Journal of Multivariate Analysis 136 (2015) 126–146

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Semi-parametric modeling of excesses above high multivariate thresholds with censored data

Anne Sabourin

Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI, Département TSI, 37-38, rue Dareau, 75014 Paris, France

ARTICLE INFO

Article history:
Received 9 July 2014
Available online 25 January 2015

AMS 2010 subject classifications:
62G32
62H12
62G07
62F15
62N01
62N02
62P12

Keywords:
Multivariate extremes
Censored data
Data augmentation
Semi-parametric Bayesian inference
MCMC algorithms

ABSTRACT

How to include censored data in a statistical analysis is a recurrent issue in statistics. In multivariate extremes, the dependence structure of large observations can be characterized in terms of a non parametric *angular measure*, while marginal excesses above asymptotically large thresholds have a parametric distribution. In this work, a flexible semi-parametric Dirichlet mixture model for angular measures is adapted to the context of censored data and missing components. One major issue is to take into account censoring intervals overlapping the extremal threshold, without knowing whether the corresponding hidden data is actually extreme. Further, the censored likelihood needed for Bayesian inference has no analytic expression. The first issue is tackled using a Poisson process model for extremes, whereas a data augmentation scheme avoids multivariate integration of the Poisson process intensity over both the censored intervals and the failure region above threshold. The implemented MCMC algorithm allows simultaneous estimation of marginal and dependence parameters, so that all sources of uncertainty other than model bias are captured by posterior credible intervals. The method is illustrated on simulated and real data.

1. Introduction

Data censoring is a commonly encountered problem in multivariate statistical analysis of extreme values. A 'censored likelihood' approach makes it possible to take into account partially extreme data (non concomitant extremes): coordinates that do not exceed some large fixed threshold are simply considered as left-censored. Thus, the possibly misleading information carried by non-extreme coordinates is ignored, only the fact that they are not extreme is considered [26,18,27], see also Thibaud and Opitz [31] or Huser et al. [16]. However, there are other situations where the original data is incomplete. For example, one popular way to obtain large sample sizes in environmental sciences in general and in hydrology in particular, is to take into account data reconstructed from archives, which results in a certain amount of left- and right-censored data, and missing data. As an example, what originally motivated this work is a hydrological data set consisting of daily water discharge recorded at four neighboring stations in the region of the Gardons, in the south of France. The extent of systematic recent records is short (a few decades) and varies from one station to another, so that standard inference using only 'clean' data is unfeasible (only 3 uncensored multivariate excesses of 'large' thresholds – fixed after preliminary uni-variate analysis – are recorded). Historical information is available, starting from the 17th century, a large part of it being censored: only major floods are recorded, sometimes as an interval data (*e.g.* 'the water level exceeded the parapet but the Mr. X's house was spared'). These events are followed by long 'blank' periods during which the previous record was

E-mail address: anne.sabourin@telecom-paristech.fr.

http://dx.doi.org/10.1016/j.jmva.2015.01.014
0047-259X/

not exceeded. Uni-variate analysis for this data set has been carried on by Neppel et al. [19] but a multivariate analysis of extremes has never been accomplished, largely due to the complexity of the data set, with multiple censoring.

While modeling multivariate extremes is a relatively well marked out path when 'exact' (non censored) data are at stake, many fewer options are currently available for the statistician working with censored data. The aim of the present paper is to provide a flexible framework allowing multivariate inference in this context. Here, the focus is on the methodology and the inferential framework is mainly tested on simulated data with a censoring pattern that resembles that of the real data. A detailed analysis of the hydrological data raises other issues, such as, among others, temporal dependence and added value of the most ancient data. These questions are addressed in a separate paper, intended for the hydrological community [24].[1]

Under a standard assumption of multivariate regular variation (see Section 2), the distribution of excesses above large thresholds is characterized by parametric marginal distributions and a non-parametric dependence structure that is independent from threshold. Since the family of admissible dependence structures is, by nature, too large to be fully described by any parametric model, non-parametric estimation has received a great deal of attention in the past few years [6,7,12]. To the best of my knowledge, the non parametric estimators of the so-called *angular measure* (which characterizes the dependence structure among extremes) are only defined with exact data and their adaptation to censored data is far from straightforward.

For applied purposes, it is common practice to use a parametric dependence model. A widely used one is the Logistic model and its asymmetric and nested extensions [13,4,29,28,8]. In the logistic family, censored versions of the likelihood are readily available, but parameters are subject to non linear constraints and structural modeling choices have to be made *a priori*, *e.g.*, by allowing only bi-variate or tri-variate dependence between closest neighbors.

One semi-parametric compromise consists in using mixture models, built from a potentially infinite number of parametric components, such as the Dirichlet mixture model (DM), first introduced by Boldi and Davison [2]. They have shown that it can approach arbitrarily well any valid angular measure for extremes. A re-parametrized version of the DMmodel [23], allows for consistent Bayesian inference – thus, a straightforward uncertainty assessment using posterior credible sets – with a varying number of mixture components *via* a *reversible-jumps* algorithm. The approach is appropriate for data sets of moderate dimension (typically, $d \approx 5$).

The purpose of the present work is to adapt the DM model to the case of censored data. The difficulties are two-fold: First, from a modeling perspective, when the censoring intervals overlap the extremal thresholds (determined by preliminary analysis), one cannot tell whether the event must be treated as extreme. The proposed approach here consists in reformulating the *Peaks-over-threshold* (POT) model originally proposed by Boldi and Davison [2] and Sabourin and Naveau [23], in terms of a *Poisson model*, in which the censored regions overlapping the threshold have a well-defined likelihood. The second challenge is numerical and algorithmic: for right-censored data above the extremal threshold (not overlapping it), the likelihood expression involves integrals of a density over rectangular regions, which have no analytic expression. The latter issue is tackled within a data augmentation framework, which is implemented as an extension of Sabourin and Naveau [23]'s algorithm for Dirichlet mixtures.

An additional issue addressed in this paper concerns the separation between marginal parameters estimation and estimation of the dependence structure. Performing the two steps separately is a widely used approach, but it boils down to neglecting marginal uncertainty, which confuses uncertainty assessment about joint events such as probabilities of failure regions. It also goes against the principle of using regional information together with the dependence structure to improve marginal estimation, which is the main idea of the popular *regional frequency analysis* in hydrology. In this paper, simultaneous inference of marginal and dependence parameters in the DM model is performed, which amounts in practice to specifying additional steps for the marginal parameters in the MCMC sampler.

The rest of this paper is organized as follows: Section 2 recalls the necessary background for extreme values modeling. The main features of the Dirichlet mixture model are sketched. This POT model is then reformulated as a Poisson model, which addresses the issue of variable threshold induced by the fluctuating marginal parameters. Censoring is introduced in Section 3. In this context, the Poisson model has the additional advantage that censored data overlapping threshold have a well defined likelihood. The lack of analytic expression for the latter is addressed by a data augmentation scheme described in Section 4. The method is illustrated by a simulation study in Section 5: marginal performance in the DM model and in an independent one (without dependence structure) are compared, and the predictive performance of the joint model in terms of conditional probabilities of joint excesses is investigated. The model is also fitted to the hydrological data. Section 6 concludes. Most of the technicalities needed for practical implementation, such as computation of conditional distributions, or details concerning the data augmentation scheme and its consistency are relegated to the Appendix.

2. Model for threshold excesses

2.1. Dependence structure model: angular measures

In this paper, the sample space is the d-dimensional Euclidean space $\mathbb{R}^d$, endowed with the Borel σ-field. In what follows, bold symbols denote vectors and, unless otherwise mentioned, binary operators applied to vectors are defined component-wise. Let $(\mathbf{Y}_t)_{t\in\mathbb{N}}$ be independent, identically distributed (*i.i.d.*) random vectors in $\mathbb{R}^d$, with joint distribution $\mathbf{F}$ and margins

[1] Preprint available at https://hal.archives-ouvertes.fr/hal-01087687.