Exploratory plots in the analysis of extremes

Bikramjit Das\textsuperscript{1} and Souvik Ghosh\textsuperscript{2,3}
\textsuperscript{1}Singapore University of Technology and Design, Singapore
\textsuperscript{2}LinkedIn Corporation, Mountain View, California, USA
\textsuperscript{3}Corresponding author: Souvik Ghosh, e-mail: ghosh.souvik@gmail.com

Abstract

Exploratory data analysis is often used to test the goodness-of-fit of sample observations to specific target distributions. A few such graphical tools have been extensively used to detect heavy-tailed behavior or extramal behavior in observed data. In this paper we discuss recent theoretical advancements in the understanding of asymptotic behavior of two such plotting tools: the Quantile-Quantile plot and the Mean Excess plot. As an application we suggest better practices in the use of these plots in practice.

Key words: extreme values, heavy-tailed distributions, quantile-quantile plots, mean excess plots

1. Introduction

Graphical tools play an important role in statistical analysis. Various kinds of plots are used in most, if not all, data analysis. The Quantile-Quantile (QQ) plot and the Mean Excess (ME) plot are among the most popular plots used in the analysis of heavy-tailed data. The ME plot is more broadly used in studying extreme-valued data. In this paper we survey the recent developments that help us better understand the large sample properties of these plots. We use that knowledge to demonstrate how we can use these plots to conduct better analysis and make better decisions.

1.1. QQ plots for heavy tails

The QQ plot is popular graphical tool for detecting the goodness-of-fit for the observations in a dataset to some known distribution $F$. The QQ plot is a plot of the empirical quantiles from the data against the theoretical quantiles of $F$. If the true distribution of the sample is $F$ then the QQ plot should converge to a straight line. The QQ plot used in studying heavy-tailed data is a little different and specifically designed to check for distributions $F$ where $\bar{F} := 1 - F$ is regularly varying with some index $-1/\xi$, $\xi > 0$, also denoted $\bar{F} \in RV_{-1/\xi}$, cf. Resnick (2007), de Haan and Ferreira (2006) or Bingham, Goldie and Teugels (1987). For a sample $X_1, X_2, \ldots, X_n$, the QQ plot in the context of heavy-tailed distributions is defined by

$$Q_n = \left\{ \left( -\log \frac{j}{k}, \log \frac{X(j)}{X(k)} \right) : 1 \leq j \leq k \right\}, \quad k < n,$$

(1.1)
where \(X^{(1)} \geq X^{(2)} \geq \ldots \geq X^{(n)}\) denote the decreasing order statistics from the sample. We concentrate on the top \(k\) quantiles of the data and that is justified by the fact that \(\bar{F} \in RV_{-1/\xi}\) is only an asymptotic property of the right tail of the distribution \(F\). The value \(k\) in (1.1) is a function of the data size \(n\) and is chosen such that \(k\) is large \((k \to \infty)\) but small compared to \(n\) \((k/n \to 0)\). Throughout the paper we will assume that \(k\) satisfies this property.

### 1.2. ME plots

The generalized Pareto distribution (GPD) is an important class of distributions and is the centerpiece for the peaks-over-threshold method used in extreme value analysis (Davison and Smith, 1990). The GPD is characterized by its cumulative distribution function \(G_{\xi, \beta}\):

\[
G_{\xi, \beta}(x) = \begin{cases} 1 - (1 + \xi x/\beta)^{-1/\xi} & \text{if } \xi \neq 0, \\ 1 - \exp(-x/\beta) & \text{if } \xi = 0, \end{cases}
\]

(1.2)

The ME plot is often used as a graphical test to check if data conform to a GPD. The ME plot is the plot of the points \(\{(X^{(k)}, \hat{M}(X^{(k)})) : 1 < k \leq n\}\) where,

\[
\hat{M}(u) := \frac{\sum_{i=1}^{n} (X_i - u)I[X_i > u]}{\sum_{i=1}^{n} I[X_i > u]}, \quad u \geq 0.
\]

(1.3)

is the empirical estimator of the ME function

\[
M(u) := E[X - u | X > u],
\]

(1.4)

provided \(EX_+ < \infty\). For a random variable \(X \sim G_{\xi, \beta}\), we have \(E(X) < \infty\) if and only if \(\xi < 1\) and in this case, the ME function of \(X\) is linear in \(u\):

\[
M(u) = \frac{\beta}{1 - \xi} + \frac{\xi}{1 - \xi} u,
\]

(1.5)

where \(0 \leq u < \infty\) if \(0 \leq \xi < 1\) and \(0 \leq u \leq -\beta/\xi\) if \(\xi < 0\). The linearity of the ME function characterizes the GPD class, and it suggests that if the ME plot for a data set is close to a straight line for high values of the threshold then there is no evidence against the use of a GPD model. See also McNeil, Frey and Embrechts (2005), Embrechts, Klüppelberg and Mikosch (1997) and Hogg and Klugman (1984) for the traditional implementation of this plot.

### 2. QQ plots and ME plots: theory and practice

Historically, we have depended a lot on heuristics while interpreting the results of exploratory plots. While using the traditional QQ plot, in order to test the null hypothesis that a data set is an IID sample from a distribution \(F\), we check if the QQ plot is close to a diagonal straight line. This is an artifact of the Glivenko-Cantelli theorem, cf. Shorack and Wellner (1986). Still there is a gap that remains to be bridged given that a plot is a random closed set in \(\mathbb{R}^2\) and the Glivenko-Cantelli theorem describes convergence of the empirical CDF to the theoretical CDF in the space of functions on \(\mathbb{R}\). The heuristic arguments have also been applied in interpreting the results of the QQ plots and ME plots for the analysis of extremes.
2.1. Convergence in probability

Das and Resnick (2008) study the convergence of the QQ plot (1.1) in an appropriate topology of random sets. Molchanov (2005) is a good reference on the topic of convergence of random sets. Das and Resnick (2008) show that under the null hypothesis of $\bar{F} \in RV_{-1/\xi}$ for some $\xi > 0$, the QQ plot converges in probability to a straight line with slope $1/\xi$.

Ghosh and Resnick (2010) discusses convergence in probability of the ME plots. They show that if $F$ is a distribution in the domain of attraction of a GPD $G_{\xi, \beta}$ with $\xi < 1$, i.e. belongs to the maximal domain of attraction of an extreme value distribution, then the suitably normalized ME plot converges in probability to a straight line with slope $\xi/(1-\xi)$. The slope of the limiting line is positive, negative or zero, depending on whether $\xi$ is positive, negative or zero, which is equivalent to $F$ being in domain of attraction of the Fréchet, the Weibull or the Gumbel distributions respectively. Ghosh and Resnick (2011) proved the converse that if the ME plot converges to a straight line for some normalization then $F$ must be in the domain of attraction of a GPD.

The advantage of the ME plot over the QQ plot is that it works when $-\infty < \xi < 1$ whereas the QQ plot works for $\xi > 0$ only. Hence the ME plot can be used whenever the sample is in the maximal domain of attraction of any generalized extreme value distribution with finite mean (Fréchet, Weibull or Gumbel distribution). The QQ plot is restricted to the domain of attraction of Fréchet distribution only. The disadvantage of the ME plot is that it requires $\xi < 1$ to make proper sense of the result, i.e., the underlying distribution should have a finite mean. Still, limits can and have been obtained for the ME plots even when the distributional mean is not finite, see Ghosh and Resnick (2010).

Fig 1. QQ Plot for right-skewed stable random variables with $\xi = 2/3$.

2.2. Weak convergence

One data set leads to just one single QQ or ME plot and a single plot is often not enough to statistically detect proximity between the plot and the intended fixed
limit set. Creating appropriate confidence bounds around these plots though can help us to test the null hypothesis with some degree of confidence.

Das and Ghosh (2013a) proved the weak convergence of the QQ and the ME plots for samples from heavy-tailed distributions in the space of random sets of $\mathbb{R}^2$. This study is very helpful as the results can be used to construct confidence bounds around the plots and test the null hypothesis that the sample conforms to the class of GPD. In Das and Ghosh (2013b), the authors extend the results to random variables in the maximal domain of attraction of the Gumbel and the Weibull distributions.

It is important to note that using the weak limit to obtain the confidence bounds is a more stable method compared to other statistical tools like bootstrap. The difficulty of using bootstrap for data from heavy-tailed distribution is well documented in Resnick (2007).
3. Examples and conclusion

Figure 1 shows the QQ plot for a sample of simulated IID stable(1.5) distribution ($\xi = 2/3$). Figure 2 shows the ME plot for the same data set. In this example the distribution is heavy-tailed and belongs to the maximal domain of attraction of the Fréchet distribution. Figure 3 shows the ME plot for a Beta(2,2) distribution ($\xi = -0.5$) which is in the maximal domain of attraction of the Weibull distribution. In both these cases the confidence bounds of the plots allow us to accept the null hypothesis.

We observe a data set, freely available from [www.eea.europa.eu](http://www.eea.europa.eu). The data set contains daily maxima of ozone concentration (in $\mu g/m^3$) from one station in Zurich, Switzerland (station code CH 0010A, Zurich-Kaserne) located 410 mts above sea-level. Data is observed from January 1, 1992 to December 31, 2009. Measurements were unavailable for 22 days, which we impute by the average value of ozone concentration on the same day for other available years. As seen in the plots in Figure
the data clearly admits periodicity. We model the seasonality and the stationary component using an AR process. We then analyze the extremal behavior of the residuals of the model. The confidence bounds indicates that we should not reject the null hypothesis that the residuals belong to the maximal domain of attraction of the Gumbel distribution.

The figures clearly demonstrate the usefulness of having the confidence bounds along with the plots. In many situations (Figure 2 for example) we might wrongly reject the null hypothesis in absence of the confidence bounds. The bounds show us which portion of the plot we should concentrate on when we make the decision. When on one hand we should concentrate on the tail portion of the data, i.e., take $k$ to be small, on the other hand we must take $k$ to be large enough so as to see any large sample behavior in the plot.

References


