Below is a list of my current research interests, together with some references. Within these domains, I am happy to propose specific topics tailored to the interests and skills of the student, both in statistics and data science. Feel free to contact me so that we can discuss.

**Extreme value analysis.** Rather than in mean effects, one is often interested in high quantiles, univariate or multivariate, with or without covariates. Extreme value analysis provides models and inference tools to estimate such quantiles \([3]\). Recently, methods from extreme value analysis are being applied in machine learning too, for instance in open set classification and anomaly detection \([11,13]\). On the other hand, machine learning methods such as quantile random forests are finding their way into extreme value analysis \([8]\). Extreme value analysis is also the domain for the project on “Solar flares – Pareto versus lognormal distribution”, jointly with the Royal Observatory, described in a separate document.

**Monte Carlo integration.** For integrating a function of several variables, Monte Carlo methods are about the only ones that do not suffer too much from the curse of dimensionality, thanks to the central limit theorem. The variance of the Monte Carlo estimator of the integral can be reduced by the use of control variates, a method with a surprising connection to linear regression \([10]\). Variable selection methods (lasso) can significantly boost the performance of Monte Carlo integration algorithms based on control variates \([6]\). Other regularization methods such as ridge regression, principal component regression, and partial least squares, remain to be explored.

**Optimal transport.** The earth mover distance or more generally the Wasserstein metric is a way to quantify the distance between two probability distributions \([9]\). It provides a powerful tool to estimate parameters, compare models, test hypotheses, and more \([2, 5]\). The Wasserstein barycenter provides a notion of mean or average that is often more appropriate for complex data structures than traditional means \([7]\). Identifying images or pictures with (discrete) bivariate distributions, methods from optimal transport can be used to define distances between pictures and thereby apply classification and clustering algorithms to samples of pictures \([12]\). For circular data, optimal transport is a promising avenue for defining ranks and constructing dependence models.

\[\text{johan.segers@uclouvain.be}\]
Graphical models. The dependence between random variables in a system can often be thought of as being organized in a network. Graphical models form a way to model such dependence structures using conditional independence relations. Gaussian graphical models have been and still are very popular. In the context of extreme value analysis, their role is being taken over by the multivariate Hüsler–Reiss distribution [4]. For such models, questions of interest are how to learn the graph structure and how to calibrate the parameters [1].

References


Solar flares are large and abrupt increases of the photon flux, observed in a wide spectral range of electromagnetic radiation. They result from the sudden release of stored magnetic energy, which may reach up to $10^{26}$ joule in a matter of a few hours or even minutes.

It is generally accepted that the distribution of solar flare parameters such as the peak intensity of a flare is described by a power law or Pareto distribution. The exact behavior of flare distributions has important implications for large-scale science questions such as coronal heating and the nature of solar flares. Studying a recent data set of 8274 flares, we found clear signs that the peak flare intensity distribution is not well-described by a Pareto distribution, but instead a lognormal fit describes the data significantly better. Figure 1 shows the data set and Figure 2 shows the empirical Complementary Cumulative Distribution Function (CCDF, black) along with the Pareto (red) and lognormal (green) fits.

To process the data, we wrote R code employing the poweRlaw package (Gillespie 2015), based on (Clauset et. al. 2009). This method calculates:

- the Maximum Likelihood Estimation (MLE) fit of a Pareto or lognormal model to the data,
- a goodness of fit to evaluate an individual fit,
- a log likelihood ratio test to compare how well two different distributions describe the data.

Our lower limit $x_{\text{min}}$ for the (Pareto or lognormal) model is selected as follows: for every value of $x_{\text{min}}$ we calculate the Kolmogorov-Smirnov (KS) distance between the empirical distribution and the MLE fit. We then select the $x_{\text{min}}$ which minimizes this KS distance.

Our goodness of fit is obtained in the following way: starting from the MLE fit for the selected $x_{\text{min}}$, we generate 5000 synthetic data sets of the same size with the same parameters. For every such data set, we repeat our procedure, i.e., we determine the optimal lower limit $x_{\text{min}}$ and MLE fit for that data set, and we calculate the KS distance between the synthetic data set and its MLE fit. The goodness of fit is the fraction of synthetic data sets for which the KS distance is larger than the KS distance between the original data set and its MLE fit. If this fraction is below 0.1, less than 10% of the KS distances between our model distribution and samples from this distribution are expected to exceed the KS distance between the original data set and its MLE fit.
Hence, the KS distance between the original data and its MLE fit cannot be explained by statistical variation due to sampling, and the hypothesis that our model describes the original data set well, is rejected.

Figure 1: Overview of the data points as a function of time. Every data point corresponds to a unique flare detected by Solar Demon and shows the peak intensity of the corresponding flare. The darkness of any pixel in this plot is proportional to the number of data points that overlap with the pixel. The data points were corrected for the degradation of the SDO/AIA 9.4 nm channel over time.

The goal of the MSc thesis is to investigate which statistical tools are best fitted for this data set, and to evaluate the merits of the poweRlaw package as compared to alternative methods and software. To understand the phenomena behind solar flares, it is important to know which kind of theoretical distribution fits the data best. We want to evaluate this question in a robust statistical way, which is why we value a comparison of different methods.

The following aspects could be investigated both at the theoretical level and in terms of comparison with our present method:

- Alternative methods to select the lower limit $x_{\text{min}}$
- Alternative fitting methods, (e.g., nonparametric estimation)
- Alternative goodness of fit methods

In addition, it will be interesting to investigate:
- Uncertainty quantification, i.e., estimation of the errors made when estimating $x_{\text{min}}$ and other parameters, in case of MLE and other fitting methods.
- Alternative methods for a direct comparison between different models: Kullback-Leibler divergence, Akaike Information Criterion, …

Figure 2: The black curve shows the CCDF of the SDO/AIA peak flare intensity. The red and green curves are the CCDF of the MLE power law respectively lognormal fit to the data. The vertical red and green lines, respectively, are the lower limits for which the power law and lognormal fit are valid.

References