EXTREMAL REGRESSION WITH APPLICATIONS TO ECONOMETRICS, ENVIRONMENT AND FINANCE

Plan de gestion de données créé à l'aide de DMP OPIDoR

Créateurs du PGD : Abdelaati Daouia, Jean-Pierre Florens, Laurent Gardes, Stéphane Girard, Armelle Guillou, Thibault Laurent, Gilles Stupfler, Antoine Usseglio-Carleve.

Affiliation du créateur principal : Université Toulouse Capitole

Modèle du PGD : Science Europe - DMP template (english)

Dernière modification du PGD : 10/04/2020

Financeur : ANR

Numéro de subvention : ANR-19-CE40-0013-01

Résumé du projet :
The project concentrates around three themes that are central to the area of modern extreme value statistics. First, we contribute to the expanding literature on non-regular regression models where the regression function describes some frontier or boundary curve. This is motivated from many applications especially in production econometrics. We address several issues related to this topic, namely polynomial spline fitting under shape constraints, estimation from noisy data using inverse problems, and estimation of locally stationary, one-sided autoregressive processes. Second, we further investigate the recent extreme value theory built on the use of asymmetric least squares instead of order statistics. We focus on two least squares analogues of quantiles (expectiles and extremiles) which have gained increasing interest in risk management. Finally, we explore the important problem of estimating conditional and joint extremes in high dimension, which is still in full development. Our research is interdisciplinary in nature, involving statistics and risk management in environment, actuarial science and finance. We plan to (re)use existing data that are publicly available online (e.g. Demographic and Health Survey, Chicago air pollution database, Federal Insurance and Mitigation Administration National Flood Insurance Program, National Water Information System, Tsunami Causes and Waves, Yahoo Finance, Global Centroid Moment Tensor database, etc) or that can be found in some known R packages (e.g. CASdatasets, insuranceData, npbr, quantreg.nonpar).

Chercheur Principal : Abdelaati Daouia

Contact pour les Données : Abdelaati Daouia

1. DATA DESCRIPTION AND COLLECTION OR RE-USE OF EXISTING DATA

1a. How will new data be collected or produced and/or how will existing data be re-used?

We employ in our numerical illustrations existing data that are available and secured online. There are no constraints on (re)use of these data. Their provenance is documented in our papers by refereeing to the institutions that collected the data and to their URL, as well as the articles and/or books where the data were initially studied:

- The childhood malnutrition data (india) in Daouia, A. and D. Paindaveine (2020), "Multivariate Expectiles, Expectile Depth and Multiple-Output Expectile Regression", can be found in the R package *quantreg.nonpar:*

  https://cran.r-project.org/web/packages/quantreg.nonpar/.

- The French postal data (post) in Daouia, A., Florens, J-P. and L. Simar (2020), "Robustified expected maximum production frontiers", can be found in the R package npbr:

  https://cran.r-project.org/web/packages/npbr/index.html.

- The Air Data (airdata) in Goegebeur, Guillou, Le Ho and Qin (2020), "Robust nonparametric estimation of the conditional tail dependence coefficient", and in Escobar-Bach, M., Goegebeur, Y. and A. Guillou (2020), "Bias correction in conditional multivariate extremes", are publicly available at:

  https://aqsdr1.epa.gov/aqsweb/aqstmp/airdata/download_files.html.

- The Chicago air pollution database (NMMAPS) in Gardes, L (2020), "Nonparametric confidence interval for conditional quantiles with large-dimensional covariates", is available on the R package NMMAPS Data Lite:

  http://www2.uaem.mx/r-mirror/web/packages/NMMAPSlite/.

- The fire losses data (Norfire) in Gardes, L. and S. Girard (2019), "On the estimation of the variability in the distribution tail", are available on the R package CASdatasets:

  http://cas.uqam.ca.

- The insurance dataset (dataOhlsson) in Girard, S., Stupfler, G. and A. Usseglio-Carleve (2019), "Nonparametric extreme conditional expectile estimation", can be found in the R package insuranceData:

  https://cran.r-project.org/web/packages/insuranceData/.

- The flood claim data (NFIP) in Goegebeur, Y., Guillou, A., Le Ho, N.K. and J. Qin (2019), "Conditional marginal expected shortfall", are publicly available at:

  https://www.fema.gov/media-library/assets/documents/180374.

- The database (waterdata) of daily river flows of the Salt River near Roosevelt-Arizona in Girard, S., Stupfler, G. and A. Usseglio-Carleve (2020), "An Lp−quantile methodology for estimating extreme expectiles", is available at:

  https://waterdata.usgs.gov/nwis/inventory (station 09497500).

- The Tsunami Causes and Waves (tsunami) dataset in Ahmad, A., Deme, E., Diop, A. Girard, S. and A. Usseglio-Carleve (2020), "Estimation of extreme quantiles from heavy-tailed distributions in a location-dispersion regression model", is available at:

  https://www.kaggle.com/noaa/seismic-waves.

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

Making use mainly of R and Matlab softwares, the utilized data consist of numeric databases under formats that facilitate sharing and long-term re-use of data such as csv, txt, xls, RData, mat, etc. The volumes of our datasets are variable. The sample sizes typically vary between some hundreds and tens of thousands. Yet, they can easily be stored and manipulated on our own laptops:

- india (Daouia and Paindaveine, 2020): one R Data file, 7.3 mega bytes, rows=37623, columns=35.

- post (Daouia, Florens and Simar, 2020): one R Data file, 0.06 mega bytes, 4000 rows, 4 columns.

- airdata (Goegebeur, Guillou, Le Ho and Qin (2020) and Escobar-Bach, Goegebeur and Guillou (2020)) : 1 file, 2 mega bytes, 5 variables, 57303 rows.

- nmmaps (Gardes, 2020): one R Data file, 4.4 mega bytes, 5114 rows, 14 columns.

- norfire (Gardes and Girard, 2019): one R Data file, 0.074 mega bytes, 9181 rows, 3 columns.

- dataOhlsson (Girard, Stupfler and Usseglio-Carleve, 2019): 2 584 248 bytes, 1 object, 1 file, 9 columns, 64 548 rows.

- nfip (Goegebeur, Guillou, Le Ho and Qin, 2019): 1 file, 574.1 mega bytes, 40 variables, 2431534 rows.

- waterdata (Girard, Stupfler and Usseglio-Carleve, 2020): 1 object, 1 file, 2 columns, 31 777 rows, 2 659 304 bytes.

- tsunami (Ahmad, Deme, Diop, Girard and Usseglio-Carleve, 2020): 1 object, 2 files. The first file "sources" is not used (45 columns, 2 583 rows, 249.82 KB). The second file we use is "waves": 30 columns, 26 204 rows, 2.51 mega bytes.

2. DOCUMENTATION AND DATA QUALITY

2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

For each database considered in our articles, the corresponding metadata and documentation are already provided in the source URLs and packages pointed out above.

2b. What data quality control measures will be used?

The consistency and quality of data collection were controlled and documented as explained in the links/websites above, provided by the organisms that originally collected the data and/or the R packages where the datasets are available.

3. STORAGE AND BACKUP DURING THE RESEARCH PROCESS

3a. How will data and metadata be stored and backed up during the research?

The online available databases we use in our research activities are stored and backed up in the links described above and can be downloaded in any time for reuse. Yet, in addition to our laptops and stand-alone hard drives, copies of our codes and data are stored in the secure servers that are managed by IT services of our home institutions.

3b. How will data security and protection of sensitive data be taken care during the research?

Given that the data we are using are publicly available and already shared online, they can be recovered anytime in the event of an incident in, e.g., our laptops or the servers of our home institutions.

4. LEGAL AND ETHICAL REQUIREMENTS, CODES OF CONDUCT

4a. If personal data are processed, how will compliance with legislation on personal data and on security be ensured?

We do not deal with personal data or any data that require compliance with data protection laws.

4b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

The databases we are using are publicly available and already shared online. Their (re)use does not affect intellectual property rights.

4c. What ethical issues and codes of conduct are there, and how will they be taken into account?

Ethical issues and codes of conduct do not apply in the case of our (re)utilized data.

5. DATA SHARING AND LONG-TERM PRESERVATION

5a. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

The data are already available online and can easily be downloaded. There are no constraints on their use.

5b. How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)?

The data are preserved long-term in the links described above (e.g. Cran repository) and in the servers of our home institutions.

5c. What methods or software tools are needed to access and use data?

Like us, potential users do not need specific tools to access and use the data.

5d. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

As indicated above, we use already available data.

6. DATA MANAGEMENT RESPONSIBILITIES AND RESOURCES

6a. Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?

Again we are not the owners of the data we use in our numerical illustrations and applications. The authors of each article, produced within the ANR project "ExtremReg", are responsible for the management of the dataset they have freely downloaded from the sites given above.

6b. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

As a matter of fact, we just download the data that were already prepared by their owners for sharing/preservation.