Reproducible research

The place of probability distributions in statistical learning. A commented book review of “Distributions for modeling location, scale, and shape using GAMLSS in R” by Rigby et al. (2021)

Raydonal Ospina

2022-03-31

Robust estimation with GAMLSS: Dealing with outliers

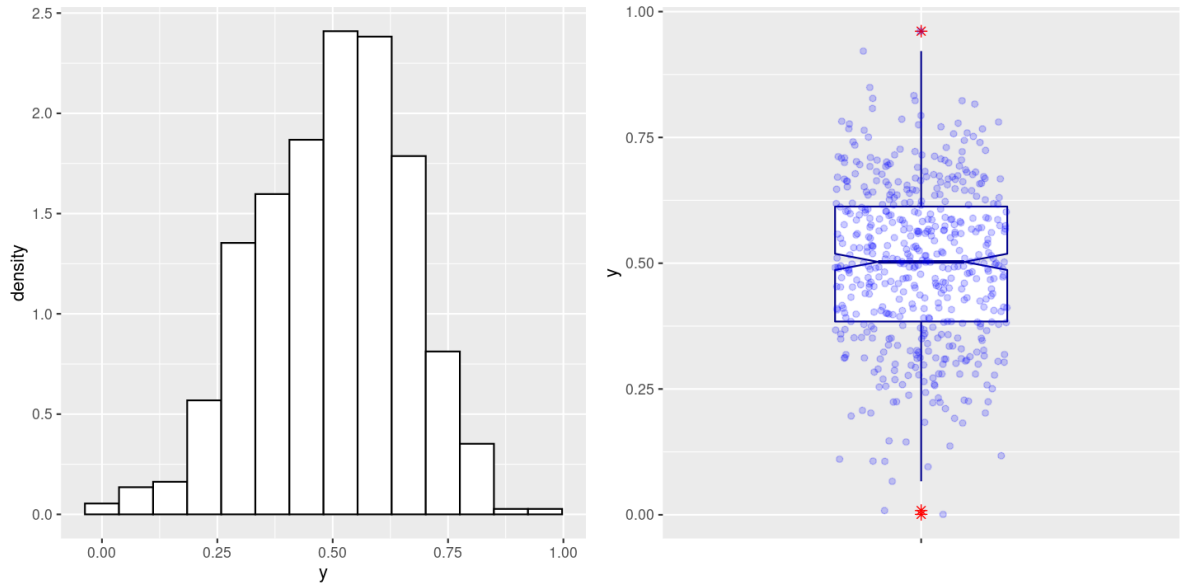
The presence of outliers may provoke big errors in the estimation procedures (Aeberhard et al. 2021; Hawkins 1980; Maronna et al. 2019; Rousseeuw and Hubert 2011). Since robustness is the ability to perform well when the data obey the assumed model and to not provide completely useless results when the observations do not exactly follow it. In these example studies, we consider contaminated models, where samples $\{X_1, \ldots, X_n\}$ are identically distributed random variables.

Example 1: Bounded data in (0, 1) (Mixture model with beta distribution)

The example use an mixture distribution to fit robust model by using GAMLSS with bounded data. The reference distribution is the Beta distribution $Beta(5, 5)$ with symmetric shape around the value 0.5. The pattern of the contamination two-side is defined as follow:

$$\alpha_1 \mathcal{U}(0, 0.1) + \alpha_2 \mathcal{U}(0.9, 1) + (1 - \alpha_1 - \alpha_2) Beta(5, 5).$$

Here, $n = 500$, $\alpha_1 = 0.01 = 1\%$ (contamination to lower-tail), $\alpha_2 = 0.01 = 1\%$ (contamination to upper-tail). On the contamination model we expect to generate potential extreme values that can affect the estimation of the Beta model parameters.



Histogram (left panel) and Boxplot (rigth panel) of simulated data

Outliers detection

Now, we extract the values of the potential outliers based on the IQR criterion. According to the IQR, there are 3 outliers

IQR criterion

index	value
131	0.0008640
258	0.0084364
416	0.9611100

Now we use the Hampel (Hampel 1974; Liu et al. 2004) criterion for the detection of outliers. According to the Hampel filter, there are 14 outliers.

Hampel criterion

index	value
25	0.1067264
119	0.9215307
131	0.0008640
170	0.0665794
177	0.1468574
248	0.1173745
258	0.0084364
310	0.1062264
368	0.8493144
377	0.1447119

index	value
383	0.1366603
409	0.1105287
416	0.9611100
486	0.0954537

By the symmetry of the data, we now use the outlier informations in setting the value k of potential outliers for the Rosner's test (Rosner 1983).

Rosner test - k based on IQR criterion

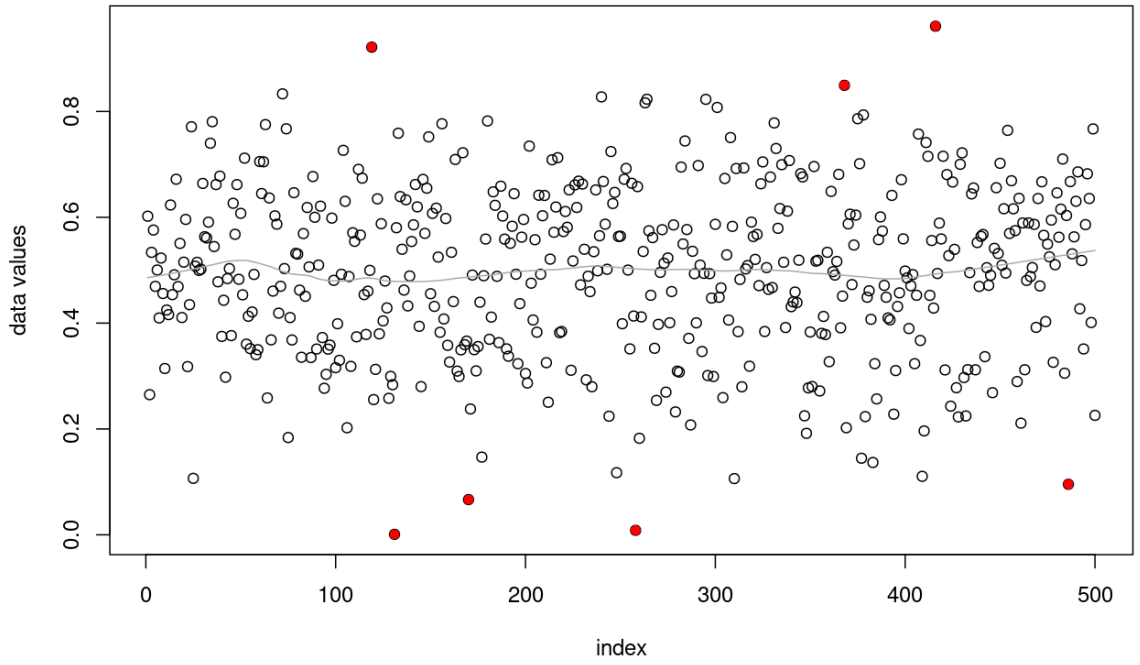
i	Mean.i	SD.i	Value	Obs.Num	R.i+1	lambda.i+1	Outlier
0	0.4982551	0.1589753	0.0008640	**131**	3.128732	3.863127	**FALSE**
1	0.4992519	0.1575631	0.0084364	**258**	3.115041	3.862597	**FALSE**
2	0.5002375	0.1561743	0.9611100	**416**	2.951015	3.862066	**FALSE**

Rosner test - k based on Hampel criterion

i	Mean.i	SD.i	Value	Obs.Num	R.i+1	lambda.i+1	Outlier
7	0.5009480	0.1510447	0.1067264	**25**	2.609967	3.859392	**FALSE**
4	0.5001826	0.1538828	0.9215307	**119**	2.738110	3.861000	**FALSE**
0	0.4982551	0.1589753	0.0008640	**131**	3.128732	3.863127	**FALSE**
3	0.4993102	0.1549531	0.0665794	**170**	2.792656	3.861534	**FALSE**
12	0.5048184	0.1468551	0.1468574	**177**	2.437511	3.856688	**FALSE**
9	0.5025461	0.1492542	0.1173745	**248**	2.580641	3.858314	**FALSE**
1	0.4992519	0.1575631	0.0084364	**258**	3.115041	3.862597	**FALSE**
6	0.5001490	0.1519329	0.1062264	**310**	2.592740	3.859929	**FALSE**
13	0.5055534	0.1461048	0.8493144	**368**	2.352839	3.856143	**FALSE**
11	0.5040820	0.1476056	0.1447119	**377**	2.434664	3.857231	**FALSE**
10	0.5033321	0.1483859	0.1366603	**383**	2.471070	3.857773	**FALSE**
8	0.5017493	0.1501459	0.1105287	**409**	2.605602	3.858853	**FALSE**
2	0.5002375	0.1561743	0.9611100	**416**	2.951015	3.862066	**FALSE**
5	0.4993314	0.1528651	0.0954537	**486**	2.642052	3.860465	**FALSE**

The Rosner's Tests show that any of observation are outlier based on the two criteria of detection. However, this result can be masked by normality violation of data.

Finally, we apply the semiparametric method based on kernel smoothing and extreme value theory (Čampulová et al. 2018; Holešovskỳ et al. 2018; Holešovskỳ and Fusek 2020) for outlier detection. The outliers are identified as observations whose values are exceeded on a certain average by using the function `RDetect.outliers.EV()` of the `envoutliers` package in `R` . According to the kernel approach, there are 9 outliers.

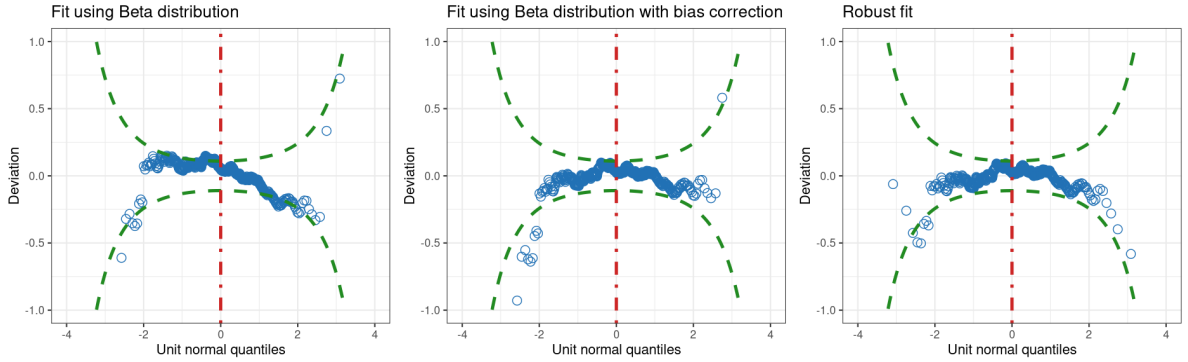


Kernel criterion

index	value
119	0.9215307
131	0.0008640
170	0.0665794
258	0.0084364
368	0.8493144
416	0.9611100
486	0.0954537

GAMLSS fit

We are looking for any lawful curves in the data that may suggest that the fitted model is better or worse at predictions. Worm-plots (Buuren and Fredriks 2001) can be used to identify some characteristics of the data that are not adequately captured by the fitted model via GAMLSS. We fit the data as a [rigby2019distributions] subsection 12.2.2.

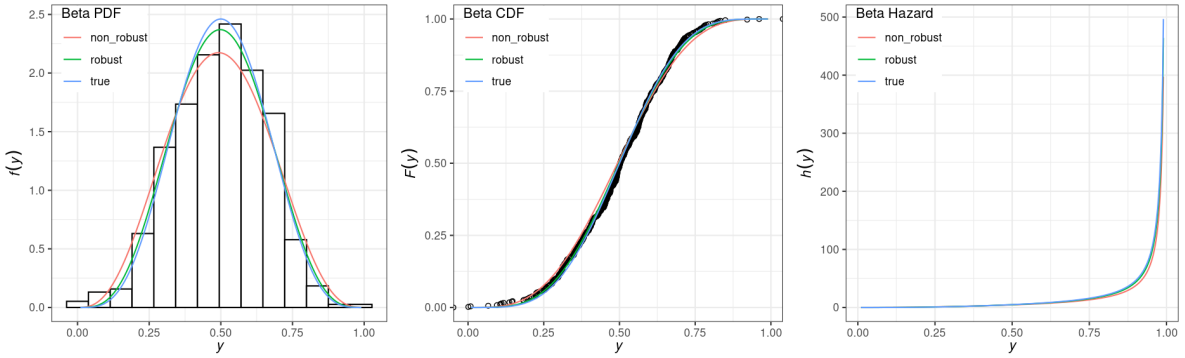


By visual inspection of the worm plots we observe that the robust fit produces the better fit of the data.

Fit results

Parameter	mu	sigma
True	5	5
Fit Beta	3.90406254832177	3.99792977763716
Fit beta with bias correction	4.64045354269377	4.67037735070456
Robust fitted	4.89278855626971	4.90656331419097

Comparison of the fits to the contaminated beta data



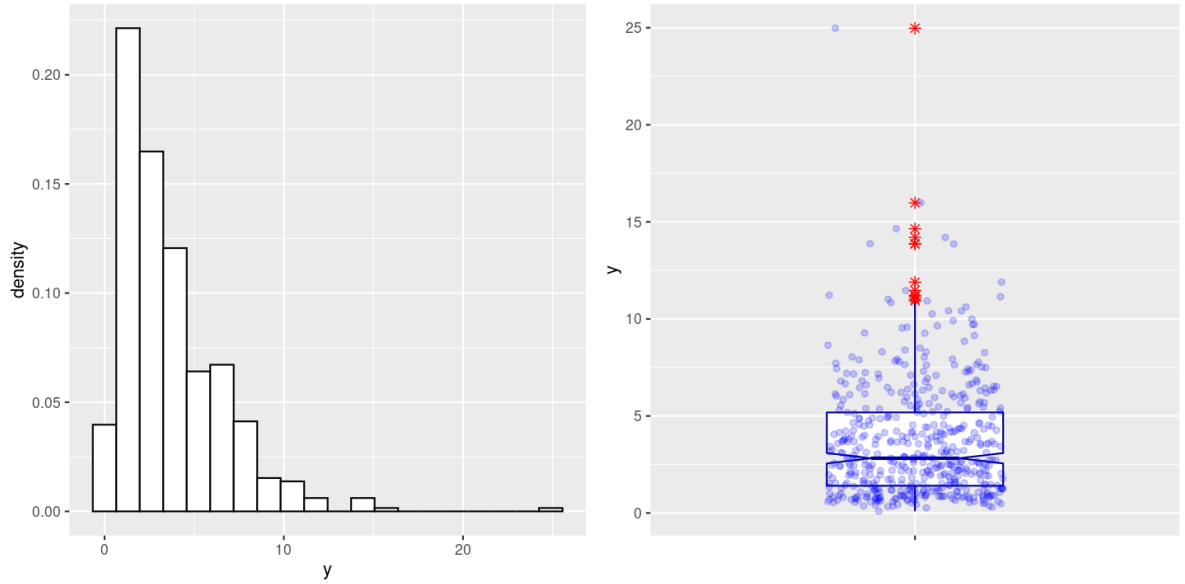
Example 2: Reaction time distributions

The example uses a mixture distribution to generate contaminated Reaction times. We fit known different Reaction time distribution and use robust estimation via GAMLSS. The reference distribution is the ExGaussian distribution $ExGaussian(\mu, \sigma, \nu)$. The parameters μ and σ are the mean and standard deviation from the normal distribution variable while the parameter ν is the rate control of the exponential component (the right-skew of the distribution).

The pattern of the contamination two-side is defined as follow:

$$(1-\alpha)ExGaussian(\mu, \sigma, \nu) + \alpha\mathcal{U}(L, U).$$

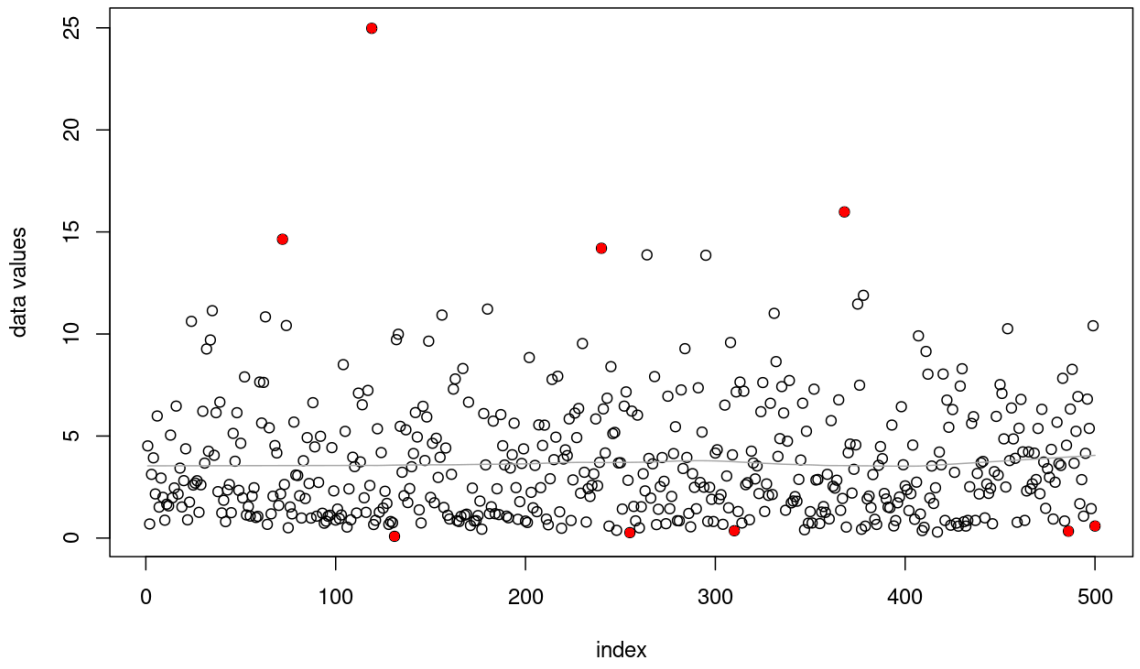
Here, α is the contamination fraction, L is the smallest observed reaction time, and U is the largest. On the contamination model we expect to generate potential extreme values that can affect the estimation of the ExGaussian parameters. In this example, $n = 500$, $\alpha = 0.1 = 10\%$ (contamination parameter), $\mu = 0.5$, $\sigma = 0.1$, $\nu = 3$, $L = 0$ and $U = 5$.



Histogram (left panel) and Boxplot (rigth panel) of simulated data

Outlier detection

Here, according to the kernel approach, there are 9 outliers.

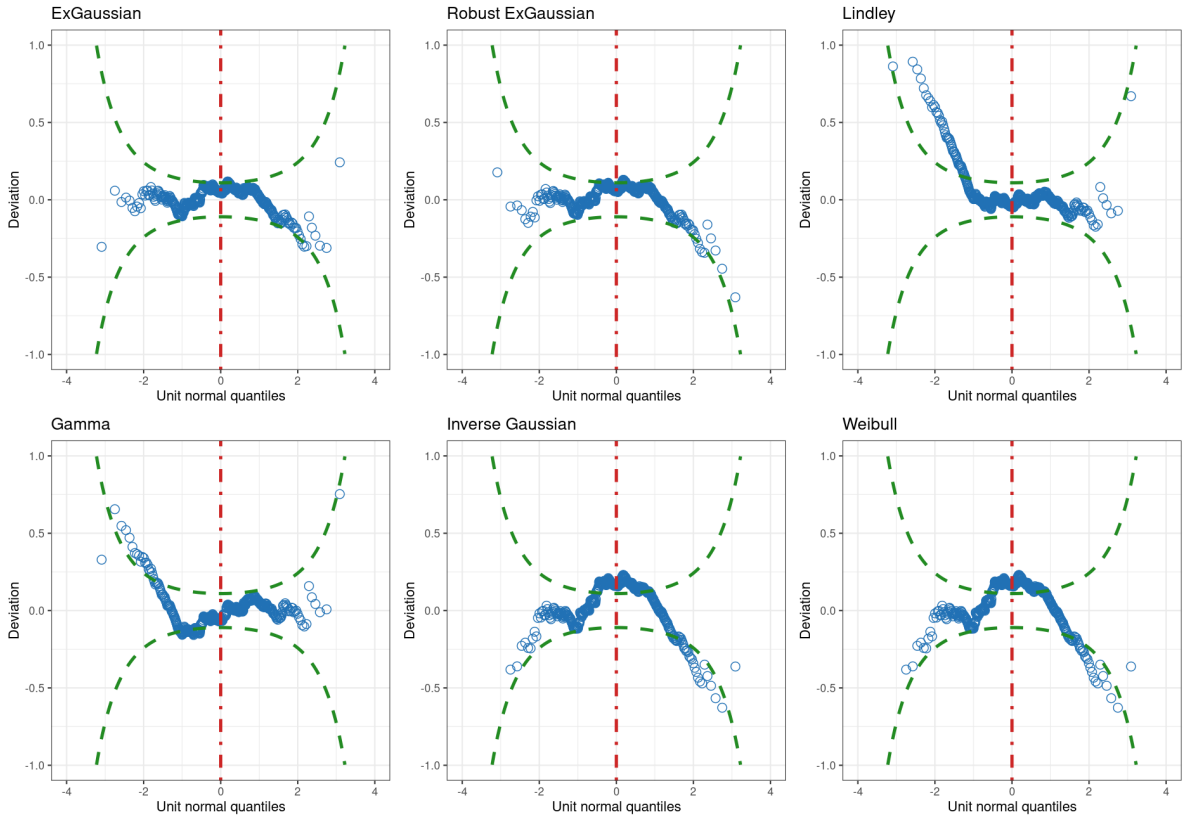


Kernel criterion

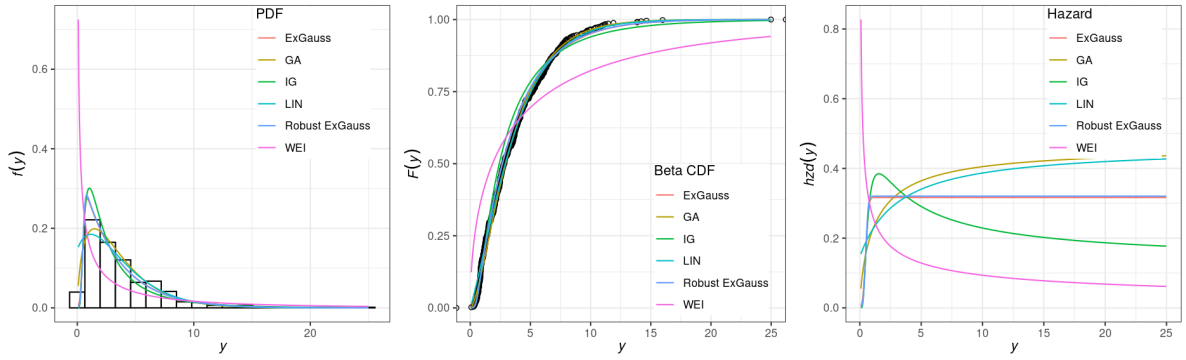
index	value
72	14.6415115
119	24.9775266
131	0.0863967
240	14.2002396

index	value
255	0.2672522
310	0.3589354
368	15.9806354
486	0.3372852
500	0.5898589

We fit the Exgaussian, Lindley, Gamma and Weibull models for data using GAMLSS. Also, ExGaussian is robust fitted by use `gamlssRobust`. The generalized joint regression modelling implemented in the `GJRM` package of `R` for robus estimation produce wrong results, we decided not to use it. Here, the Lindley distribution is in `RelDists` package. See here.



By visual inspection of the worm plots we observe that the robust fit produces the better fit of the data.



References

Aeberhard, W. H., Cantoni, E., Marra, G., and Radice, R. (2021), "Robust fitting for generalized additive models for location, scale and shape," *Statistics and Computing*, Springer, 31, 1–16.
Buuren, S. van, and Fredriks, M. (2001), "Worm plot: A simple diagnostic device for modelling growth reference curves," *Statistics in medicine*, Wiley Online Library, 20, 1259–1277.
Čampulová, M., Michálek, J., Mikuška, P., and Bokal, D. (2018), "Nonparametric algorithm for identification of outliers in environmental data," *Journal of Chemometrics*, Wiley Online Library, 32, e2997.
Hampel, F. R. (1974), "The influence curve and its role in robust estimation," *Journal of the american statistical association*, Taylor & Francis, 69, 383–393.
Hawkins, D. M. (1980), *Identification of outliers*, Springer.
Holešovskỳ, J., Čampulová, M., and Michalek, J. (2018), "Semiparametric outlier detection in nonstationary times series: Case study for atmospheric pollution in brno, czech republic," *Atmospheric Pollution Research*, Elsevier, 9, 27–36.
Holešovskỳ, J., and Fusek, M. (2020), "Estimation of the extremal index using censored distributions," *Extremes*, Springer, 23, 197–213.
Liu, H., Shah, S., and Jiang, W. (2004), "On-line outlier detection and data cleaning," *Computers & chemical engineering*, Elsevier, 28, 1635–1647.
Maronna, R. A., Martin, R. D., Yohai, V. J., and Salibián-Barrera, M. (2019), *Robust statistics: Theory and methods (with r)*, John Wiley & Sons.
Rosner, B. (1983), "Percentage points for a generalized ESD many-outlier procedure," *Technometrics*, Taylor & Francis, 25, 165–172.
Rousseeuw, P. J., and Hubert, M. (2011), "Robust statistics for outlier detection," *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, Wiley Online Library, 1, 73–79.