Robust Estimation of the Parameters of the GPD A Case Study

Prepared by René Stephan

Presented to the ASTIN COLLOQUIUM
LISBOA 2016
31 MAY - 03 JUNE 2016

René Stephan, Klemmstein & Stephan GmbH

Abstract

The paper presents the results of a case study fitting the generalized Pareto distribution to insurance industry claims data. Besides classical parametric procedures, robust statistical concepts are considered. The latter provide instruments to assess the characteristics of estimators also in the neighborhood of parametric models.

A demand for robust methods may arise in cases of fitting distribution functions to large claims or extreme events, that is, in situations, in which quite a few data points may have a considerable impact on the estimate. Special areas of application are the calibration of individual large claims in internal models and reinsurance pricing.

Keywords: Generalized Pareto distribution, M-estimator, influence curve, contamination, neighborhood, gross error sensitivity.

0.1 Introduction

Topic of this article is the fitting of the generalized Pareto distribution (GPD) to insurance industry claims data. The interest is in estimating the tail of large claims or extreme events.

Classical statistical procedures act on strict parametric model assumptions which may not always be fulfilled in real world situations. Tukey (1960) has shown that already under tiny deviations from model assumptions, here from normality, the mean deviation may perform better than the standard deviation, the latter being efficient under strict normality. This conjecture goes back to A.S. Eddington, see Fisher (1920).

Robust statistical concepts deal with the stability of statistical procedures in the neighborhood of ideal models. The results of a case study are presented in which classical and robust estimators are applied to fit the GPD to insurance industry data. We investigate Danish fire insurance claims data, that have been analysed extensively in the literature, see for example McNeil (1997). New in this case study is the consideration of several estimators with quite different characteristics in terms of efficiency and robustness. To test the performance of the estimators the dataset is contaminated by adding a new largest claim. The following estimators are considered in some detail: maximum likelihood estimator, method of moments, optimal bias-robust estimator and Cramér-von Mises minimum distance estimator. The efficiency and (local) robustness of these estimators are briefly discussed. The probability weighted moments estimator and the Kolmogorov minimum distance estimator are included in the case study. The final selection of an estimator may be a decision on the trade-off between efficiency and robustness.

0.2 The Estimation Problem

A parametric model $\mathcal{P}$ consists of a family of probability measures or distribution functions F_θ on a sample space $(\mathcal{X}, \mathcal{A})$ with θ belonging to a parameter space Θ. Here, F_θ is the generalized Pareto distibution (GPD) given by

$$F_\theta(x) \ = \ 1 - \left(1 + \xi \frac{x}{\beta}\right)^{-1/\xi}$$

on the sample space $\mathcal{X} = [0, \infty)$ with unknown parameter $\theta = (\beta, \xi) \in \Theta = (0, \infty) \times (0, \infty)$. The parameters β and ξ are referred to as scale and shape parameter. The family $\mathcal{P}$ is L_2-differentiable (smooth) at $\theta \in \Theta$ with L_2-derivative or scores function

$$\Lambda_\theta \left(x\right) = \frac{d}{d\theta} \ln f_\theta \left(x\right)$$

with $E_\theta \Lambda_\theta = 0$ and Fisher information of full rank $\mathcal{I}_\theta = E_\theta \Lambda_\theta \Lambda_\theta^t$.

Assume that $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) observations belonging to the sample space $\mathcal{X} = [0, \infty)$. The X_i may be interpreted as exceedances over some threshold. Classical statistics assume that the observations X_i are distributed exactly like one of the F_θ and estimate θ based on the data that is available.

Robust statistics[1] introduce the concept of neighborhoods of distribution functions or probability measures. P_θ will generally differ from the real probability measure which in turn may be at least in the neighborhood of P_θ. Such a neighborhood can be described by contamination balls $U_c\left(\theta, r\right) = B_c\left(P_\theta, r\right)$ that allow for convex combinations between P_θ and arbitrary probability measures,

$$B_c\left(P_\theta, r\right) = \left\{\left(1-r\right)^+ P_\theta + \min\left(1, r\right) Q \mid Q \in \mathcal{M}_1\left(\mathcal{A}\right)\right\}.$$

Influence curves, or then called influence functions, have been introduced as Gateâux derivatives of statistical functionals, see for example Hampel et al (1986). They also appear as summands of asymptotically linear estiamtors, see Rieder (1994). The set $\Psi_2\left(\theta\right)$ of all square integrable influence curves is defined as

$$\Psi_2\left(\theta\right) = \left\{\psi_\theta \in L_2^k\left(P_\theta\right) \mid E_\theta \psi_\theta = 0,\ E_\theta \psi_\theta \Lambda_\theta^t = \mathbb{I}_k\right\}.$$

(In our case $k = 2$ for the two parameters β and ξ.)

[1]For the theoretical background we refer to Rieder (1996). Here, just some rough ideas of the theory are indicated.

An estimator $\hat{\theta}_n = \hat{\theta}_n(X_1 \ldots, X_n)$ is called asymptotically linear at P_θ if there is an influence curve $\psi_\theta \in \Psi_2(\theta)$ such that

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_\theta(X_i) + o_{P_\theta^n}\left(n^0\right).$$

(o is the usual Landau symbol, $X_n = o_P(R_n)$ is short for $X_n/R_n \to 0$ in probability.) This expansion determins the influence curve ψ_θ uniquely, see Rieder (1996). If the estimator $\hat{\theta}_n$ is asymptotically linear with square integrable influence curve ψ_θ, then

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \to \mathcal{N}\left(0, Cov_\theta(\psi_\theta)\right)$$

in distribution.

All estimators that are considered here, M estimators and minimum distance estimators, are asymptotically linear.

0.3 M Estimators

M estimators are of special interest in the context of robust estimation. They generalize the concept of maximum likelihood estimators.

Consider the classical estimation problem. The maximum likelihood (ML) etimator $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ maximizes the likelihood or equivalently the log likelihood

$$l_n(\theta; X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} \ln f(X_i, \theta).$$

M estimators are a generalization of this concept introduced by Huber (1964). The idea is to replace the function $\ln f$ by one that does not deviate too much from $\ln f$ to maintain the good characteristics of the ML estimator and that reacts more robust to outliers. M estimators are defined as maximizing an expression like

$$M_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} m(X_i, \theta)$$

with some function $m : \mathcal{X} \times \mathbb{R}^k \to \mathbb{R}^l$. Suppose m has a set of partial derivatives $\Psi(x, \theta) = \frac{\partial}{\partial \theta} m(x, \theta)$, then $\hat{\theta}_n$ is called an M-estimator if it satisfies the equation

$$\Psi_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \Psi_\theta(X_i, \theta) = 0.$$

Assume that the M estimator $\hat{\theta}_n$ is consistent, that is, the sequence $\hat{\theta}_n$ converges in probability to a zero θ_0 of $\Psi = E_\theta \Psi_\theta$. Under mild regularity conditons $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$ is asymptotically normal with mean zero and covariance matrix $\left(E_{\theta_0} \frac{d}{d\theta} \Psi_{\theta_0}\right)^{-1} E \Psi_{\theta_0} \Psi_{\theta_0}^t \left(E_{\theta_0} \frac{d}{d\theta} \Psi_{\theta_0}\right)^{-t}$, see van der Vaart (1998). Then,

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{\theta_0}(X_i) + o_{P_{\theta_0}^n}\left(n^0\right)$$

with $\psi_{\theta_0}(x) = -\left(E_{\theta_0} \frac{d}{d\theta} \Psi_{\theta_0}\right)^{-1} \sqrt{n} \Psi_{\theta_0}(x)$.

0.4 Criteria to assess an Estimator

The asymptotic covariance matrix of an asymptotically linear estimator $\hat{\theta}_n$ (converging to θ) may be determined as $Cov_\theta(\psi_\theta) = E_\theta \psi_\theta \psi_\theta^t$. To measure the efficiency of an estimator and compare it to other estimators we either consider the trace of the covariance matrix, see Hampel et al. (1986),

$$\text{trace } Cov_\theta(\psi_\theta) = E_\theta \psi_\theta^t \psi_\theta = E_\theta \parallel \psi_\theta \parallel^2$$

or the standard error of the components of $\hat{\theta}_n$.

The gross error sensitivity (GES) is a measure for the asymptotic bias of an estimator caused by contamination, see Hampel et al. (1986),

$$\text{GES}\hat{\theta}_n = \sup_x \parallel \psi_\theta(x) \parallel .$$

The GES provides a measure of the (local) robustness of an estimator. If $\text{GES}\hat{\theta}_n < \infty$ the estimator $\hat{\theta}_n$ is called bias robust (b-robust).

0.5 Estimators

The following estimators are considered in the case study analysing the industry data,

- Maximum likelihood estimator (MLE)
- Method of moments estimator (MOM)
- Optimal bias robust estimator (OBRE)
- Cramér-von-Mises minimum distance estimator (MDE CvM)

Some basic ideas of these estimators are provided, estimation procedures and characteristics are indicated.

0.5.1 Maximum Likelihood Estimator

The maximum likelihood estimator $\hat{\theta}_n^{MLE} = \hat{\theta}_n^{MLE}(X_1 \ldots, X_n)$ maximizes the likelihood $L(\theta; X_1, \ldots, X_n) = \prod_{i=1}^n f_\theta(X_i)$ or equivalently the log likelihood

$$l_n(\theta; X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^n \ln f_\theta(X_i).$$

The estimator satiefies the implicit equation,

$$\Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^n \Lambda_\theta(X_i) = 0,$$

with scores function $\Lambda_\theta(x) = d/d\theta \ln f_\theta(x)$. A zero of Ψ_n can be determined iteratively applying the Newton-Raphson algorithm,

$$\theta_n^{(i+1)} = \theta_n^{(i)} + \mathcal{I}_{\theta_n^{(i)}}^{-1}\frac{1}{n}\sum_{i=1}^n \Lambda_{\theta_n^{(i)}}(X_i),$$

with Fisher information $\mathcal{I}_\theta = -E_\theta \frac{d}{d\theta}\Lambda_\theta = E_\theta \Lambda_\theta \Lambda_\theta^t > 0$ and some starting value $\theta_n^{(0)}$. The influence function is given by

$$\psi_\theta(x) = \mathcal{I}_\theta^{-1}\Lambda_\theta(x).$$

The MLE is asymptotically normal with asymptotic covariance matrix $\mathcal{I}_\theta^{-1}/n$. The estimator is efficient regarding the partial ordering of covariance matrices,

$$U \leq V \Leftrightarrow V - U \text{ positive semidefinite.}$$

The Influence function is unbounded and the asymptotic bias is infinite. The MLE is not b-robust.

0.5.2 Method of Moments

The method of moments or moment estimator $\hat{\theta}_n^{MOM}$ for θ is the solution of the system of equations

$$T_n = \left(\overline{X}, \overline{X^2}\right) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{i=1}^{n} X_i^2\right) = e(\theta) := \left(E_\theta X, E_\theta X^2\right).$$

With $e_i = E_\theta X^i$, $i = 1, 2$, the moment estimator can be specified as

$$\hat{\theta}_n^{MOM} = \left(\hat{\beta}, \hat{\xi}\right) = e^{-1}\left(E_\theta X, E_\theta X^2\right)$$

provided that $\xi < 0.5$. The influence function of $T_n = \left(\overline{X}, \overline{X^2}\right)$ is given by

$$\psi_\theta^{T_n}(x) = \begin{pmatrix} x - EX \\ x^2 - EX^2 \end{pmatrix},$$

see for example Hampel et al. (1986). The influence function of $\hat{\theta}_n^{MOM}$ can be determined by applying the delta method, see van der Vaart (1998), as $\psi_\theta = \mathcal{D}\psi_\theta^{T_n}$ with $\mathcal{D}$ as Jacobi matrix of e^{-1}. The moment estimator is asymptotically normal for $\xi < 0.25$, i.e. if the fourth moment exists. The asymptotic covariance matrix is given by $\mathcal{D}\Sigma\mathcal{D}^t/n$ with $\Sigma = Cov_\theta\left(X, X^2\right)$, the covariance matrix of the first two moments. The influence function is unbounded. The estimator is not b-robust.

0.5.3 Optimal B-Robust Estimator

The intention associated with the optimal bias-robust estimator (OBRE) $\hat{\theta}_n^{OBRE}$ is to construct an estimator that is just slightly less efficient than the

MLE and that has at the same time a bounded asymptotic bias in the neighborhood of the ideal model. There is a trade-off between these objectives. The OBRE minimizes the trace of the asymptotic covariance matrix subject to a bias bound b,

$$E_\theta \parallel \psi_\theta \parallel^2 = \min! \quad \text{subject to} \quad \sup_x \parallel \psi_\theta \parallel \leq b.$$

The solution to this otimization problem coincides with the solution to the asymptotic mean square error (MSE) problem,

$$\text{maxMSE}_\theta\left(\psi_\theta, r\right) := E_\theta \parallel \psi_\theta \parallel^2 + r^2 \sup_x \parallel \psi_\theta \parallel^2 = \min!$$

To estimate the parameter θ, determine a starting value $\theta_n^{(0)}$ for the iteration and set $a = 0$ and $A = \mathcal{I}^{-1}_{\theta_n^{(0)}}$ with Fisher information $\mathcal{I}_\theta$. $\hat{\theta}_n^{OBRE}$ can be determined iteratively,

$$\theta_n^{(i+1)} = \theta_n^{(i)} + A \frac{1}{n} \sum_{i=1}^{n} w_c^{A,a}\left(X_i\right)\left(\Lambda_{\theta_n^{(i)}}\left(X_i\right) - a\right),$$

with

$$\begin{aligned}
w_b^{A,a}\left(x\right) &= h_b\left(A\left(\Lambda_\theta\left(x\right) - a\right)\right), \quad h_b\left(x\right) = \min\left(1, \frac{b}{\parallel x \parallel}\right), \\
a &= \frac{\int w_b^{A,a}\left(s\right)\Lambda_\theta\left(s\right) dF_\theta\left(s\right)}{\int w_b^{A,a}\left(s\right) dF_\theta\left(s\right)}, \\
A &= \left(\int w_b^{A,a}\left(s\right)\left(\Lambda_\theta\left(s\right) - a\right)\left(\Lambda_\theta\left(s\right) - a\right)^t dF_\theta\left(s\right)\right)^{-1}.
\end{aligned}$$

$w_b^{A,a}$ is a weighting function. The bulk of the observations will usually be assigned the weight 1, while very large claims or outliers may be assigned a weight between 0 and 1. a has an auxiliary function. It ensures consistency of the estimator. Matrix A can be interpreted as inverse of the generalization of the Fisher information. b and radius r are related by $r^2 b = E_\theta\left(\parallel A\Lambda_\theta - a \parallel - b\right)_+$. The influence function is given by

$$\psi_\theta\left(x\right) = w_c^{A,a}\left(x\right) A\left(\Lambda_\theta\left(x\right) - a\right).$$

It is bounded. The OBRE is relative efficient. The smaller the constant b is, the less efficient the estimator is. The asymptotic bias does not exceed b. The OBRE is b-robust.

0.5.4 Cramér-von Mises Minimum Distance Estimator

The Cramér-von-Mises (CvM) minimum distance between empirical df $\hat{F}_n$ and theoretical df F_θ is defined as

$$d_{CvM}\left(\hat{F}_n, F_\theta\right) := \sqrt{\int \left(\hat{F}_n(x) - F_\theta(x)\right)^2 dF_\theta(x)}.$$

The minimun distance estimator $\hat{\theta}_n^{MDE}$ minimizes the distance d_{CvM}. For the parameter estimation we refer to Rieder (1994). The family $\mathcal{P}$ of generalized Pareto distributions is Cramér-von Mises (CvM) differentiable at θ with CvM derivative Δ_θ,

$$\Delta_\theta(y) = \int \left(I(x \leq y) - \Lambda_\theta(x)\right) P_\theta(dx).$$

The influence function is given by

$$\psi_\theta(x) = \int \left(I(x \leq y) - F_\theta(y)\right) \varphi_\theta(y)\, \mu(dy),$$

with $\varphi_\theta(x) = \mathcal{J}_\theta^{-1}\Delta_\theta(x)$ and CvM information $\mathcal{J}_\theta$ of $\mathcal{P}$ at θ, $\mathcal{J}_\theta = \int \Delta_\theta \Delta_\theta^t d\mu > 0$. The estimator is quite efficient and b-robust, the influence curve is bounded.

0.6 Case Study: Danish Fire Insurance Data

The data analysed in this case study comprises of 2,167 fire claims of a Copenhagen reinsurance company collected in 1980 to 1990. The data is adjusted for inflation reflecting 1985 values. (The claims represent aggregated claims from the sub-lines building, contents and loss of profits.)

First, we study the data considering a threshold of 10m DKK (Model 1). Then, we introduce an additional large claim of 350m DKK and investigate how the estimators react to this contamination. In the original data set the largest claim amounts to 263m DKK. The threshold of 10m DKK and the contamination have been considered by McNeil (1997).

Model 1: Exceedances over threshold of 10m DKK

Model 2: Introduction of new largest claim of 350 m DKK to dataset

0.6.1 Model 1: Exceedances over threshold of 10m DKK

Estimator	Scale parameter		Shape parameter		GES	Quantiles	
	β	s.e. β	ξ	s.e. ξ		99.5%	99.9%
MLE	6.975	1.156	0.497	0.143	∞	181	421
MOM	8.520		0.395		∞	153	309
OBRE ($b = 10$)	6.970	1.161	0.494	0.149	10	179	413
OBRE ($b = 9$)	6.982	1.164	0.488	0.150	9	176	403
OBRE ($b = 8$)	7,032	1.162	0.454	0.150	8	156	341
MDE CvM	7.696	1.382	0.333	0.202	28	112	208

The MLE is efficient. The standard errors of scale and shape parameter are smaller than the standard errors of the other estimators. The GES is infinite, the MLE is not b-robust. The MOM provides a poor fit. Since the shape parameter exceeds 1/4, the estimator is not asymptotically normal. The MOM is not b-robust. OBRE and MDE CvM are both efficient and b-robust ($GES < \infty$). With decreasing constant b the shape parameter of the OBRE decreases as well. The OBRE achieves that by assigning a weight lower than 1 to the largest claims. The radius r depends on the constant b and equals 5.1% ($b = 10$), 8.1% ($b = 8$) and 14.5% ($b = 6$).

0.6.2 Model 2: Introduction of new claim of 350m DKK

For this analysis we have considered two more estimators, the Kolmogorov minimum distance estimator (MDE Kol) and the probability weighted moments estimator (PWM). The former minimizes the Kolmogorov distance defined by

$$d_{Kol}\left(\hat{F}_n, F_\theta\right) := \sup_x \mid \hat{F}_n(x) - F_\theta(x) \mid,$$

for the latter we refer to Hosking and Wallis (1987).

Estimator	Model 1			Model 2			Delta
	ξ	Quantiles		ξ	Quantiles		99.9%-
		99.5%	99.9%		99.5%	99.9%	Quantile
MLE	0.497	181	421	0.597	257	690	64%
PWM	0.517	191	455	0.613	266	732	61%
OBRE ($b = 10$)	0.494	179	413	0.592	252	672	63%
OBRE ($b = 9$)	0.488	176	403	0.571	234	603	50%
OBRE ($b = 8$)	0.454	156	341	0.551	218	545	60%
MDE CvM	0.333	112	208	0.370	127	248	19%
MDE Kol	0.457	158	347	0.489	179	410	18%

We leave the moment estimator (MOM) aside because of the poor fit and the restriction to the parameter space associated with the MOM ($\xi < 1/4$). The 99.9%-quantile of MLE and OBRE increases by approximately 60% switching from model 1 to model 2. An exeption is the OBRE with constant $b = 9$. For this estimator the increase is still 50%. The behaviour of the minimum distance estimators stands out. The increase is slightly less than 20%.

0.7 Conclusions

In this paper several estiamtors, classical and robust ones, have been applied to insurance industry data. It turned out that the selection of the estimator may have a significant impact on the shape parameter and the quantiles of the fitted distribution. Contaminating the dataset with a new largest claim results in a considerable increase of the shape parameter not just for classical estimators but also for robust ones like the OBRE. The behaviour of the minimum distance estimators turned out to be more resilient in this respect. It should be avoided to use robust estimators to optimize the fit to the bulk of the data at the cost of weighting down or excluding the largest claims (outliers) without precaution for an appropriate tail of the fitted distribution. After all, robust estimators should be treated with knowledge and care.

0.8 Literatur

Hampel, F.R., et al. (1986). Robust Statistics - The Approach Based on Influence Functions. Wiley, New York.

Hosking, J.R.M. and J.R. Wallis (1987). Parameter and quantile estimation for the generalized Pareto distribution. Technometrics 29 339-349 Huber, P.J. (1964). Robust estimation of a location parameter. Ann. Math. Stat. 35 73-101.

Huber, P.J. (1981). Robust Statistics. Wiley, New York.

Fisher, R.A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error. Monthly Not. Roy. Astr. Soc. 80 758-770.

McNeil, A.J. (1997). Estimating the Tails of Loss Severity Distributions using Extreme Value Theory. ASTIN Bulletin 27 117-137.

Rieder, H. (1994). Robust Asymptotic Statistics. Springer-Verlag New York.

Tukey, J.W. (1960). A survey of sampling from contaminated distributions. In Contributions to Probability and Statistics, I. Olkin (ed.). Stanford University Press, Calif., 448-485.

van der Vaart, A.W. (1998). Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge Univ. Press, Cambridge.