23, 42] or anomaly detection [8, 31], the information carried by extremes is crucial. In a multivariate context, the dependence structure of the joint tail is of particular interest, as it gives access to probabilities of a joint excess above high thresholds or to multivariate quantile regions. Also, the distributional structure of extremes indicates which components of a multivariate quantity may be simultaneously large while the others remain small, which is a valuable piece of information for multi-factor risk assessment or detection of anomalies among other non abnormal extreme data.

In a multivariate "Peak-Over-Threshold" setting, realizations of a d-dimensional random vector $\mathbf{Y} = (Y_1, \ldots, Y_d)$ are observed and the goal pursued is to learn the conditional distribution of excesses, $\mathbf{Y} \mid \|\mathbf{Y}\| \geq r$, above some large threshold $r > 0$. The dependence structure of such excesses is described via the distribution of the directions formed by the most extreme observations, the so-called angular measure, hereafter denoted by Φ. The latter is defined on the positive orthant of the $(d-1)$-dimensional hyper-sphere. To wit, for any region A on the unit sphere (a set of "directions"), after suitable standardization of the data (see Section 2), $C\Phi(A) \simeq \Pr(\|\mathbf{Y}\|^{-1}\mathbf{Y} \in A \mid \|\mathbf{Y}\| > r)$, where

*Corresponding author.
Email address: `stephan.clemencon@telecom-paristech.fr` (Stephan Clémençon)