Statistical Theory of Extremes
Homepage: http://www.gathacognition.com/book/gcb14
http://dx.doi.org/10.21523/gcb1

Part 5
Complements

Chapter 16

GEA: General Extremes Analysis: A guide in natural language

J. Tiago de Oliveira
Academia das Ciências de Lisboa (Lisbon Academy of Sciences), Lisbon, Portugal.

Abstract

This chapter reviews the step-by-step writing of a program for the computer in the appropriate computer language. The model explained in four steps: Statistical Choice of Models (SCM), Statistical Analysis of Models (SAM), Statistical Decision for Multivariate Models (SDM) and Stochastic Processes Analysis (SPA). The exact distribution of extremes is approximated by one of the asymptotic distributions.

Published Online

23 June 2017

Keywords

Non-diagram fluxogram; Statistical Choice of Model; Multivariate Model; Stochastic Processes.

Editor(s)

J.C. Tiago de Oliveira

16.1 Introduction

This chapter is not a program for *GEA*, which would differ according to the language used, nor a set of recipes for the concrete analysis of data. It presupposes knowledge of the corresponding chapters and the writing of a program for the computer used in the appropriate computer language. This guide is, at most, a non-diagram fluxogram of the procedures to be implemented in each computer + language system. In fact, with every system, we can obtain (to some decimal points of accuracy) the results that will be analysed in the Case Studies chapter. In general we will think of maxima (largest values) of samples.

Originally published in 'Statistical Analysis of Extremes', 1997, 2016
http://dx.doi.org/10.21523/gcb1.1727

This sequence of procedures depends, evidently, on the *formulae* to be found in Parts 2, 3 and 4. *We assume, throughout, that the exact distribution of extremes is approximated by one of the asymptotic distributions*.

16.2 Step 1 — Statistical Choice of Models (SCM)

This initial step, for univariate distributions (or margins in the multivariate case) leads to the (preliminary) model to be fitted to the data and subsequently used. It can be omitted if there exist justifications (such as theoretical results, previous experience, etc.) to justify some assumption and, thus, to act in such a way. Although in this phase we integrate a geometrical plotting of data, its purpose is not to justify the statistical choice made analytically but to *support intuition* and to clarify for experimentalists the use of the assumed model.

Thus, after filing the data — for each margin if the data are multivariate — we can do the following types of *GEA*, which are not mutually exclusive.

For graphical choice, after ordering the univariate (or univariate margins) data, we plot them on a Gumbel probability paper, as said in Chapter 4 and choose one of the 3 models (Weibull, Gumbel and Fréchet) according to the "eye-curvature" of the plotted points.

This graphical choice, although attractive and intuitive, must be confirmed by the use of the analytical choice, through the test statistic $\widehat{V}_n$ and the associated decision rule (Chapter 8). Recall that $\widehat{V}_n$ presupposes that the location-dispersion parameters of a possible Gumbel distribution are estimated as said before.

This trilemma decision, analogous to an LMPU two-sided test, splits the next step in three directions, according to the chosen model or alternative (Weibull, Gumbel or Fréchet). In this case we choose between $\theta < 0$ (Weibull), $\theta = 0$ (Gumbel) and $\theta > 0$ (Fréchet) distributions (in the von Mises-Jenkinson formula), excluding thus the distributions of extremes that are not attracted to any of the limiting distributions.

In the cases where we can assume, with sufficient information, that $\theta \geq \theta_0 \geq 0$ or $\theta \leq \theta_0 \leq 0$, we can have a usual LMP test, also using the test statistic $\widehat{V}_n$.

Another (weak) check of the sign of $\theta(< 0, = 0, > 0)$ can be made by using the statistic Q_n (Chapter 8), intuitive but not as efficient as $\widehat{V}_n$.

A comment: if the graphical choice and the analytical choice (with $\widehat{V}_n$ and its ancillary, Q_n) show a clear-cut discrepancy we can conjecture that the asymptotic approximation was not obtained for the sample sixe(n) used, that the i.i.d. condition is not (approximately) verified, or even that there does not exist an attraction to any of the asymptotic distributions. A pragmatical approach must then be tried.

16.3 Step 2— Statistical Analysis of Models (SAM)

After *assuming* (i.e., having made the statistical choice of) one of the asymptotic distributions (of the univariate margins, in the multivariate case) the next essential step is to obtain estimators of the parameters. If the Gumbel distribution is assumed then $(\hat{\lambda}, \hat{\delta})$ have been obtained in the first step, and they can be used for the different problems described in Chapter 5 and possibly others. If the Gumbel distribution is not assumed (i.e., $\theta \neq 0$) the techniques relative to the assumption of the Fréchet distribution ($\theta > 0$) are described in Chapter 6 and those for the assumption of the Weibull distribution ($\theta < 0$) are dealt with (but for minima) in Chapter 7 .

In all these cases, parameters and quantiles are estimated as well as probabilities of over- and underpassing threshold levels, and tests of hypotheses, prediction, discrimination, tolerance intervals and multisample analysis have also been described when possible.

Recall that if the location parameter is known we can easily reduce the Weibull and Fréchet models to the Gumbel one by exp/log transformations.

The tail estimation must be made according the third approach given in Annex 4.

16.4 Step 3 — Statistical Decision for Multivariate Models (SDM)

Let us suppose that we have vectors of extremes, whose margins have been studied (Step 1) and assumed to have one of the asymptotic distributions. The case for bivariate extremes is, in part, dealt with theoretically but the multivariate case is very partially covered.

Various bivariate models exist (Chapter 10) but it *seems*, for the moment, that essentially the logistic and in second place the biextremal and natural models most often fit to the data; the last two are very easily generalized to the multivariate case.

In the bivariate case, the most important initial step is to test for independence as a reference pattern, although we can in general expect dependence. Once the dependence is confirmed and a *bivariate model assumed* we should estimate the dependence parameters and act accordingly. Chapter 11 gives the essential formulae. But it should be recalled, if we are dealing with the multivariate case, that if there exists independence in the bivariate margins than we should always have independence, and we are returning to Step 2.

Sometimes, and at this state of knowledge, the estimation of the dependence function can be made intrinsically and compared with the supposed and assumed bivariate distribution, in an intuitive way.

For the multivariate models, after the bivariate studies are made, we can *try* to assume one the bivariate models for the corresponding margins. The logistic, the biextremal, and the natural model may be the best fit to data.

16.5 Step 4 — Stochastic Processes Analysis (SPA)

In principle, although it is not necessarily always the case, we may suppose that every stochastic process (of extremes), by a choice of observation times, is reduced to a convenient random sequence (originated from random or periodic sampling); then, with a discrete time data, we will follow the procedures for random sequences; otherwise we will follow directly the procedures for stochastic processes.

As strong assumptions are made about the margins in general Step 2 may be of interest.

As a first approach, after filing the data and plotting it against a time variable (the data may be a sequence or a process) we try to "see" if it follows globally one of the patterns described in Annex 6.

Then we try to fit of curvilinear regression to the observed sequence: linear in $\log n$ for extremal sequences, constant for EMS sequences, asymptotically linear in n or constant according to $a_0 > 0$ or $a_0 < 0$ for EME sequences, and linear in n for sliding sequences; recall that in all cases variance is constant. For stochastic processes fix a net of points, at equal steps, and consider the sequence of observations at each point of the net; it will be convenient to use various nets to *reinforce the assumed model*.

When dealing with random sequences:

a) if the extremal sequence model is assumed, use formulae of Chapter 14;

b) if the EMS sequence model is assumed, use formulae of Chapter 15;

c) if the EME sequence model is assumed, use formulae of Chapter 15;

d) if the sliding sequence model is assumed, use formulae of Chapter 15.

When dealing with stochastic processes:

a) if the extremal process model is assumed, use formulae of Chapter 14;

b) if the EMS process is assumed, use formulae of Chapter 15.

*****