TP1 Statistical learning with extreme values

Anne Sabourin, Antoine Doizé

October, 10, 2024

1 Fire Damage Data in Denmark

The dataset `danish` contains the largest insurance claims related to fire damages between 1980 and 1990 (from Thursday 3rd January 1980 until Monday 31st December 1990).

1.1 Stationarity and Gaussianity

Check quickly that a stationary, independent model is suitable and that a Gaussian model is not. To do so:

- Plot the time series.
- Compute the auto-correlation function (ACF). Refer to this documentation for more details.
- Use the Dickey-Fuller stationarity test `ts.adfuller` (documentation link) or other known tests for stationarity.
- For the Gaussianity test, use, for example, a QQ-plot and a Shapiro test (documentation link).

1.2 Exploratory Data Analysis of the Peak Over Threshold (POT)

Make different Exploratory Data Analysis (EDA) plots to check if a POT analysis is relevant and adapted. You can try:

- Quantile plot for Pareto model (use the `pareto_plot` function provided at the beginning of the notebook). *Warning: takes a* `numpy.array` *as input.*

- Mean residual life plot (use the plot_mean_residual_life from the pyextremes library: documentation link). *Warning: takes a pandas.Series as input.*

- Hill plot (use the hill function provided at the beginning of the notebook). *Warning: takes a numpy.array as input.*

1.3 Generalized Pareto Distribution (GPD) Fit

- Find a relevant threshold to fit your model using the diagnostic methods from question 2.

- Use the plot_parameter_stability function from pyextremes (documentation link). *Note: This library is designed for time series. You may need to adapt the index to make it a time series and adjust the r parameter, which is the declustering window.*

  For example:

  ```
  new_index = pd.to_datetime(df_danish.index)
  df_danish.index = new_index
  delta_declustering = df_danish.index[1] - df_danish.index[0]
  ```

- To fit a GPD, you can use scipy.stats.genpareto.fit (documentation link):

  ```
  # Demo of genpareto library
  data_simu = stats.genpareto.rvs(c=0.5, loc=5, scale=10, size=10000)
  xi_fit, mu_fit, sigma_fit = stats.genpareto.fit(data_simu)
  print(xi_fit, mu_fit, sigma_fit)
  ```

1.4 Comparison with GPD Fit with $\xi = 0$ Constraint

- First, recall the limit of the GPD density (or CDF) when $\xi \to 0$. You should find a known distribution.

- Fit a GPD with $\xi = 0$ to the data.

- Compare both models (model $\mathcal{M}_0$ with constraint $\xi = 0$ and model $\mathcal{M}_1$ with free ξ). For this you could use an AIC based analysis, or use a test such as the likelihood ratio (see reminder below). Can you tell how relevant is it to have $\xi = 0$?

***Reminder for the likelihood-ratio test**: Let $\mathcal{M}_1$ be a model, and $\mathcal{M}_0$ a subset model of $\mathcal{M}_1$ (which means that $\forall \theta_0 \in \Theta_0$, the parameter set of model $\mathcal{M}_0$, we have $\theta_1 \in \Theta_1$ (the parameter set of model $\mathcal{M}_1$), such that $\theta_1 = (\theta_0, \theta^{(1)})$).*
Then under $\mathbb{H}_0$ and suitable regularity conditions (we will suppose those conditions are met):

$$T := 2\ln\left(\frac{\sup_{\theta\in\Theta_1} \mathcal{L}(\theta) \mid x_n}{\sup_{\theta\in\Theta_0} \mathcal{L}(\theta) \mid x_n}\right) \xrightarrow{n\to\infty,d} \chi^2_k$$

with

- $\sup_{\theta\in\Theta} \mathcal{L}(\theta) \mid x_n$ *the maximum likelihood on the set Θ, on n observed values x_n*
- $k = \dim(\Theta_1) - \dim(\Theta_0)$, *the difference of degrees of liberty of the two models.*

1.5 Return Levels

Provide an estimate for the 50-year return level based on the estimated parameters. Additionally, provide another estimate using the Weissman estimator.

1.6 Probability of High Quantiles

Provide an estimate for the probability of occurrence over a year of an excess above:

- Twice the maximum observed over the considered period.
- The maximum observed over the considered period.

Compare these with the empirical estimator.

2 Annual Maxima of Sea Level at Port Pirie (Australia)

This dataset consists of daily sea level recordings for a period of 48 years.

2.1 Data Analysis

- Perform the same data analysis as in the previous exercise (such as stationarity, Gaussianity, exploratory data analysis, etc.).

2.2 Generalized Extreme Value (GEV) Fit

- Try to fit a Generalized Extreme Value (GEV) distribution to the data. To fit a GEV, you can use the `genextreme` package. **Caution:** Refer to the documentation for details, as the shape parameter c is defined such that $c = -\xi$. You can find the documentation here: scipy.stats.genextreme.
- Then fit a GEV distribution with the constraint $\xi = 0$.
  You may use the Gumbel function provided in the Util section of the notebook, or use the gumbel_r library of scipy : check the documentation here: scipy.stats.gumbel_r.
- Discuss and provide insights on which model might be more appropriate to choose (you could use an AIC based analysis, or use a test such as the likelihood ratio test)

2.3 One Thousand Year Return Level Estimate

Provide an estimate of the one thousand-year return level using both models (GEV and GEV with $\xi = 0$).

3 Rain Data in South England

3.1 Exploratory Data Analysis

- Perform the same Exploratory Data Analysis (EDA) as in previous questions to check if the data shows any seasonality and whether it can be modeled by a Gaussian distribution.

3.2 Return Level Estimation

Propose two methods to estimate the 100-year return level:

- Based on block-maxima analysis.
- Based on peaks over threshold (POT) analysis.

3.3 Comparison of Models

Compare the estimations obtained from both models (block-maxima and POT).