Thesis for the Degree of Doctor of Philosophy

# Moment estimation using extreme value techniques

N. C. Joachim Johansson

CHALMERS | GÖTEBORG UNIVERSITY


Department of Mathematical Statistics
Chalmers University of Technology and Göteborg University
Göteborg, Sweden, 2003

Moment estimation using extreme value techniques
N. C. Joachim Johansson
ISBN 91-7291-305-3



Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 1987
ISSN 0346-718X

Department of Mathematical Statistics
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31 772 1000

Printed and bound at the School of Mathematical Sciences
Chalmers University of Technology and Göteborg University
Göteborg, 2003

Abstract

The thesis is composed of three papers, all dealing with the application of extreme value methods to the problem of moment estimation for heavy-tailed distributions.

In Paper A, an asymptotically normally distributed estimate for the expected value of a positive random variable with infinite variance is introduced. Its behavior relative to estimation using the sample mean is investigated by simulations. An example of how to apply the estimate to file-size measurements on Internet traffic is also shown.

Paper B extends the results of Paper A to a situation where the variables are $m$-dependent. It is shown how this method can be applied for estimating covariances and be put to use as a diagnostic tool for estimating the order of an ARMA-processes with heavy-tailed innovations.

Paper C further extends the methodology to the case of regression through the origin with heteroscedastic errors. In a simulation study, the estimate is compared to some standard alternatives and used for estimating the population total in a superpopulation sampling framework.

Keywords: Pareto distribution, mean estimation, heavy tailed distributions, telecommunication, $m$-dependence, superpopulation sampling, peaks over threshold

MSC2000 classification: primary – 62G32
secondary – 62F10, 62F12, 62P30, 62F35, 62D05

## Acknowledgements

As a kid, I was real good at adding numbers. Even quite big numbers – in some cases almost extreme. As the numbers grew ever larger, my teachers patience in checking the answers I got grew proportionately smaller. I did not understand this at the time. Surely it was of great interest to see what five thirteen figure numbers added up to? This, I thought, is what math researchers do.

Anyways, I enjoyed solving math problems and, since I wasn't very good at football, I didn't share my friends dreams of becoming a professional athlete. Instead, I decided I would become a mathematics professor when I grew up. This might not be very common amongst seven year old boys but, as I said, I did know my math, I didn't know my football.

Now, some twenty-odd years later, I have a somewhat broader view on what mathematics is, and I still enjoy the area. My main teacher nowadays, my supervisor professor Holger Rootzén, even looks at those long calculations of mine. Sometimes he does this whilst shaking his head and sighing – just like my first grade teacher did – but not so often recently. Throughout the years, he's always been there with encouraging comments and clues for how to proceed when I got stuck somewhere. The thesis you hold in your hands wouldn't be here if it wasn't for him and the way he pointed out a path for me to follow. For this I am very grateful.

Over the years I have learned a lot from many at the department and outside of it. Among those who stand out in this respect are Sture Holm who got me starting out as a PhD student; Olle Nerman, who has the most amazing mathematical intuition and Patrik Albin, a man whose energy and enthusiasm seemingly knows no bounds. There are many more, too many to mention in this limited space – thank you all.

I would also like to thank my fellow PhD students for all the good times and for being all around good sports. I would especially like to thank Sharon Kühlmann-Berenzon, Ulrica Olofsson, Mikael Knutsson and Fredrik Altenstedt for many fun conversations and good laughs.

It still remains to be seen if I'll ever grow up to be a math professor or not, I don't have a seven year old's narrow view of available career options anymore. But I still like solving math problems.

Joachim Johansson
Göteborg, 2003

# Contents

## List of papers

The thesis is composed of the following papers

- **Paper A:** N. C. J. Johansson *Estimating the mean of heavytailed distributions*, submitted for publication.
- **Paper B:** N. C. J. Johansson *Estimating the mean of heavytailed distributions in the presence of dependence*
- **Paper C:** N. C. J. Johansson *An extreme value approach to regression through the origin with an application to superpopulation sampling*

# 1 Introduction

What are the people at UC Berkely doing on the Internet in the wee hours of the morning? They are sure up to something, at least they were between November 6 and 9 in 1996, as is apparent from Figure 1 below.

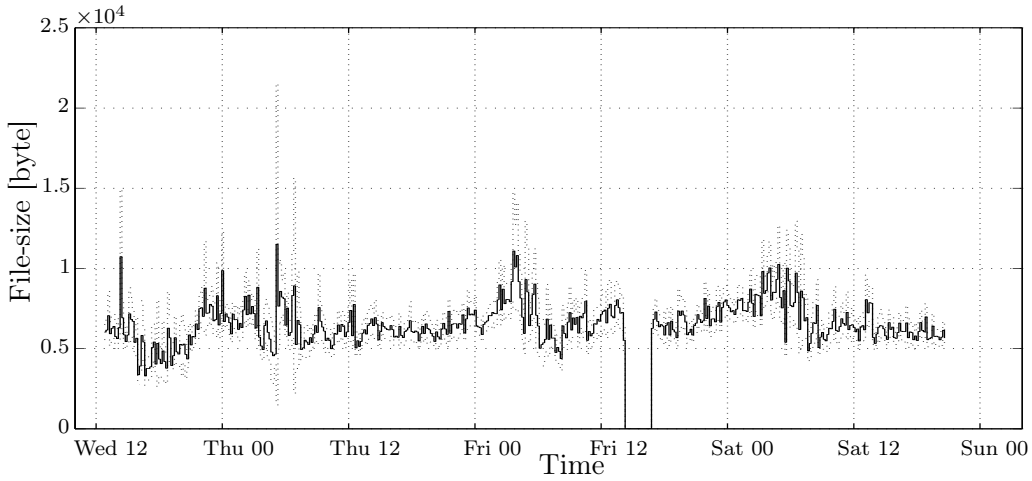


Figure 1: UC Berkeley mean file-sizes over 10 minute intervals. Dotted lines are 95% pointwise confidence intervals. Note that the network was down for approximately 2.5 hours on Friday afternoon.

The graph shows estimates of the average file-size downloaded on the Home IP service offered by UC Berkely to its students, faculty and staff. The dataset was divided into ten-minute intervals and the average estimated over each such interval. (The complete dataset is available on-line at http://ita.ee.lbl.gov/html/contrib/UCB.home-IP-HTTP.html). Note how the file-sizes seem to increase early in the morning.

A closer inspection of the data (Johansson, 2001) shows that the number of downloads decrease at night, at the same time as the files grow larger. A possible explanation for this could be that users wait until the traffic has slowed down before downloading large files – thus making the downloads faster. If this is the case, it is an interesting observation of human behaviour on the Internet which might be of use for the network administrators.

This investigation may sound like a straightforward application of basic statistics, however there are mathematical elements in this that may cause trouble. It turns out that the distribution of the file-sizes is heavy-tailed. Here, the term heavy-tailed refers to a distribution lacking finite variance. This type of distributions are common, not only in telecommunications (Crovella et al., 1998; Resnick and Rootzén, 2000; Deng, 2001) but also in finance (Davis and Mikosch, 2001), insurance (Tajvidi, 1996; Embrechts et al., 1997; Rootzén and Tajvidi, 2001) and hydrology (Naveau et al., 2001) to name but a few areas. See also Leadbetter et al. (1983); Reiss and Thomas (2001); Adler et al. (1998); Uchaikin and Zolotarev (1999) and Coles (2001). The list could be made substantially longer.

The reason the lack of finite variance may be a bit of a snag is that the common method for estimating the mean does not work in the usual manner. Recall that if $X_1, \ldots, X_n$ are iid random variables (r.v.'s) with common distribution function $F$ and with $\mathsf{E}[X_1] = \mu$ and $\mathsf{Var}(X_1) = \sigma^2 < \infty$, the well known method of estimating $\mu$ is by looking at the sample mean $\bar{X}_n = n^{-1}(X_1 + \ldots + X_n)$, which has the property that

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \rightarrow_d \mathcal{N}(0,1), \quad \text{as } n \rightarrow \infty. \tag{1}$$

This result does not hold for the case when $\mathsf{Var}(X_1) = \infty$, hence an alternative approach is needed. The two main alternatives are described below, followed by a suggestion of a third, new methodology.

## 1.1 The stable distribution approach

We begin by looking at an analogue of (1) for the case where the underlying distribution is heavy tailed.

Whether a random variable $X$ has finite variance or not is a property of the tail of its distribution. Assume that

$$\begin{aligned} &(i) \quad \lim_{x\rightarrow\infty} \mathsf{P}(X > x)/\mathsf{P}(|X| > x) = c \in [0,1] \\ &(ii) \quad \mathsf{P}(|X| > x) = x^{-\alpha}L(x), \end{aligned}$$

where $L$ is a slowly varying function at infinity, i.e $L(tx)/L(x) \rightarrow 1$ as $x \rightarrow \infty$ for all $t > 0$. See Bingham et al. (1987) for further properties of such functions.

For $\alpha \in (1,2)$ the mean of $X$ is finite but the variance is not. In this case $\bar{X}_n$ is still a valid estimate of the mean in the sense that $\bar{X}_n \rightarrow_{\text{a.s.}} \mu$. For the estimate to be useful for inferential purposes, its distribution needs to be determined. It turns out that (see e.g. Theorem 7.7 in Chapter 2 in Durrett (1996)) under assumptions $(i)$ and $(ii)$

$$\frac{\bar{X}_n - \mu}{a_n} \rightarrow_d Y, \quad \text{as } n \rightarrow \infty,$$

where $Y \sim S_\alpha(\sigma, \beta, 0)$ is a stable random variable and

$$a_n = n^{-1} \inf\{x : \mathsf{P}(|X_1| > x) \leq n^{-1}\}.$$

There is in general no closed form of the distribution of $Y$, the theory of which goes back as far as to Cauchy (1853), but its characteristic function can be written as

$$\mathsf{E}[\exp\{itY\}] = \begin{cases} \exp\{-\sigma^\alpha |t|^\alpha (1 - i\beta(\text{sign } t)\tan\frac{\pi\alpha}{2}) + it\mu\} & \text{if } \alpha \neq 1 \\ \exp\{-\sigma|t|(1 + i\beta\frac{2}{\pi}(\text{sign } t)\ln|t|) + it\mu\} & \text{if } \alpha = 1 \end{cases},$$

where $\alpha \in (0,2]$ is the index of stability, $\sigma \geq 0$ is the scale parameter, $\beta \in [-1,1]$ is the skewness parameter and the shift parameter $\mu \in \mathbb{R}$ is equal to the mean if $\alpha \in (1,2]$.

Other equivalent parametrisations exist, see Samorodnitsky and Taqqu (1994) for an overview of the theory of stable distributions. Nolan (1998) suggests a parametrisation, $S^*$, which is more suitable for computational and modelling purposes. It has the feature that the characteristic function is jointly continuous in all four parameters. Further, the mode is equal to the location parameter and, as $\alpha \to 2$, the scale tends to the variance for a normal random variable and as $\alpha \to 1$, $\sigma$ is the standard scale parameter.

In practice it is necessary to estimate the parameters of the asymptotic stable distribution of the normed mean. If $X_1, \ldots, X_n$ are iid with stable distributions then there are basically three main methods of estimation available (Nolan, 1999). The first one is the quantile method, which uses sample quantiles and match these to stable distributions, see McCulloch (1986). A second approach is to use the fact that the stable distributions have a closed form for their characteristic function (chf). Methods for using the sample chf have been suggested for instance by Kogon and Williams (1998). The third approach is to use numerical maximisation of the likelihood function, an approach made possible with the rapid development of powerful computers, see Nolan and Rajput (1995); Nolan and Panorska (1997); Nolan (1997, 1998, 1999, 2001); Nolan et al. (2001) and Uchaikin and Zolotarev (1999).

If the variables $X_1, \ldots, X_n$ are not stable, then the index of stability (or tail index) $\alpha$ may be estimated by a tail estimator. Graphically, this can be done by looking at the slope of a log-log plot of $1 - F_n$, where $F_n$ is the empirical distribution function. Another alternative is to use the popular Hill-estimator (Hill, 1975). The estimation of $\alpha$ has received a lot of attention, see for instance Embrechts et al. (1997) for an introduction and overview of the area. Note that the performance of tail estimators depends heavily on the nature of the slowly varying part $L(\cdot)$ of the distribution (see condition $(ii)$ above). For $X_1 \sim S_\alpha(\sigma, \beta, \mu)$ the polynomial tail behaviour sets in very late when $\alpha > 1.2$, see Fofack and Nolan (1999), which implies that tail estimators of $\alpha$ are not likely to be very useful for such distributions.

The skewness parameter $\beta$ can easily be estimated in some special cases. If the distribution is symmetric, then $\beta = 0$. If it has a finite right (left) end point, then $\beta = -1$ ($\beta = 1$) and $\alpha < 1$. And finally if $x^\alpha \mathsf{P}(X_1 > x) \to c_1$ and $x^\alpha \mathsf{P}(X < -x) \to c_2$ as $x \to \infty$, then $\beta = (c_1 - c_2)/(c_1 + c_2)$.

Estimation of the mean could also be made by dividing the sample $X_1, \ldots, X_n$ into parts of length $k$ and computing $\bar{X}_{k,j} = k^{-1}(X_{jk} + \ldots + X_{(j+1)k-1})$. The $\bar{X}_{k,j}$ will be approximately stable and can thus be used for estimating the parameters $\alpha$, $\beta$, $\sigma$ and $\mu$, possibly in conjunction with a bootstrapping scheme. More on bootstrap below. A difficulty would be how to select the block size $k$, see Crovella and Taqqu (1999) for one possible graphical approach for estimating $\alpha$. No other, non-bootstrap, results have been found in this area.

In summary, one of the difficulties for making inference about the mean $\mu$ is that there does not seem to exist any obvious pivotal statistic, i.e. a statistic involving only $\mu$ as the population parameter in its definition and whose limiting distribution is parameter free (Athreya et al., 1998). There are methods available for generating confidence intervals for the mean in case the observations follow a stable law – methods which require estimation of several parameters of the limiting distribution. In case the observations are not stable, more work is needed for creating a unified estimation framework.

## 1.2 Bootstrap

Bootstrapping is a resampling method that goes back to Efron (1979). It has gained increased popularity with the advent of ever faster computers. Below, we describe the basic methodology for the finite and infinite variance cases.

### 1.2.1 The bootstrap idea

A brief description of the original bootstrap idea, which does not rely on any particular class of distributions, is as follows. Assume that $X_1, \ldots, X_n$ are iid random variables with distribution function $F$ and $\mathsf{Var}(X_1) < \infty$. Further assume that there exists an estimator $\hat{\theta}$ of a parameter $\theta$, (how this estimator was discovered is unimportant to the bootstrap methodology), and that $\theta$ can be defined as $\theta = g(F(\cdot))$ for some function $g$, in principle. Then the corresponding bootstrap estimate is $\tilde{\theta} = g(F_n(\cdot))$, where $F_n$ is the empirical distribution function for the data. Since $F_n$ is close to $F$, we also expect that $\hat{\theta}$ is close in distribution to $\tilde{\theta}$.

The bootstrap method can be summarised as

1. Let $X_1^*, \ldots, X_n^*$ be iid r.v's with distribution function $F_n$. This is the bootstrap sample.
2. Compute the parameter estimate $\hat{\theta}^* = \hat{\theta}(X_1^*, \ldots, X_n^*)$.
3. Repeat 1 and 2.

This simple scheme allows us get simulated estimates of quantities such as $\mathsf{E}_*[\hat{\theta}^* - \tilde{\theta}] = \mathsf{E}_*[\hat{\theta}^* - \tilde{\theta}|F_n(x)]$, $\mathsf{P}_*(\hat{\theta}^* - \tilde{\theta} \leq x)$ and so on. The $*$ denotes that the quantities above only take the variations in the bootstrap simulations into account, while $F_n$ is kept fixed.

For the sample mean estimate $\bar{X}_n$, Singh (1981) and Bickel and Freedman (1981) showed that if

$$H_n(x, \omega) := \mathsf{P}\Big( \frac{\bar{X}_n^* - \bar{X}_n}{s_n} \leq x | X_1, \ldots, X_n \Big)$$

where

$$\bar{X}_n^* = \frac{1}{n}\sum_{k=1}^{n} X_k^*, \quad \bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k \quad \text{and} \quad s_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2$$

then $\sup_x |H_n(x,\omega) - \Phi(x)| \to 0$ with probability one. Singh (1981) showed that $H_n(\cdot,\omega)$ is a better approximation to the distribution of $(\bar{X}_n - \mu)/\sigma$ than the Edgeworth approximation up to the first order term. Further, Bickel and Freedman (1981) showed that, under second moments, the bootstrap asymptotics is the same as that supplied by the normal theory for a variety of statistics. Hence it is possible to draw conclusions from the behaviour of the bootstrap estimate and apply these to the original one. For a more thorough introduction to the area, see for instance the excellent books by Hall (1992) and Hjorth (1994).

### 1.2.2 Bootstrapping with heavy tails

However, Athreya (1987) showed that if $X_1$ is in the domain of attraction of a stable law with index $0 < \alpha \leq 2$ and $\mathsf{E}[X_1^2] = \infty$, applying the bootstrap methodology in the way described above breaks down in the sense that, for any set of real numbers $x_1, \ldots, x_n$, the sequence of random vectors $(H_n(x_i,\omega), i = 1, \ldots, k)$ converges in distribution to a non-degenerate random vector $(H(x_i,\omega), i = 1, \ldots, k)$. This means that the limiting distribution of the bootstrap mean is random (see Cuesta-Albertos and Matrán (1998) for more on this).

All is not lost, however. One method suggested by Athreya et al. (1998) uses the possibility of having a smaller order bootstrap sample $X_1^*, \ldots, X_m^*$ where $m = o(n)$. In this article the authors work under the assumption that $X_1, X_2, \ldots$ is an iid sample from $F$, where

$$1 - F(x) \sim px^{-\alpha}L(x) \quad \text{and} \quad F(-x) \sim qx^{-\alpha}L(x)$$

with $0 \leq p = 1 - q \leq 1$, $1 \leq \alpha < 2$ and $L$ is a slowly varying function at infinity.

Building on the results of Resnick (1986, 1987), let $a_n \to \infty$ be a sequence such that $n(1 - F(a_n) + F(-a_n)) \to 1$, $M_n = \max\{|X_j| : 1 \leq j \leq n\}$ and $\mu = \mathsf{E}[X_1]$. Then

$$\Big(\frac{n(\bar{X}_n - \mu)}{a_n}, \frac{M_n}{a_n}\Big) \to_d (Z_1, Z_2),$$

where $Z_1$ is a stable random variable and $Z_2 \sim G$ with

$$G(x) = \begin{cases} \exp\{-|x|^{-\alpha}\} & x > 0 \\ 0 & x \leq 0 \end{cases}.$$

Denote the distribution of $Z_1/Z_2$ by $K(\cdot, \alpha, \beta)$. The authors then show that if

$$\bar{X}_m^* = \frac{1}{m}\sum_{k=1}^{m} X_k^*, \quad M_m^* = \max\{X_j^* : 1 \leq j \leq m\} \quad \text{and} \quad T_n^* = \frac{m(\bar{X}_m^* - \bar{X}_n)}{M_m^*}$$

then

$$\sup_x |\mathsf{P}(T_n^* \leq x | X_1, \ldots, X_n) - K(x, \alpha, \beta)| = o_p(1),$$

a result which can be readily used for generating bootstrapping confidence intervals for the mean $\mu$.

There are a number of other results available on the bootstrap method in the presence of heavy tails and various forms of dependence, see for instance Lahiri (1995), Feigin and Resnick (1997), Davis and Wu (1997), Bickel et al. (1997), Hall and Jing (1998), Pictet et al. (1998), Romano and Wolf (1999), Politis et al. (1999), del Barrio and Matrán (2000) and the references therein. The examples above only serve to illustrate the methodology and are by no means an exhaustive treatment of this area.

A drawback with the subsample method is that, even when the subsample size is chosen optimally, the error between the subsample bootstrap approximation and the true distribution is often an order of magnitude larger than that of an asymptotic approximation (Hall and Jing, 1998). So, even if neither the bootstrap nor the asymptotic approximation succeeds in capturing the first term in an Edgeworth-type expansion of error, the asymptotic approach will be more accurate.

To be more specific, assume that (following Hall and Jing (1998)) the random variables $X_k$ are distributed as $\text{sign}(Y_k)|Y_k|^{-1/\alpha}$ where $Y_k$ has a continuous distribution $F_Y$. Let the distribution of $S_n = n^{-1/\alpha}\sum_{k=1}^n (X_k - \mu)$ be $G_n$. Then $G_n$ tends to a stable distribution $H(\cdot|\alpha,\sigma)$ with characteristic function $\phi(t) = \exp(-\sigma|t|^\alpha)$.

If $F_Y$ has three continuous derivatives in a neighborhood of zero, then $\mathsf{P}(|X_1| > x) = 2x^{-\alpha}F'(0) + x^{-3\alpha}F'''(0)/3 + o(x^{-3\alpha})$. The Hill estimator can be used for estimating $\alpha$ and

$$\sigma = 2F'(0)\int_0^\infty x^{-\alpha}\sin x dx = \frac{2\pi F'(0)}{\Gamma(\alpha)\sin(\alpha\pi/2)} > 0.$$

In their article, Hall and Jing (1998) show that the difference $G_n - H(\cdot|\hat\alpha,\hat\sigma)$ is of size $n^{1-2/\alpha} \vee (n^{-2/5}\log n)$.

For the subsample bootstrap method, let $\hat G_m$ be the bootstrap distribution of $m^{-1/\alpha}\sum_{k=1}^m (X_k^* - \bar X)$ conditional on the observations $X_1,\ldots,X_n$ and where $\bar X = n^{-1}\sum_{k=1}^n X_k$. That is, $\hat G_m$ is the percentile bootstrap distribution of $G_m$. If $m = o(n)$, then $\hat G_m$ also tends to $H$. In the article it is shown that, with an optimal choice of $m = m(n)$, $G_n - \hat G_m$ is of order $n^{-(\alpha-1)(2-\alpha)/\alpha}$ which is of larger order than for the asymptotic approximation.

The conclusion is that, even when a comparatively simple estimator of $\alpha$ is used, the asymptotic method outperforms the subsample bootstrap in this important class of distributions.

## 2 Semi-parametric estimation

The thesis explores a third, new, idea for estimating the mean of $F$. It can roughly be described as follows. Assume that the tails of the unknown distribution $F$ start at some levels $\pm u$. (In the papers, the thresholds for the upper and lower tails can differ). Fit a parametric model to the upper and lower tails, respectively, and weigh this together with the sample mean of the truncated variables at this level

$$\hat{\nu} = \frac{1}{n}\sum_{k=1}^{n} X_k \mathbf{1}_{\{-u<X_k<u\}}.$$

More specifically, we may write

$$\mathsf{E}[X] = \int_{-\infty}^{\infty} x dF(x) = \int_{-\infty}^{-u} x dF(x) + \int_{-u}^{u} x dF(x) + \int_{u}^{\infty} x dF(x).$$

Now make a parametric estimate of the distribution for the excesses over the thresholds $\pm u$, call these $\hat{F}$. For the middle part, $F$ is approximated by the empirical distribution function $F_n$. Hence we arrive at

$$\hat{\mathsf{E}}[X] = \int_{-\infty}^{-u} x d\hat{F}(x) + \hat{\nu} + \int_{u}^{\infty} x d\hat{F}(x).$$

A natural model for $\hat{F}$ can be found by applying the so called Peaks Over Threshold (POT) method. See for instance Smith (1987), Leadbetter (1991) or Embrechts et al. (1997). It is based on the observation that, if $F$ has polynomially decaying tails, e.g. $1 - F(x) = cx^{-1/\xi}$, for constants $c$ and $\xi$, then

$$\mathsf{P}(X - u > x | X > u) = \frac{1 - F(x+u)}{1 - F(u)} = \Big(1 + \frac{x}{u}\Big)^{-1/\xi} = \Big(1 + \xi\frac{x}{\beta}\Big)^{-1/\xi},$$

where $\beta = u\xi$. This is the generalised Pareto (GP) distribution. Hence, a natural parametric estimate $\hat{F}$ is

$$\hat{F}(x) = 1 - \hat{\mathsf{P}}(X > u)\Big(1 + \hat{\xi}\frac{x-u}{\hat{\beta}}\Big)^{-1/\hat{\xi}},$$

for suitable estimates of $\beta$ and $\xi$. See for instance Smith (1987). The resulting estimate is then of the form

$$\hat{\mathsf{E}}[X] = \hat{\mathsf{P}}(X < -u)\Big(u - \frac{\hat{\beta}}{1-\hat{\xi}}\Big) + \hat{\nu} + \hat{\mathsf{P}}(X > u)\Big(u + \frac{\hat{\beta}}{1-\hat{\xi}}\Big),$$

where $\beta/(1-\xi)$ is the expected value of the GP distribution.

The thesis investigates the semi-parametric approach to estimating the mean in three different papers which are summarised below.

## 2.1 Summary of Paper A

The approach described above was used in Paper A, which is based on the results in Johansson (2001). It is assumed that the random variables, $X_k$, are positive with distribution function $1 - F(x) = cx^{-1/\xi}(1 + x^{-\delta}L(x))$, as $x \to \infty$, where $c > 0$, $\xi \in (0, 1)$ and $\delta > 0$ are constants and $L$ is a slowly varying function at infinity.

In the article, a two parameter GP distribution was fitted to the tail and it was shown that the resulting estimator is asymptotically normally distributed and unbiased. The result is easily extendable to distributions with unbounded support and different tail-indices for the upper and lower tails.

A similar approach was considered by Peng (2001) who worked under the tail balance condition

$$\lim_{t\to\infty} \frac{\mathsf{P}(|X_1| > tx)}{\mathsf{P}(|X_1| > t)} = x^{-1/\xi}, x > 0 \quad \text{and} \quad \lim_{t\to\infty} \frac{\mathsf{P}(X_1 > t)}{\mathsf{P}(|X_1| > t)} = p \in [0, 1].$$

This time a one parameter tail was fit to the distribution.

One of the main advantages of the semi-parametric approach is that the estimate has a normally distributed limit which makes it possible to construct confidence intervals and tests for the mean. This may not always be as easily done in the bootstrap and other asymptotic approaches, as described above. Other advantages include ease of computability and that the estimate has an intuitive form.

Further, it is shown that a confidence interval based on the semi-parametric approach has a width proportional to $n^{\alpha(\xi-1/2)-1/2}$, where $\alpha \in ((2\delta\xi)^{-1}, 1)$. This can be compared to the stable distribution approach described above, where the interval width is proportional to $n^{\xi-1}$. This means that, for the infinite variance case where $\xi \in (1/2, 1)$, the semi-parametric estimator converges at a faster rate than the one based on the sample mean.

One of the problems, however, is how to select the threshold values on which the estimates of the tail parameters are based. The specific problem of threshold selection for estimating the parameters in the GP distribution is discussed e.g. in Dupuis (1998).

In Paper A the performance of the estimator is also assessed through a small simulation study. In particular, it displayed a smaller median bias than the sample mean. The methodology was also applied for estimating the mean file-size of downloaded files on a computer network. That this kind of traffic is often heavy-tailed is also described for instance in Crovella et al. (1998).

## 2.2 Summary of Paper B

Plotting the autocorrelation function (acf) is a common diagnostics tool for assessing dependence, or testing for independence, in a time series $\{X_t\}_{t=1}^n$. For the heavy tailed case the sample acf can be defined as

$$\hat{\rho}(h) = \frac{n^{-1}\sum_{t=1}^{n-h} X_t X_{t+h}}{n^{-1}\sum_{t=1}^{n} X_t^2}, \quad h = 0, 1, 2, \ldots.$$

See for instance Brockwell and Davis (1987). However, it may happen that the distribution of $\hat{\rho}(h)$ is random, see Resnick et al. (1999), which might make the use of the sample acf unreliable.

The numerator in the expression for $\hat{\rho}(h)$ is an estimate of $\mathsf{E}[X_1 X_{1+h}]$ and this quantity may be estimated using the same methodology as in Paper A. The difference is that, even if the random variables $X_k$ are independent, the sequence $\{X_1 X_{1+h}\}$ will be dependent.

Paper B extends the methodology from Paper A to the case where the random variables are $m$-dependent. This model for the dependence arises when looking at products such as $X_1 X_{1+m}$ of independent r.v's $X_k$, as described above. It may also be a natural first attempt at modelling dependence, based on graphical diagnostics methods such as the acf.

The resulting estimator is asymptotically normally distributed and unbiased. In the paper, its properties are examined by simulations for a variety of different cases and it is also used for estimating covariances. Finally, it is described how the technique may be used as a diagnostics tool for estimating the order of ARMA processes.

## 2.3 Summary of Paper C

Paper C examines the problem of regression through the origin with heavy-tailed heteroscedastic errors. More specifically, estimation of the parameter $a$ in the model

$$Y_k = aX_k + \epsilon_k \sigma(X_k)$$

is investigated. It is assumed that $\sigma(\cdot)$ is a positive function and, in the asymptotic analysis, that the constants $X_k$ and $\sigma(X_k)/X_k$ are bounded away from zero and infinity. The errors $\epsilon_k$ are iid with regularly varying tails, zero mean and infinite variance.

The estimate is based on the observation that $\mathsf{E}[Y_k/X_k] = a$ and that very few assumptions about $\sigma(\cdot)$ and $X_k$ are needed in order to apply techniques which are similar in principle to those in Paper A. The resulting estimator, $\hat{a}_M$, is asymptotically normally distributed and unbiased. It is also robust against changes in the underlying variance structure, described by $\sigma(\cdot)$.

These properties contrast those of the standard least squares estimate

$$\hat{a}_{LS} = \Big(\sum_{k=1}^{n} X_k Y_k / \sigma(X_k)^2\Big) \Big/ \Big(\sum_{k=1}^{n} X_k^2 / \sigma(X_k)^2\Big),$$

which is sensitive to outliers (Chambers, 1986). Further, $\hat{a}_{LS}$ is not asymptotically normally distributed, unless the errors $\epsilon_k$ have finite variances. Finally, it requires knowledge about the function $\sigma$.

In a small simulation study, the performance of the estimator is compared to that of some standard alternatives. It is also shown how to apply the technique for using a sample to estimate a population total under superpopulation assumptions.

## 2.4 Further work

It would be interesting to extend the model in Paper C to a superpopulation sampling context, i.e. take the sampling variability into account for different kinds of sampling distributions. It is not clear how this should be done at this time, but it is work in progress.

Another extension of Paper C is to try to apply the same ideas in a more general regression setting. A first step would be to examine how to estimate the parameters $a$ and $b$ in a model such as $Y_k = a + bX_k + \epsilon_k \sigma(X_k)$, where the $\epsilon_k$'s are heavy-tailed.

It would also be interesting to try and find better estimators for the variance of the proposed estimators. One possibility might be to apply the methodologies in conjunction with bootstrapping schemes.

## References


Adler, R. J., Feldman, R. E., and Taqqu, M. S. (eds.) (1998), *A practical guide to heavy tails*, Boston, MA: Birkhäuser Boston Inc., statistical techniques and applications, Papers from the workshop held in Santa Barbara, CA, December 1995.

Athreya, K. B. (1987), "Bootstrap of the mean in the infinite variance case," *Annals of statistics*, 15, 724–731.

Athreya, K. B., Lahiri, S. N., and Wei, W. (1998), "Inference for heavy tailed distributions," *Journal of Statistical Planning and Inference*, 66, 61–75.

Bickel, P. J. and Freedman, D. A. (1981), "Some asymptotic theory for the bootstrap," *Annals of statistics*, 9, 1196–1217.

Bickel, P. J., Götze, F., and van Zwet, W. R. (1997), "Resampling fewer than $n$ observations: gains, losses, and remedies for losses," *Statist. Sinica*, 7, 1–31, empirical Bayes, sequential analysis and related topics in statistics and probability (New Brunswick, NJ, 1995).

Bingham, N., Goldie, C., and Teugels, J. (1987), *Regular Variation*, no. 27 in Encyclopedia of Mathematics and its Applications, Cambridge University Press.

Brockwell, P. and Davis, R. (1987), *Time Series: Theory and Methods*, Springer, 2nd ed.

Cauchy, A. (1853), "Sur les résultats moyens d'observations de même nature, et sur les résultats les plus probable," *Comptes Rendus de l'Académie des Sciences de Paris*, 37, 198–206.

Chambers, R. (1986), "Outlier robust finite population estimation," *Journal of the American Statistical Association*, 81, 1063–1069.

Coles, S. (2001), *An introduction to statistical modeling of extreme values*, Springer Series in Statistics, London: Springer-Verlag London Ltd.

Crovella, M. E. and Taqqu, M. S. (1999), "Estimating the heavy tail index from scaling properties," *Methodol. Comput. Appl. Probab.*, 1, 55–79.

Crovella, M. E., Taqqu, M. S., and Bestavros, A. (1998), "Heavy-Tailed Probability Distributions in the World Wide Web," in *A Practical Guide to Heavy Tails*, eds. Adler, R. J., Feldman, R. E., and Taqqu, M. S., Birkhäuser, pp. 3–25.

Cuesta-Albertos, J. A. and Matrán, C. (1998), "The asymptotic distribution of the bootstrap sample mean of an infinitesimal array," *Ann. Inst. H. Poincaré Probab. Statist.*, 34, 23–48.

Davis, R. A. and Mikosch, T. (2001), "Point process convergence of stochastic volatility processes with application to sample autocorrelation," *J. Appl. Probab.*, 38A, 93–104, probability, statistics and seismology.

Davis, R. A. and Wu, W. (1997), "Bootstrapping $M$-estimates in regression and autoregression with infinite variance," *Statist. Sinica*, 7, 1135–1154.

del Barrio, E. and Matrán, C. (2000), "The weighted bootstrap mean for heavy-tailed distributions," *J. Theoret. Probab.*, 13, 547–569.

Deng, Q. (2001), *Queues with regular variation*, Beta. Technische Universiteit Eindhoven. Research School for Operations Management and Logistics, D44, Eindhoven: Eindhoven University of Technology Department of Mathematics and Computing Science, dissertation, Technische Universiteit Eindhoven, Eindhoven, 2001.

Dupuis, D. J. (1998), "Exceedances over High Thresholds: A Guide to Threshold Selection," *Extremes*, 1, 251–261.

Durrett, R. (1996), *Probability: Theory and Examples, Second Edition*, Duxbury Press.

Efron, B. (1979), "Bootstrap methods: another look at the jackknife," *Ann. Statist.*, 7, 1–26.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997), *Modelling Extremal Events*, Springer.

Feigin, P. D. and Resnick, S. I. (1997), "Linear programming estimators and bootstrapping for heavy tailed phenomena," *Adv. in Appl. Probab.*, 29, 759–805.

Fofack, H. and Nolan, J. P. (1999), "Tail behavior, modes and other characteristics of stable distributions," *Extremes*, 2, 39–58.

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, Springer.

Hall, P. and Jing, B.-Y. (1998), "Comparison of bootstrap and asymptotic approximations to the distribution of a heavy-tailed mean," *Statist. Sinica*, 8, 887–906.

Hill, B. M. (1975), "A simple general approach to inference about the tail of a distribution," *Ann. Statist.*, 3, 1163–1174.

Hjorth, J. S. U. (1994), *Computer intensive statistical methods*, Chapman & Hall.

Johansson, N. C. J. (2001), "A semiparametric estimator of the mean of heavytailed distributions," Licentiate thesis, Chalmers University of Technology, Sweden.

Kogon, S. M. and Williams, D. B. (1998), "Characteristic Function Based Estimators of Stable Distribution Parameters," in *A Practical Guide to Heavy Tails*, eds. Adler, R. J., Feldman, R. E., and Taqqu, M. S., Birkhäuser, pp. 311–335.

Lahiri, S. N. (1995), "On the asymptotic behaviour of the moving block bootstrap for normalized sums of heavy-tail random variables," *Ann. Statist.*, 23, 1331–1349.

Leadbetter, M. R. (1991), "On a basis for "peaks over threshold" modeling," *Statist. Probab. Lett.*, 12, 357–362.

Leadbetter, M. R., Lindgren, G., and Rootzén, H. (1983), *Extremes and related properties of random sequences and processes*, Springer Series in Statistics, New York: Springer-Verlag.

McCulloch, J. H. (1986), "Simple consistent estimators of stable distribution parameters," *Comm. Statist. B—Simulation Comput.*, 15, 1109–1136.

Naveau, P., Katz, R., and Moncrieff, M. (2001), "Extremes and Climate: an Introduction and a Case Study," in *Notes de l'Institut Pierre Simon Laplace*, vol. 11, pp. 4–22.

Nolan, J. P. (1997), "Numerical calculation of stable densities and distribution functions," *Comm. Statist. Stochastic Models*, 13, 759–774, heavy tails and highly volatile phenomena.

— (1998), "Parameterizations and modes of stable distributions," *Statist. Probab. Lett.*, 38, 187–195.

— (1999), "Fitting Data and Assessing Goodness-of-fit with Stable Distributions," Available from http://academic2.american.edu/ jpnolan/stable/stable.html.

— (2001), "Maximum likelihood estimation and diagnostics for stable distributions," in *Lévy processes*, Boston, MA: Birkhäuser Boston, pp. 379–400.

Nolan, J. P. and Panorska, A. K. (1997), "Data analysis for heavy tailed multivariate samples," *Comm. Statist. Stochastic Models*, 13, 687–702, heavy tails and highly volatile phenomena.

Nolan, J. P., Panorska, A. K., and McCulloch, J. H. (2001), "Estimation of stable spectral measures," *Math. Comput. Modelling*, 34, 1113–1122, stable non-Gaussian models in finance and econometrics.

Nolan, J. P. and Rajput, B. (1995), "Calculation of multidimensional stable densities," *Comm. Statist. Simulation Comput.*, 24, 551–566.

Peng, L. (2001), "Estimating the mean of a heavy tailed distribution," *Statistics and probability letters*, 52, 255–264.

Pictet, O. V., Dacorogna, M. M., and Müller, U. A. (1998), "Hill, bootstrap and jackknife estimators for heavy tails," in *A practical guide to heavy tails (Santa Barbara, CA, 1995)*, Boston, MA: Birkhäuser Boston, pp. 283–310.

Politis, D. N., Romano, J. P., and Wolf, M. (1999), *Subsampling*, Springer Series in Statistics, New York: Springer-Verlag.

Reiss, R.-D. and Thomas, M. (2001), *Statistical analysis of extreme values*, Basel: Birkhäuser Verlag, 2nd ed., from insurance, finance, hydrology and other fields, With 1 CD-ROM (Windows).

Resnick, S. and Rootzén, H. (2000), "Self-similar communication models and very heavy tails," *Ann. Appl. Probab.*, 10, 753–778.

Resnick, S., Samorodnitsky, G., and Xue, F. (1999), "How misleading can sample ACFs of stable MAs be? (Very!)," *Ann. Appl. Probab.*, 9, 797–817.

Resnick, S. I. (1986), "Point processes, regular variation and weak convergence," *Advances in applied probability*, 18, 66–138.

— (1987), *Extreme values, regular variation and point processes*, Springer.

Romano, J. P. and Wolf, M. (1999), "Subsampling inference for the mean in the heavy-tailed case," *Metrika*, 50, 55–69.

Rootzén, H. and Tajvidi, N. (2001), "Can losses caused by wind storms be predicted from meteorological observations?" *Scand. Actuar. J.*, 162–175.

Samorodnitsky, G. and Taqqu, M. S. (1994), *Stable Non-Gaussian Random Processes*, Chapman & Hall.

Singh, K. (1981), "On the asymptotic accuracy of Efron's bootstrap," *Annals of statistics*, 9, 1187–1195.

Smith, R. (1987), "Estimating Tails of Probability Distributions," *The Annals of Statistics*, 15, 1174–1207.

Tajvidi, N. (1996), "Characterisation and Some Statistical Aspects of Univariate and Multivariate Generalised Pareto Distributions," Ph.D. thesis, Chalmers University of Technology.

Uchaikin, V. V. and Zolotarev, V. M. (1999), *Chance and stability*, Modern Probability and Statistics, Utrecht: VSP, stable distributions and their applications, With a foreword by V. Yu. Korolev and Zolotarev.

# Paper A

# Estimating the mean of heavytailed distributions

Joachim Johansson*

May 6, 2003

**Abstract**

An asymptotically normally distributed estimate for the expected value of a positive random variable with infinite variance is introduced. Its behavior relative to estimation using the sample mean is investigated by simulations. An example of how to apply the estimate to file-size measurements on Internet traffic is also shown.

Keywords: Pareto distribution, mean estimation, heavy tailed distributions, telecommunication

MSC2000 classification: primary – 62G32
secondary – 62F10, 62F12, 62P30

## 1 Introduction

When modelling phenomena in telecommunications, finance, insurance or sociology, to name but a few areas, heavy-tailed distributions are sometimes encountered. Heavy-tailed is used here to describe a distribution with the property that some of its moments are infinite. In telecommunications it has been observed that the distribution of file sizes on the Internet has this property, see e.g [2]. And in insurance, heavy tails are encountered when modelling for instance fire and storm damages [4, 14]. In sociology it might be interesting to estimate the average income in a population. It would be reasonable to assume that only a small number of individuals have a large income and thus that the distribution has a pareto like tail. There are many other situations where similar reasoning could be applied.

A large body of work is available on estimating quantiles of heavy-tailed distributions, see [4] for an excellent introduction and overview of this area. In this paper, however, we attempt to estimate the expected value of the distribution instead.

*Chalmers University of Technology, Göteborg, Sweden. E-mail: joachimj@math.chalmers.se

In the second section of the paper, the estimate is introduced and its properties examined. This is followed by a simulation study of its behaviour relative to estimating the mean using the sample mean. Finally, two applications of the estimate to telecommunications data are shown.

## 2 The Estimate

Let $X, X_1, X_2, \ldots, X_n$ be iid positive random variables with distribution function $F$, where

$$\bar{F}(x) := 1 - F(x) = cx^{-1/\xi}(1 + x^{-\delta}L(x)), \tag{1}$$

for $\xi \in (0, 1)$, $\delta > 0$, $L \in RV_0$ and some constant $c$. $L \in RV_0$ means that $L$ is a slowly varying function, i.e $L(tx)/L(x) \to 1$ as $x \to \infty$ for all $t > 0$. For further properties of these functions, see Chapter 0 in [10], or either of [12] or [1], where the latter is very extensive.

For $\xi \in (0, 1/2)$, $X$ has finite variance and estimation of $\mathsf{E}[X]$ could be done using the sample mean $\bar{X} = (X_1 + \ldots + X_n)/n$ which, by the Central Limit Theorem, is asymptotically normal as $n \to \infty$. If $\xi \in (1/2, 1)$, $\bar{X}$ converges to a stable distribution (see e.g. [3]) and is still a valid estimate. There are, however, some difficulties in estimating the parameters of the limiting stable distribution. If the sample is big enough, it could be partitioned into sub-samples, for which the mean of each could be calculated. Such a procedure would render $\bar{X}_1, \ldots, \bar{X}_k$, which would be iid with a distribution that is nearly stable. Inference about the unknown parameters could then for instance be based on maximum likelihood methods. If the sample size, $n$, is not big enough to accommodate this method, bootstrapping, as described in [11], is an alternative.

Below, a different procedure is suggested for the model in (1). It gives estimators which are asymptotically normally distributed and unbiased with an easily estimated variance. In the calculations, $\delta$ will not be estimated. The reason for this is that, normally, the sample size $n$ would have to be very large in order to estimate it reasonably well. Instead the properties of the mean estimate will be examined as functions of $\delta$.

The standard estimate of $\mathsf{E}[X]$ is

$$\bar{X} = \int x dF_n(x) = \frac{1}{n}\sum_{k=1}^{n} X_k,$$

where $F_n$ is the empirical distribution function. Building on this we propose an estimate of the form

$$\hat{\mathsf{E}}[X] := \hat{M} := \hat{\mu} + \hat{\tau} := \int_0^{u_n} x dF_n(x) + \int_{u_n}^{\infty} x d\hat{F}(x), \tag{2}$$

where $\hat{\tau}$ is the part of $\hat{M}$ originating from the tail of the distribution. The tail is assumed to start at some level $u_n$, which in the analysis will be assumed to tend to infinity. $\hat{F}$ is an estimate of the tail distribution function, as described below.

Let $F_u(y) = \mathsf{P}(X - u_n \leq y | X > u_n)$ be the distribution of the excesses over the threshold $u_n$. It follows from (1) that

$$\bar{F}_u(y) = \frac{\bar{F}(u_n + y)}{\bar{F}(u_n)} = \Big(1 + \frac{y}{u_n}\Big)^{-1/\xi} \frac{1 + (u_n + y)^{-\delta} L(u_n + y)}{1 + u_n^{-\delta} L(u_n)}, \tag{3}$$

and if $\beta_n = \beta(u_n) = u_n \xi$, then $\bar{F}_u(y)$ is a perturbed GPD, where the df of the generalised Pareto distribution (GPD) has the form

$$G_{\beta,\xi}(x) = \begin{cases} 1 - \Big(1 + \xi \frac{x}{\beta}\Big)^{-1/\xi}, & \xi \neq 0 \\ 1 - e^{-x/\beta}, & \xi = 0 \end{cases}, \quad x \in \begin{cases} [0, \infty), & \xi \geq 0 \\ [0, -\beta/\xi], & \xi < 0 \end{cases}.$$

This means that for large values of $u_n$, $F_u(y) \approx G_{\beta(u_n),\xi}(y)$ in the sense that

$$\lim_{u_n \uparrow y_F} \sup_{0 < y < y_F - u_n} |F_u(y) - G_{\beta(u_n),\xi}(y)| = 0,$$

where $y_F$ is the right end point of $F$ and $\beta$ is some positive function. See also Theorem 3.4.13 in [4].

By definition $\bar{F}(u_n + y) = \bar{F}(u_n)\bar{F}_u(y)$. And, for $N = N_n = |\{i : X_i > u_n\}|$, the number of $X_i$'s which exceed $u_n$, we have $N \sim \text{Bin}(n, p_n)$ with $p_n = \mathsf{P}(X_1 > u_n)$, and estimation of $p_n = \bar{F}(u_n)$ may be done using

$$\hat{p} = \hat{\bar{F}}(u_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i > u_n\}} = \frac{N}{n}.$$

For large values of $u_n$, use

$$\hat{\bar{F}}_u(y) = \bar{G}_{\hat{\beta},\hat{\xi}}(y),$$

for appropriate estimates $\hat{\xi} = \hat{\xi}_n$ and $\hat{\beta}_n = \hat{\beta}(u_n)$. Note that $\beta$ will be estimated separately, i.e. $\beta = \xi u_n$ will not be used. The reason for this is to achieve greater flexibility in the parameter fitting, compensating for the underlying distribution not being an exact GPD.

We have now arrived at an alternative estimate, $\hat{M}$, of $\mathsf{E}[X]$,

$$\begin{aligned}\hat{M} &:= \int_0^{u_n} x dF_n(x) + \int_{u_n}^{\infty} x d\hat{F}(x) \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i \mathbf{1}_{\{X_i \le u_n\}} + \int_{u_n}^{\infty} x \frac{N}{n\hat{\beta}}\Big(1+\hat{\xi}\frac{x-u_n}{\hat{\beta}}\Big)^{-1-1/\hat{\xi}} dx \qquad (4) \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i \mathbf{1}_{\{X_i \le u_n\}} + \hat{p}\Big(u_n + \frac{\hat{\beta}}{1-\hat{\xi}}\Big), \text{ for } \hat{\xi} \in (0,1), \qquad (5)\end{aligned}$$

where $\hat{p} = N/n$. As seen above, the main interest is in the case $\xi \in (1/2, 1)$. It will be seen later that, if $\xi \in (0,1)$ then $\mathsf{P}(\hat{\xi}_n \in (0,1)) \to 1$, so the convergence of the integral will asymptotically not be a problem.

If $\hat{\xi} \ge 1$, then $\hat{M}$ should be set to $\infty$ since this would indicate that the first moment of the distribution in (1) is infinite.

There are different ideas on how to select the threshold level $u_n$. In the applications shown later, $u_n$ was selected so that the estimates of $\beta$ and $\xi$ were stable around that value of $u_n$, see Section 3, Figure 6.

**Theorem 2.1 (Distribution of $\hat{M}$)**
Let $X_1, \ldots, X_n$ be positive iid random variables with distribution function $F$, such that $\bar{F}(x) = cx^{-1/\xi}(1 + x^{-\delta}L(x))$, for some constants $c > 0$, $\xi \in (0,1)$ and $\delta > 1/2\xi$, where $L(x)$ is a slowly varying function such that $x^{-\delta}L(x)$ is non-inceasing and $L(x)$ is locally bounded in $[x_0, \infty)$ for some $x_0 \ge 0$. With $u_n = O_+(n^{\alpha\xi})$ for some $\alpha \in (1/2\delta\xi, 1)$, $p_n = \mathsf{P}(X_1 > u_n)$, $\mu_n = \mathsf{E}[X_1 \mathbf{1}_{\{X_1 \le u_n\}}]$, $\gamma_n^2 = \mathsf{Var}(X_1 \mathbf{1}_{\{X_1 \le u_n\}})$, $\beta$ as in (3), $M = \mathsf{E}[X_1]$ and $\hat{M}$ as in (5),

$$\frac{\sqrt{n}}{\gamma_n\sqrt{k_n}}(\hat{M} - M) \to_d \mathcal{N}(0,1),$$

where

$$k_n = 1 + \frac{p_n(1-p_n)}{\gamma_n^2}\Big(u_n + \frac{\beta}{1-\xi}\Big)^2 + \frac{p_n\beta^2}{\gamma_n^2}\frac{(1+\xi)^2}{(1-\xi)^4} = O_+(1),$$

and $O_+(1)$ denotes a sequence bounded away from zero and infinity.

Proof: See Appendix A. □

Note that there is no closed expression for the asymptotic variance of the estimate $\hat{M}$ under the conditions in Theorem 2.1. This is because only the tail behaviour of the distribution function $F$ is specified and hence the assumptions do not determine the value of $\gamma_n^2$. An exact limit could be obtained, should $F$ be more closely specified, but this would in turn mean placing extra conditions on $L(x)$. Since $L(x)$ is not known in practice, such conditions would be of limited practical use. In applications, $\gamma_n^2$ is estimated by the sample variance for the truncated variables $\{X_k \mathbf{1}_{\{X_k \le u_n\}}\}$.

Also note that a confidence interval for the expected value, based on Theorem 2.1, would have a width proportional to $\gamma_n/\sqrt{n} \propto n^{\alpha(\xi-1/2)-1/2}$. This can be compared to a confidence interval based on the sample mean, which has width proportional to $n^{\xi-1}$. This means that, for the infinite variance case $\xi \in (1/2, 1)$, $\hat{M}$ converges faster than $\bar{X}_n$.

# 3 Simulations

Usually the expected value is estimated using the sample mean, $\bar{X}_n$ which, properly normed and in the infinite variance case, tends to a stable distribution, see e.g. Theorem 7.7 in Chapter 2 in [3].

Simulations were made in order to determine the behaviour of $\hat{M}$, and its relation to $\bar{X}_n$. The parameters $\beta$ and $\xi$ were estimated using the Splus program-package EVIS, Version 3, written by Alexander J McNeil. Refinements could be made for these estimates, of course, and there is a large body of work done in this area, see for instance [13] or [14] and the references therein. Improvements to $\bar{X}_n$ could also be made, for instance by bootstrapping from the sample as described in [11].

Depending on the value of $\hat{\xi}$, different estimates of the mean were made, as illustrated in Figure 1. $\hat{\xi} > 1$ would indicate that the mean is infinite, and so the estimate should be infinity in this case. Further, if $\hat{\xi} < 0$ this would indicate a distribution with finite tail and then the ordinary Central Limit Theorem will be used.

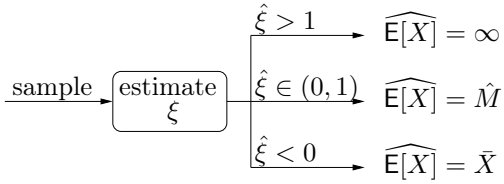


Figure 1: The figure shows how $\hat{M}$ is used in the simulations.

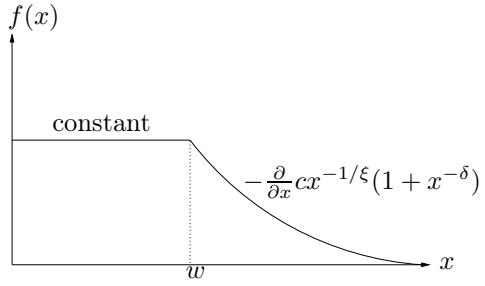


Figure 2: The distribution used in the simulations.

A few simulations were done where samples were drawn from the pdf in Figure 2 with parameters $w = 10$, $\xi = 0.7$ and $\delta = 1$, corresponding to a mean of approximately 15.7. 500 samples, varying in size from 1000 up to 100000, were generated and the estimates calculated for different values of the threshold $u_n$. If the perturbation $1 + x^{-\delta}$ was replaced by $1 - x^{-\delta}$, similar results were obtained.

An additional simulation was made with $\xi = 0.3$ and the other parameters unchanged. This case corresponds to the variance being finite, and again $\hat{M}$ was compared to $\bar{X}_n$. The results of the simulations are found in Figure 3. As can be seen in the figure, $\hat{M}$ and $\bar{X}$ give similar results in this case. This could be expected since the tail is comparatively light and does not influence the estimate all that much.

For $u_n = 5$ the estimate is very biased. This is as expected since the density, $f(x)$, is constant up to $x = 10$ and only after that decreases polynomially.

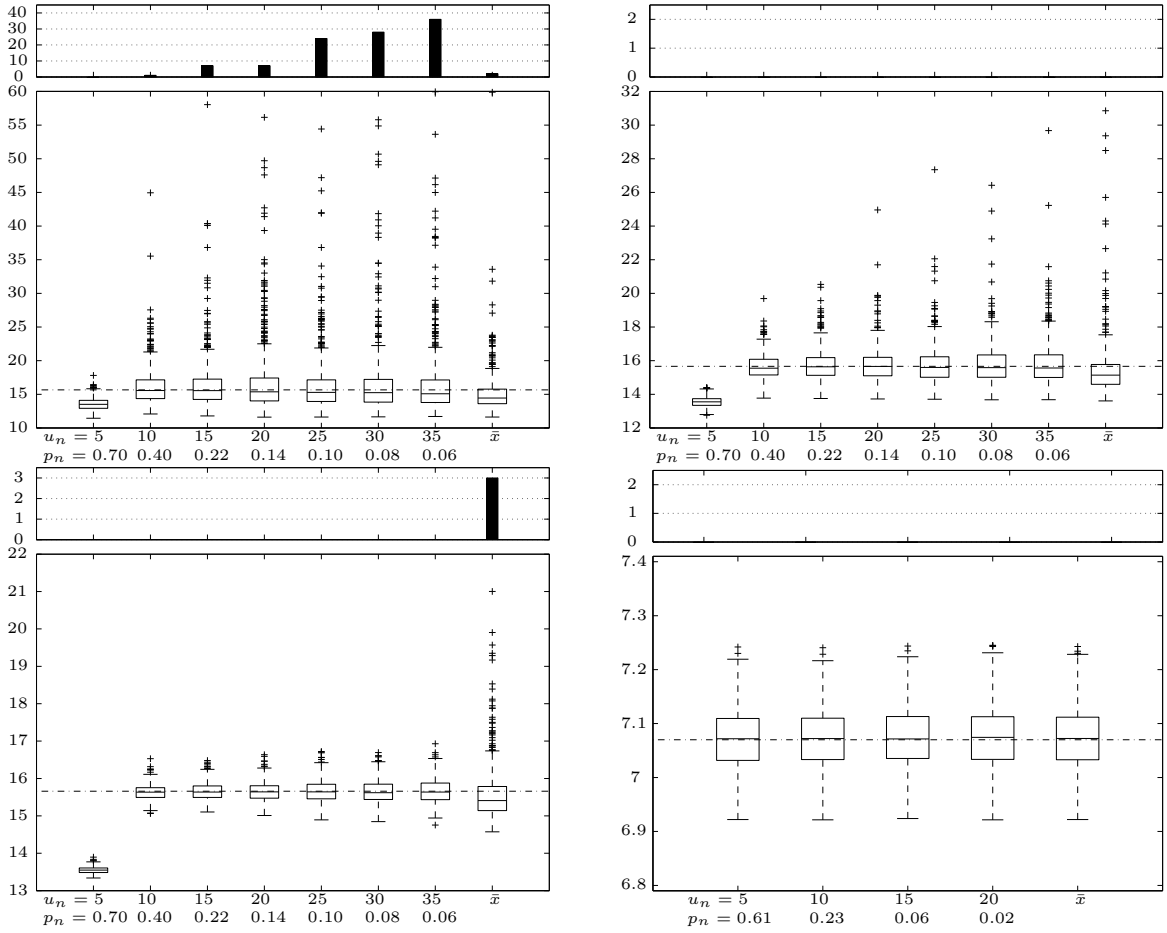


Figure 3: The results from the simulations. The sample sizes for $\xi = 0.7$ were *top left*: $10^3$, *top right*: $10^4$, *bottom left*: $10^5$ and *bottom right*: for $\xi = 0.3$ the sample size was $10^4$. $u_n$ is the selected threshold and $p_n = \bar{F}(u_n)$. The bar-charts indicate how many observations were too large to be displayed in the box-plots underneath.

## Applications

The aim of this section is to outline applications of $\hat{M}$ to telecommunication data. Focus will be on the estimate, rather than on how to model network traffic, which is a large area on its own. References to this area will therefore be very sketchy and we will not go into any details, but see for instance [6], [8], [5], [7] and its companion paper [9], for suggestions of different models.

When attempting to describe the behaviour of Internet-traffic data one often arrives at a model like

$$X(t) =_d \mu + \text{noise},$$

where $X(t)$ describes the traffic intensity into a network node at time $t$, $\mu$ is a constant mean traffic intensity and the noise could for instance be fractional Brownian noise. This model is interesting when analysing network behaviour and in particular may influence decisions about augmenting the network.

Sometimes it is more interesting to look at the accumulated traffic into a node

$$A(t) = \int_0^t X(s)ds =_d \nu t + \text{noise},$$

where $\nu$ is a constant and noise could be fractional Brownian motion. Such a process is useful for bandwidth and buffer dimensioning in the network. There are various models for the processes $X$ and $A$, but most of them have in common that $\mu$ or $\nu$ are expected values of some random variable. Hence, in order to have some idea of how $X$ or $A$ behaves, this mean will have to be estimated. In the heavy-tailed setting discussed earlier, $\hat{M}$ could be used for this purpose.

### HTTP data

The data examined is the file sizes in HTTP traffic from a modem pool. It was generously supplied by Dr Attila Vidács at the High Speed Networks Laboratories at the Technical University of Budapest, and is displayed in Figure 4 below.

Note that there are six large observations, five of which appear closely together. A more detailed look at the data shows that five of these were generated by the same user and that four appeared back to back in time and originated from the same web site. This kind of behaviour might be observed for instance when the HTTP protocol is used for file transfers instead of using FTP. It is not clear that this was the case, but in order to keep the model simple, these six large observations were taken out when the mean was estimated, so that the Pareto-tail assumption would more realistic, see Figure 5.

The mean was then estimated using the same strategy as in Figure 1. The result is displayed in Figure 6 below. Visual inspection shows that the mean might be around $1.25 \cdot 10^4$. Estimation using the sample mean, $\bar{X}_n$, without removing the six largest

observations resulted in an estimate of about $1.5 \cdot 10^4$, and when the observations were removed this number changed to slightly below $1.2 \cdot 10^4$.

Two observations can be made. First, estimation based on the sample mean is sensitive to outliers, as demonstrated by the large difference between the estimates above when 6 out of 7627 observations are removed from the sample. Secondly, it would be difficult to find a simple model for this situation. An alternative estimation strategy would be separate modelling of the six large observations and then concatenation of the resulting two estimates. When using $\bar{X}_n$, such a strategy would result in just using the whole sample. For $\hat{M}$, though, this might not necessarily be the case. The problem would be that six observations would not be enough for making a good estimate.

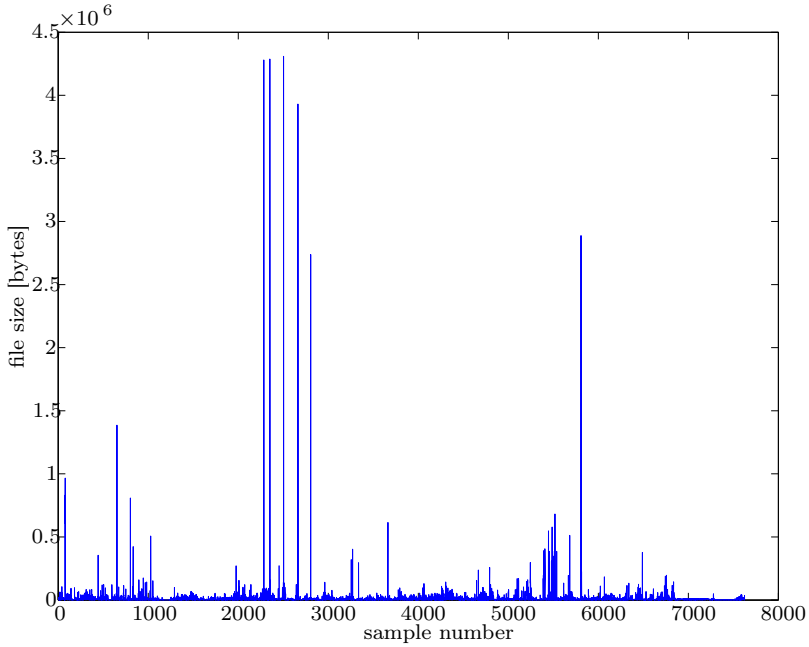


Figure 4: File sizes [bytes].

### A regression-type application

This section presents a regression-type application of the mean-estimation procedure described earlier. This procedure will, in turn, be compared to the standard methods otherwise used.

The dataset to be investigated is a portion of the publically available measurements of HTTP data gathered from the Home IP service offered by UC Berkeley to its students, faculty and staff. The dataset is described in detail on http://ita.ee.lbl.gov /html/contrib/UCB.home-IP-HTTP.html. Due to the size of the dataset, only the measurements between Wed Nov 6 12:46:59 1996 and Sat Nov 9 20:47:01 1996, were used. And of these, only the GET requests were looked at. This left 1 577 582

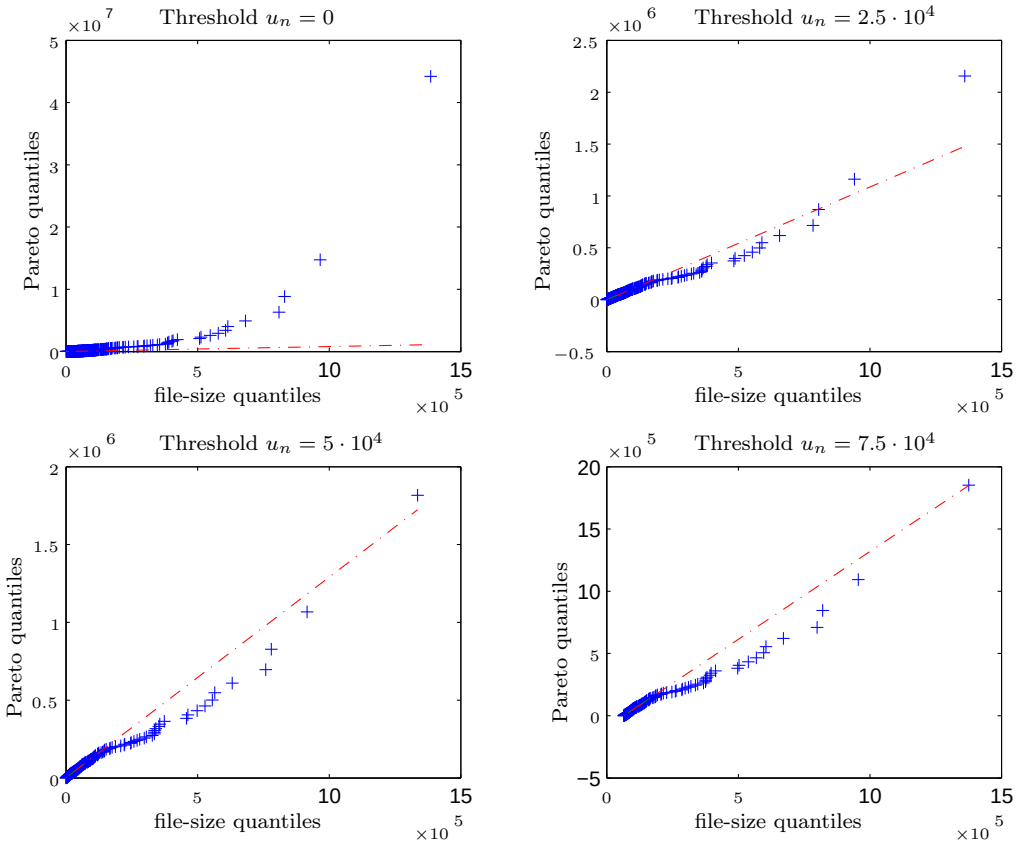


Figure 5: QQ-plot of data against a generalised Pareto distribution. The six largest observations have been removed.

observations of file down-loads to be examined.

The dataset was then divided into 10-minute intervals and the mean file-size was estimated for each of these. The division into small intervals was done so that trends and other non-stationarities would be negligible in each small interval. Then, estimates of the parameters $\beta$ and $\xi$, as in Equation (5) were calculated for different thresholds $u_n$. Which threshold to choose was based on examination of qq- and probability-plots and on comparing the excesses over the treshold to a generalised Pareto distribution using a Kolmogorov-Smirnov test. The threshold for each interval was selected so that it passed the test at the 5% significance level. The result is presented in Figure 7.

The estimation procedure used to generate Figure 7 should be compared to the one using the sample mean for the same disjoint intervals. In Figure 8 a plot of the difference between these estimates is displayed. Note that the sample average tends to yield larger estimates for this data set.

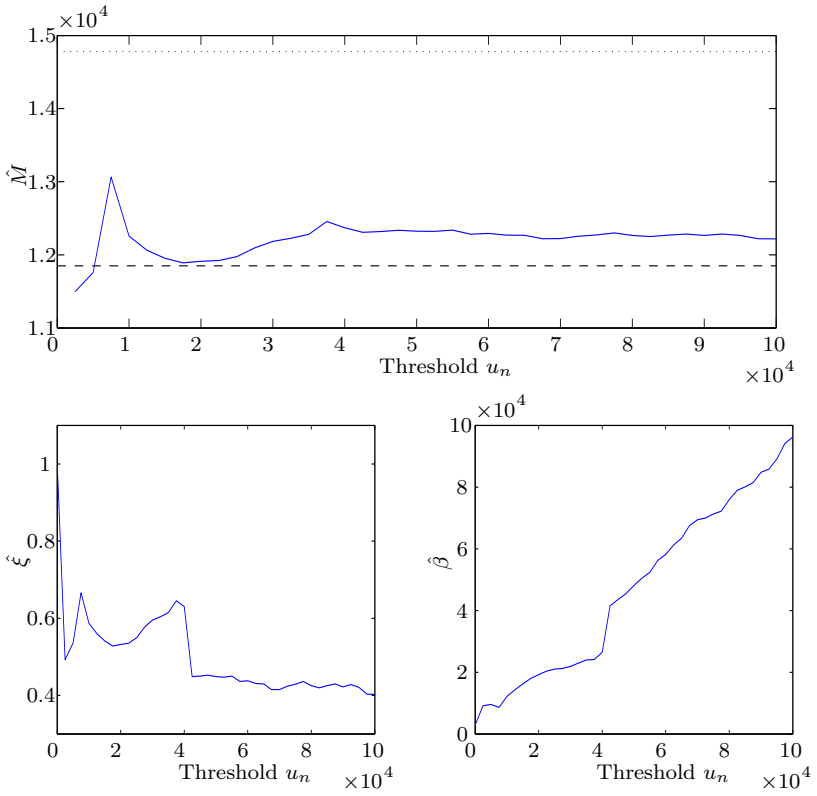


Figure 6: *Top:* Estimates of the mean for different values of the threshold $u_n$, compared to the sample mean, $\bar{X}_n$, based on all observations (dotted line) and with the six largest observations removed (dashed line). *Bottom left:* Estimation of $\xi$ for different values of $u_n$. *Bottom right:* Estimation of $\beta$ for different values of $u_n$.

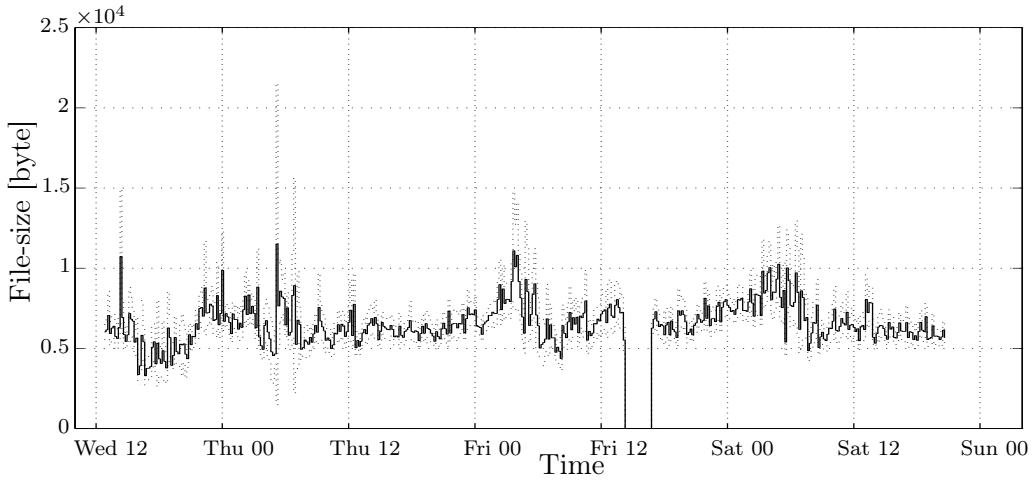


Figure 7: UC Berkeley mean file-sizes over 10 minute intervals. Dotted lines are 95% pointwise confidence intervals. Note that the network was down for approximately 2.5 hours on Friday afternoon. For the cases where $\hat{\xi} < 0$, the mean was estimated using the sample mean. No confidence interval is supplied for these observations.

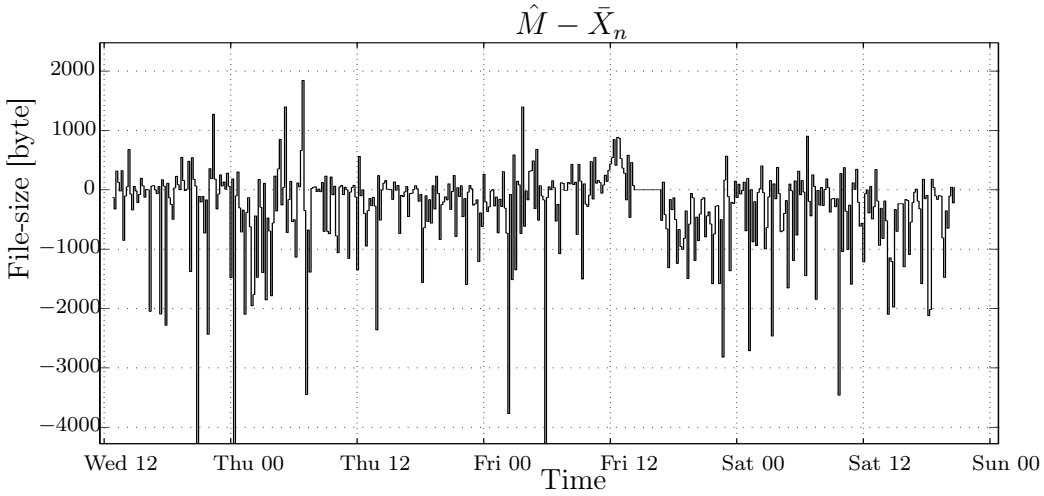


Figure 8: Difference between the two estimates, $\hat{M} - \bar{X}_n$. Not completely shown are two large negative observations of size $-3.5 \cdot 10^5$ and $-1 \cdot 10^5$ around Thursday midnight.

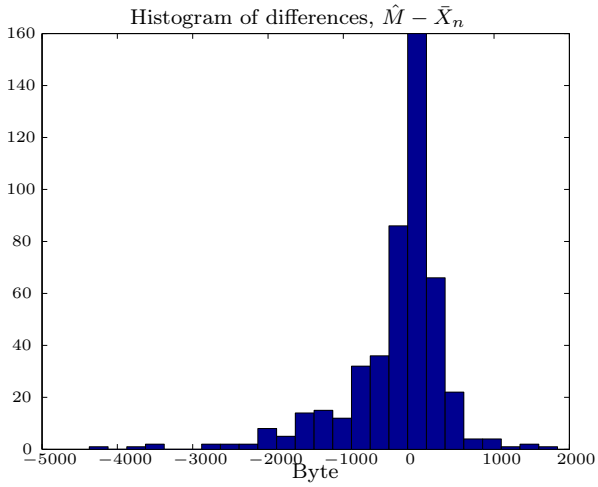


Figure 9: Histogram of the differences between the estimates, $\hat{M} - \bar{X}_n$. The two largest negative observations are not shown.

The number of observations used for estimating the tail-parameters $\xi$ and $\beta$ is displayed in Figure 11. Note that few observations were used for Wednesday afternoon, due to a drop in overall traffic. The relatively few observations, in turn, lead to greater variability in the estimate of the mean file-size for the same period.

Examining the results, it appears that file-sizes increase slightly after midnight. This might indicate that users wait for periods of low traffic intensity before downloading large files, thus making the downloads faster, or it could be due to network administration tasks.

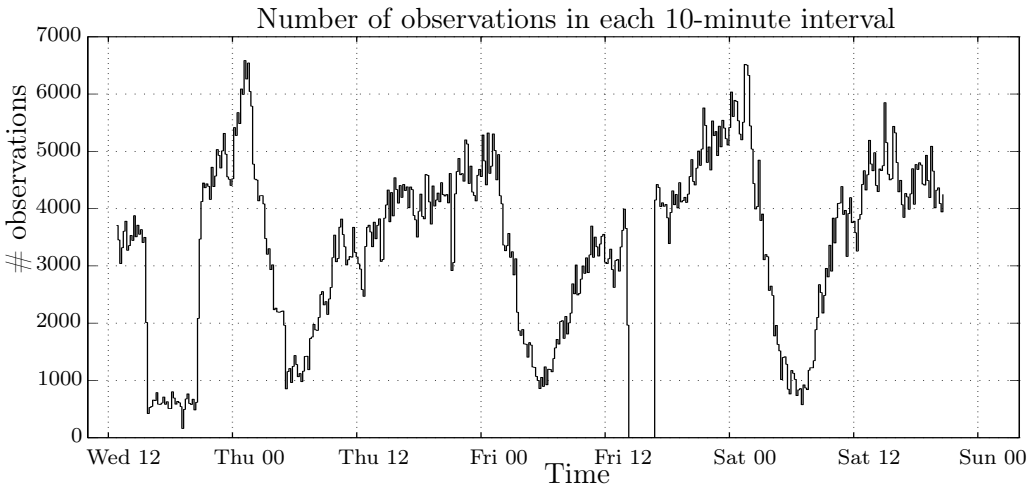


Figure 10: Number of observations in each 10-minute interval.

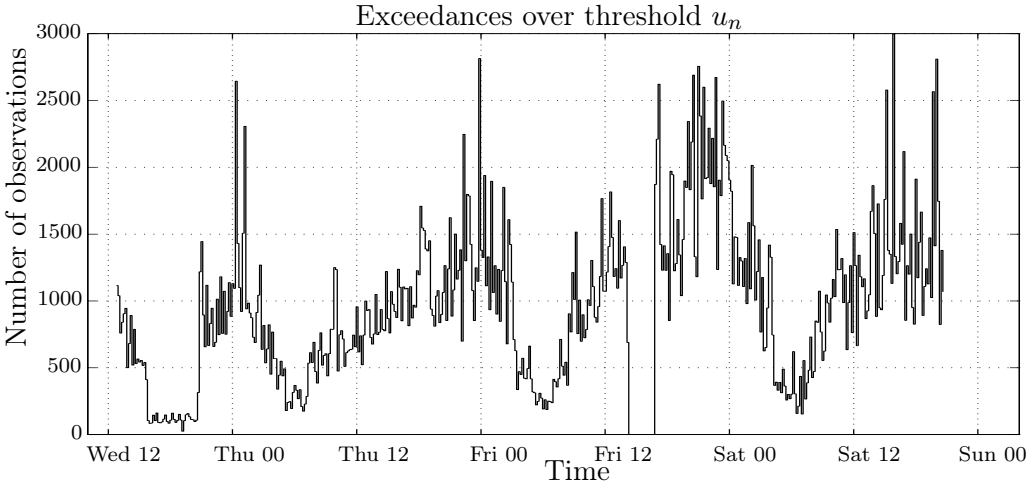


Figure 11: Number of observations exceeding the selected threshold. This corresponds to the number of observations that were used for estimating the parameters $\xi$ and $\beta$.

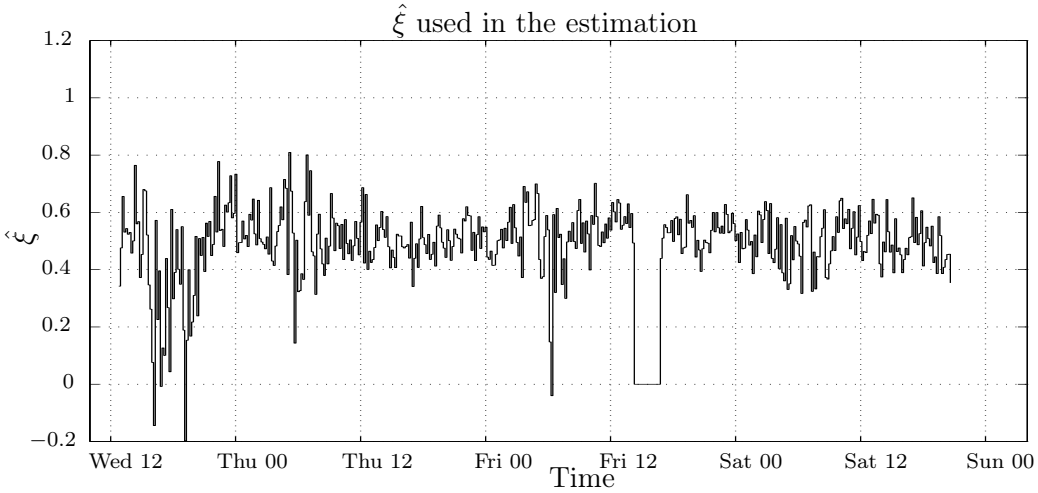


Figure 12: The estimated value of $\xi$ for each of the 10-minute intervals. When $\hat{\xi} < 0$, the mean was estimated using the sample mean.

# A Proof of Theorem 2.1

First note the following straightforward result.

**Proposition A.1**
Let $X$ be a positive random variable with distribution function $F$ as in (1) and let $u_n \to \infty$ as $n \to \infty$ be a sequence of numbers. Then

$$p_n = \mathsf{P}(X_1 > u_n) = O_+(u_n^{-1/\xi})$$
$$\gamma_n^2 = \mathsf{Var}(X_1 \mathbf{1}_{\{X_1 \le u_n\}}) = O_+(u_n^{2-1/\xi}),$$

where $r_n = O_+(a_n)$ denotes a sequence such that $r_n/a_n$ is bounded away from zero and infinity.

Proof: $\bar{F}(x) = cx^{-1/\xi}(1 + o(1)) = x^{-1/\xi}O_+(1)$ and the results follow from standard calculations, using $\mathsf{E}[X^k] = \int_0^\infty kx^{k-1}\mathsf{P}(X > x)dx$ for positive random variables $X$ and $k = 1, 2, 3, \ldots$. □

Theorem 3.2 in [13] gives us the asymptotic distribution of the tail parameters $(\hat{\beta}, \hat{\xi})$ as

$$\sqrt{n}Q^{1/2}\begin{pmatrix}\hat{\beta}_n - \beta \\ \hat{\xi}_n - \xi\end{pmatrix} \to_d \mathcal{N}(0, I), \quad \text{as } n \to \infty, \tag{6}$$

where

$$Q^{-1} = (1 + \xi)\begin{pmatrix}2\beta^2 & -\beta \\ -\beta & 1 + \xi\end{pmatrix}, \tag{7}$$

under the assumtion that $\sqrt{n}u_n^{-\delta}L(u_n) \to 0$ and that $x^{-\delta}L(x)$ is non-increasing. Especially, this condition is met if $u_n = n^{\alpha\xi}$ with $\alpha \in (1/2\delta\xi, \infty)$.

The distribution of $\hat{\mu}$ is given by the following lemma:

**Lemma A.1 (Distribution of $\hat{\mu}$)**
Let $X_1, X_2, \ldots$ be positive iid random variables with df $F$ given by (1). Further, let $\mu_n = \mathsf{E}[X_1 \mathbf{1}_{\{X_1 \le u_n\}}]$ and $\gamma_n^2 = \mathsf{Var}(X_1 \mathbf{1}_{\{X_1 \le u_n\}})$, with $u_n = O_+(n^{\alpha\xi})$ for some $\alpha \in (0, 1)$ and $\xi \in (0, 1)$, where $\delta > 0$. Then

$$\frac{\sqrt{n}}{\gamma_n}(\hat{\mu} - \mu_n) \to_d \mathcal{N}(0, 1), \quad \text{as } n \to \infty,$$

or, equivalently,

$$\mathsf{E}[\exp\{it\frac{\sqrt{n}}{\gamma_n}(\hat{\mu} - \mu_n)\}] \to e^{-t^2/2}, \quad \text{as } n \to \infty,$$

where $\hat{\mu}$ is defined in Equation (2).

Proof: The proof is straightforward in that all that has to be done is to verify that the Lindeberg-Feller Theorem (see e.g. Chapter 2 in [3]) is applicable. To this end, note that

$$\frac{\sqrt{n}}{\gamma_n}(\hat{\mu} - \mu_n) = \sum_{k=1}^n \frac{X_k \mathbf{1}_{\{X_k \le u_n\}} - \mu_n}{\gamma_n \sqrt{n}} =: \sum_{k=1}^n Y_{k,n},$$

where $\mathsf{E}[Y_{k,n}] = 0$, $\mathsf{E}[Y_{k,n}^2] = 1/n$ and hence $\sum_{k=1}^n \mathsf{E}[Y_{k,n}^2] = 1$ for all $n \ge 1$.

Further, by Proposition A.1,

$$\mu_n \pm \epsilon\gamma_n\sqrt{n} = \pm n^{\alpha(\xi-1/2)+1/2}O_+(1) \to \pm\infty, \quad \text{as } n \to \infty$$

for all $\alpha \in (0,1)$, $\delta > 0$, $\xi \in (0,1)$ and $\epsilon > 0$, where $u_n = O_+(n^{\alpha\xi})$ was used. This means that

$$\begin{aligned}
\sum_{k=1}^{n} \mathsf{E}[|Y_{k,n}|^2; |Y_{k,n}| > \epsilon] &= \frac{1}{\gamma_n^2}\mathsf{E}[(X_1\mathbf{1}_{\{X_1\le u_n\}} - \mu_n)^2; |X_1\mathbf{1}_{\{X_1\le u_n\}} - \mu_n| > \epsilon\gamma_n\sqrt{n}] \\
&\le \frac{u_n^2}{\gamma_n^2}\mathsf{P}(|X_1\mathbf{1}_{\{X_1\le u_n\}} - \mu_n| > \epsilon\gamma_n\sqrt{n}) \\
&= \frac{u_n^2}{\gamma_n^2}\Big(\mathsf{P}(X_1\mathbf{1}_{\{X_1\le u_n\}} > \mu_n + \epsilon\gamma_n\sqrt{n}) \\
&\quad + \mathsf{P}(X_1\mathbf{1}_{\{X_1\le u_n\}} < \mu_n - \epsilon\gamma_n\sqrt{n})\Big) \\
&= \frac{u_n^2}{\gamma_n^2}(\bar{F}(\mu_n + \epsilon\gamma_n\sqrt{n}) + F(\mu_n - \epsilon\gamma_n\sqrt{n})) \to 0, \quad \text{as } n \to \infty,
\end{aligned}$$

since $F(x) = 0$ for $x < 0$, $\bar{F}(x) = x^{-1/\xi}O_+(1)$ and $u_n^2/\gamma_n^2 = n^\alpha O_+(1)$. The convergence holds for all $\epsilon > 0$, $\alpha \in (0,1)$, $\delta > 0$ and $\xi \in (0,1)$. The result now follows from the Lindeberg-Feller Theorem. □

Now examine the joint distribution of the parameters.

**Lemma A.2 (Joint distribution)**
Let $X_1, X_2, \ldots$ be positive iid random variables with distribution function $F$ as in (1) and $N = |\{X_i : X_i > u_n\}| \sim \text{Bin}(n, p_n)$, where $p_n = \bar{F}(u_n)$ and $u_n = O_+(n^{\alpha\xi})$ for some $\alpha \in (1/2\delta\xi, 1)$, $\delta > 1/2$ and $\xi \in (0,1)$. Then

$$\begin{aligned}
\phi(t_1,t_2,t_3,t_4) &= \mathsf{E}[\exp\{it_1\frac{\sqrt{n}}{\gamma_n}(\hat{\mu} - \mu_n) + i\sqrt{np_n}Q^{1/2}(t_2,t_3)\begin{pmatrix}\hat{\beta}_N - \beta \\ \hat{\xi}_N - \xi\end{pmatrix} + it_4\frac{\sqrt{n}(\hat{p} - p_n)}{\sqrt{p_n(1-p_n)}}\}] \\
&\to \exp\{-\frac{t_1^2}{2} - \frac{1}{2}(t_2,t_3)\begin{pmatrix}t_2 \\ t_3\end{pmatrix} - \frac{t_4^2}{2}\} \quad \text{as } n \to \infty,
\end{aligned}$$

where $Q$ is given by (7), $\mu_n = \mathsf{E}[X_1\mathbf{1}_{\{X_1\le u_n\}}]$, $\gamma_n^2 = \mathsf{Var}(X_1\mathbf{1}_{\{X_1\le u_n\}})$ and $\hat{p} = N/n$.

Proof: Let

$$\begin{aligned}
\phi_{\mu|N}(t_1) &= \mathsf{E}[\exp\{it_1\frac{\sqrt{n}}{\gamma_n}(\hat{\mu} - \mu_n)\}|N] \\
\phi_{\beta,\xi|N}(t_2,t_3) &= \mathsf{E}[\exp\{i\sqrt{np_n}Q^{1/2}(t_2,t_3)\begin{pmatrix}\hat{\beta}_N - \beta \\ \hat{\xi}_N - \xi\end{pmatrix}\}|N] \\
\phi_{p|N}(t_4) &= \mathsf{E}[\exp\{it_4\frac{\sqrt{n}(\hat{p} - p_n)}{\sqrt{p_n(1-p_n)}}\}|N] = \exp\{it_4\frac{\sqrt{n}(\hat{p} - p_n)}{\sqrt{p_n(1-p_n)}}\},
\end{aligned}$$

then $\phi(t_1,t_2,t_3,t_4) = \mathsf{E}[\phi_{\mu|N}(t_1)\phi_{\beta,\xi|N}(t_2,t_3)\phi_{p|N}(t_4)]$. Note that, conditional on $N$, $\hat{\mu}$ is independent of $(\hat{\beta}_N, \hat{\xi}_N)$.

If $\{\phi_n\}_{n=1}^{\infty}$ is a sequence of characteristic functions such that $\phi_n(t) \to \phi(t)$, then there is a constant $n_0$ such that, for each $n > n_0$ there is a $\rho > 0$ such that $|\phi_n(t) - \phi(t)| < \rho$. This means that

$$\begin{aligned}\mathsf{P}(|\phi_N(t) - \phi(t)| < \rho) &= \mathsf{P}(|\phi_N(t) - \phi(t)| < \rho, N > n_0) \\ &\quad + \mathsf{P}(|\phi_N(t) - \phi(t)| < \rho, N \leq n_0) \\ &= \mathsf{P}(N > n_0) + \mathsf{P}(|\phi_N(t) - \phi(t)| < \rho, N \leq n_0) \to 1, \quad \text{as } n \to \infty.\end{aligned}$$

Using Equation 6 and Lemma A.1 together with the above property, means that

$$\begin{aligned}\phi_{\mu|N}(t_1) &\to_p \exp\{-\frac{t_1^2}{2}\} \quad \text{and} \\ \phi_{\beta,\xi|N}(t_2, t_3) &\to_p \exp\{-\frac{1}{2}(t_2, t_3)\begin{pmatrix} t_2 \\ t_3 \end{pmatrix}\},\end{aligned}$$

Naturally, the product $\phi_{\mu|N}(t_1)\phi_{\beta,\xi|N}(t_2, t_3)$ will also converge in probability.

Further, since $N \sim \text{Bin}(n, p_n)$, $\hat{p} = N/n$ and $np_n = n^{1-\alpha}O_+(1)$, it follows from the Lindeberg-Feller Theorem that

$$\frac{\sqrt{n}(\hat{p} - p_n)}{\sqrt{p_n(1-p_n)}} \to_d \mathcal{N}(0,1) \quad \text{and} \quad \phi_{p|N}(t_4) \to_d \exp\{it_4\mathcal{N}(0,1)\}.$$

Finally, then

$$\phi_{\mu|N}(t_1)\phi_{\beta,\xi|N}(t_2, t_3)\phi_{p|N}(t_4) \to_d \exp\{-\frac{t_1^2}{2}\}\exp\{-\frac{1}{2}(t_2, t_3)\begin{pmatrix} t_2 \\ t_3 \end{pmatrix}\}\exp\{it_4\mathcal{N}(0,1)\},$$

and the claim follows by taking expectations. □

All the tools are now in place for proving Theorem 2.1

Proof: Equation (5) states that

$$\hat{M} - M = \hat{\mu} - \mu_n + \hat{\tau} - \tau,$$

where

$$\begin{aligned}\tau &= \int_{u_n}^{\infty} y dF(y) \\ &= \left\{\bar{F}(u_n + y) = \bar{F}(u_n)\bar{F}_u(y) = p_n\bar{F}_u(y)\right\} \\ &= p_n u_n + p_n \int_0^{\infty} \bar{F}_u(y)dy \\ &= p_n u_n + \frac{p_n u_n^{1/\xi}}{1 + u_n^{-\delta}L(u_n)} \int_{u_n}^{\infty} y^{-1/\xi} + y^{-1/\xi-\delta}L(y)dy \\ &\sim \left\{\text{Karamata's theorem and using } \beta = \xi u_n\right\} \\ &\sim p_n\big(u_n + \frac{\beta}{1-\xi}\big) + R_1,\end{aligned}$$

where $a_n \sim b_n$ means that $a_n/b_n \to 1$ and where

$$R_1 = \frac{p_n\beta}{1-\xi}\frac{\delta\xi}{1-\xi+\delta\xi}\frac{u_n^{-\delta}L(u_n)}{1+u_n^{-\delta}L(u_n)}.$$

Karamata's theorem requires $L(x)$ to be locally bounded in $[x_0,\infty)$ for some $x_0 \geq 0$. Using Taylor expansion, we find that

$$\begin{aligned}\hat{M} - M &\sim \hat{\mu} - \mu_n + (\hat{p}-p_n)\big(u_n + \frac{\hat{\beta}_N}{1-\hat{\xi}_N}\big) + p_n\big(\frac{\hat{\beta}_N}{1-\hat{\xi}_N} - \frac{\beta}{1-\xi}\big) + R_1 \\ &= \hat{\mu} - \mu_n + (\hat{p}-p_n)\big(u_n + \frac{\beta}{1-\xi} - \frac{\beta}{(1-\xi)^2}(\hat{\xi}_N-\xi) + R_2\big) - \\ &- p_n\frac{\beta}{(1-\xi)^2}(\hat{\xi}_N-\xi) + p_nR_2 + R_1,\end{aligned}$$

where

$$R_2 = \beta\sum_{k=2}^{\infty}\frac{(-1)^k(\hat{\xi}_N-\xi)^k}{(1-\xi)^{k+1}} + \sum_{k=1}^{\infty}\frac{(-1)^k}{(1-\xi)^k}(\hat{\beta}_N-\beta)(\hat{\xi}_N-\xi)^{k-1}.$$

Muliplying by $\sqrt{n}/\gamma_n$ and using Lemma A.2 together with the Continuous Mapping Theorem, we find that

$$\begin{aligned}\frac{\sqrt{n}}{\gamma_n}(\hat{M}-M) \sim \underbrace{\frac{\sqrt{n}}{\gamma_n}(\hat{\mu}-\mu_n)}_{\to_d\mathcal{N}(0,1)} + \frac{\sqrt{p_n(1-p_n)}}{\gamma_n}\Big(u_n+\frac{\beta}{1-\xi}\Big)\underbrace{\frac{\sqrt{n}}{\sqrt{p_n(1-p_n)}}(\hat{p}-p_n)}_{\to_d\mathcal{N}(0,1)} - \\ - \frac{\sqrt{p_n}\beta}{\gamma_n(1-\xi)^2}\underbrace{\sqrt{np_n}(\hat{\xi}_N-\xi)}_{\to_d\mathcal{N}(0,(1+\xi)^2)} + o_p(1).\end{aligned}$$

The sequence

$$k_n = 1 + \frac{p_n(1-p_n)}{\gamma_n^2}\Big(u_n+\frac{\beta}{1-\xi}\Big)^2 + \frac{p_n\beta^2}{\gamma_n^2}\frac{(1+\xi)^2}{(1-\xi)^4} = O_+(1)$$

by Proposition A.1

□

## References


[1] N.H. Bingham, C.M. Goldie, and J.L. Teugels. *Regular Variation*. Number 27 in Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1987.

[2] M. E. Crovella, M. S. Taqqu, and A. Bestavros. Heavy-Tailed Probability Distributions in the World Wide Web. In R. J. Adler, R. E. Feldman, and M. S. Taqqu, editors, *A Practical Guide to Heavy Tails*, pages 3–25. Birkhäuser, 1998.

[3] R Durrett. *Probability: Theory and Examples, Second Edition*. Duxbury Press, 1996.

[4] Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling Extremal Events*. Springer, 1997.

[5] C. A. Guerin, H. Nyberg, O. Perrin, S. Resnick, H. Rootzén, and C. Starica. Empirical Testing of the Infinite Source Poisson Data Traffic Model. Preprint, Chalmers University of Technology, Sweden, no 2000:4.

[6] F. P. Kelly, S. Zachary, and I. Ziedins, editors. *Stochastic Networks, Theory and Applications*, volume 4 of *Royal Statistical Society Lecture Note Series*. Oxford University Press, 1996.

[7] J. B. Levy and M. S. Taqqu. Renewal Reward Processes with Heavy-Tailed Inter-Renewal Times and Heavy-Tailed Rewards. *Bernoulli*, 6(1):23–44, 2000.

[8] I. Norros. On the Use of Fractional Brownian Motion in the Theory of Connectionless Networks. *IEEE Journal on Selected Areas in Communications*, 13(6):953–962, August 1995.

[9] V. Pipiras and M. S. Taqqu. The Limit of a Renewal Reward Process with Heavy-Tailed Rewards is not a Linear Fractional Stable Motion. *Bernoulli*, 6(4):607–614, 2000.

[10] S.I. Resnick. *Extreme Values, Regular Variation and Point Processes*. Springer-Verlag, 1987.

[11] J. P. Romano and M. Wolf. Subsampling inference for the mean in the heavy-tailed case. *Metrika*, 50:55–69, 1999.

[12] E. Seneta. *Regularly Varying Functions*. Number 508 in Lecture Notes in Mathematics. Springer-Verlag, 1976.

[13] R.L. Smith. Estimating Tails of Probability Distributions. *The Annals of Statistics*, 15(3):1174–1207, 1987.

[14] N. Tajvidi. *Characterisation and Some Statistical Aspects of Univariate and Multivariate Generalised Pareto Distributions*. PhD thesis, Chalmers University of Technology, 1996.

# Paper B

# Estimating the mean of heavy-tailed distributions in the presence of dependence

Joachim Johansson*

May 6, 2003

**Abstract**

An asymptotically normally distributed estimate of the expected value for heavy-tailed $m$-dependent random variables is suggested and its behaviour relative to estimation using the sample mean is investigated. It is also shown how covariances can be estimated using the same technique, making the method suitable as a diagnostic tool for fitting the order of ARMA-processes. In a small simulation study, the suggested estimate exhibits smaller median bias compared to standard methodologies.

Keywords: Pareto distribution, mean estimation, covariance estimation, heavy tailed distributions, $m$-dependence

MSC2000 classification: primary – 62G32
secondary – 62F10, 62F12, 62P30

## 1 Background

When modelling phenomena in as diverse fields as telecommunications, finance, insurance and hydrology, heavy-tailed distributions are sometimes encountered. The term heavy-tailed means a distribution with some infinite moments. In telecommunications it has been observed that the distribution of file sizes on the Internet has this property, see Crovella et al. (1998). In insurance, heavy tails are encountered in models for fire and storm damages, see Embrechts et al. (1997) and Tajvidi (1996). And in finance, the so-called log-returns exhibit similar behaviour (Diebold et al., 1997; Longin, 2000). The list could be made much longer. For an introduction to the field of extreme value modelling, see Embrechts et al. (1997), Reiss and Thomas (2001) or Coles (2001).

*Chalmers University of Technology, Göteborg, Sweden. E-mail: joachimj@math.chalmers.se

Much emphasis has been placed on estimating quantiles in heavy-tailed distributions, which is of paramount importance for many risk-assessment applications. See Embrechts et al. (1997) for an overview of this area. Such quantile estimation has also been examined in the presence of dependence, see for instance Drees (2000). Estimating the mean of heavy-tailed distributions has not enjoyed the same attention, however.

When the estimate of the mean is the sample mean $\bar{X}_n = n^{-1}(X_1 + \ldots + X_n)$ of some random variables $X_1, \ldots, X_n$ with infinite second moment, there are two main methods available. The first is based on the well-known fact that $\bar{X}_n$, properly normalised, tends to a stable distribution. If the $X_k$'s are themselves stable, estimation of the parameters of this limiting distribution can be made, see Nolan (1999, 2001).

In case the observations $X_k$ are not stable, one alternative might be to base estimation on $\bar{X}_{k,j} = k^{-1}(X_{kj} + \ldots + X_{k(j+1)-1})$, where the $\bar{X}_k$ will be approximately stable. The difficulty here is to select the block-size $k$. No (non-bootstrap) results seem to be available for this case though. However, see Crovella and Taqqu (1999) where such an approach is used for estimating the tail index from scaling properties. There are special cases where at least some parameters of the resulting limiting distribution can be estimated; if the observations are bounded from below, the limiting distribution will also be totally skewed to the right, and so on. But more work seems to be needed for creating a unified estimation framework using this approach.

The second main possibility is to use resampling methods. In the infinite variance case, Efron's original bootstrap for the sample mean does not work (Athreya, 1987). However, sub-sampling methods, based on Politis and Romano (1994), are available and do generate consistent estimators. See also Romano and Wolf (1999) and Politis et al. (1999), where the latter gives an introduction to the area of sub-sampling.

A drawback with the subsampling method is that, even if the subsample size is chosen optimally, the error between the subsample bootstrap and the true distribution will often be an order of magnitude larger than for an asymptotic distribution, see Hall and Jing (1998). More specifically, if $G_n$ is the distribution of $n^{-\xi}\sum_{k=1}^{n}(X_k - \mu)$ where $1/\xi$ is the tail index and $\mu = \mathsf{E}[X_1]$, then the $G_n$ tends to some limit distribution $H$. The order of $G_n - H(\hat{\xi})$ will then be smaller than the order of $\hat{G}_m - G_n$, where $\hat{G}_m$ is the subsample bootstrap estimate of $G_n$ with optimal subsample size $m = m(n)$.

A third, new approach, based on extreme value theory was recently independently proposed by Peng (2001) and Johansson (2001). The idea is to fit a parametric model for the tails but use a non-parametric estimate for the distribution centre. Peng (2001) worked under a tail-balance condition and used a one-parameter model for the tail. Johansson (2001) assumed positive r.v.'s and fitted a two-parameter

model. The latter method can easily be extended to allow for distributions with unbounded support as well, and there is no need for tail-balance in that case. Both approaches assume mutually independent observations.

The aim of this paper is to extend the results of Johansson (2001) to include $m$-dependence. There are several reasons why this kind of structure is interesting. One of the standard diagnostics tests for data is to plot sample correlations. If the data seems to be dependent then $m$-dependence may serve as preliminary model. In fact, even if the r.v.'s $\{X_k\}$ are iid, the products $\{X_k X_{k+m}\}$ will be $m$-dependent. Estimation of $\mathsf{E}[X_1 X_{1+m}]$ may be used as an alternative diagnostics tool for fitting ARMA-models to data, see Brockwell and Davis (1987). A third reason is that the actual data-generating mechanism may suggest this type of model.

Section 2 of the paper introduces the estimate and examines its statistical properties. The estimate is then compared to the standard methods for estimating the mean and the covariance in Section 3. In Section 4 some concluding remarks are made and the paper ends with an appendix where most proofs have been gathered.

## 2 A semi-parametric estimate of the mean

In this section we propose a semi-parametric estimate of the mean of $m$-dependent, heavy-tailed random variables. First we introduce a model for the underlying distribution. Let $X_1, X_2, \ldots, X_n$ be stationary $m$-dependent positive r.v.'s with distribution function $F$, where

$$\bar{F}(x) := 1 - F(x) = cx^{-1/\xi}(1 + x^{-\delta}L(x)), \tag{1}$$

for $\xi \in (0,1)$, $\delta > 0$ and some constant $c$. $L$ is a slowly varying function, i.e $L(tx)/L(x) \to 1$ as $x \to \infty$ for all $t > 0$. See Bingham et al. (1987) for further properties of these functions.

The model in (1) is justified by the fact that the tail of many heavy-tailed distributions can be approximated by a series expansion. The use of the slowly varying function $L$ makes this even more general in that it captures the behaviour of the remainder term of such an expansion. We assume that $F$ is one sided purely for notational convenience. The same arguments that are made below are easily extended to the case where the distribution function has unbounded support. Our main interest is in the case $\xi \in (1/2, 1)$ for which the variance is infinite, but the mean is finite. However, the method suggested below will work also in the finite variance case where $\xi \in (0, 1/2)$.

For motivating the alternative estimate, first look at the standard estimate of $\mathsf{E}[X]$, namely the sample mean, which can be written as

$$\bar{X} = \int x dF_n(x) = \frac{1}{n}\sum_{k=1}^{n} X_k,$$

where $F_n$ is the empirical distribution function. Building on this we propose an estimate of the form

$$\hat{\mathsf{E}}[X] := \hat{M} := \hat{\mu} + \hat{\tau} := \int_0^{u_n} x dF_n(x) + \int_{u_n}^{\infty} x d\hat{F}(x),$$

where $\hat{\tau}$ is the part of $\hat{M}$ originating from the tail of the distribution. The tail is assumed to start at some level $u_n$, which in the asymptotic analysis will be assumed to tend to infinity. $\hat{F}$ is an estimate of the tail distribution function, as described below.

It is desirable to find a model for the tail. To do this, follow Pickands (1975) and let $F_u(y) = \mathsf{P}(X - u_n \leq y | X > u_n)$ be the distribution of the excesses over the threshold $u_n$. It follows from (1) that

$$\bar{F}_u(y) = \frac{\bar{F}(u_n + y)}{\bar{F}(u_n)} = \left(1 + \frac{y}{u_n}\right)^{-1/\xi} \frac{1 + (u_n + y)^{-\delta} L(u_n + y)}{1 + u_n^{-\delta} L(u_n)}, \tag{2}$$

and if $\beta_n = \beta(u_n) = u_n \xi$, then $\bar{F}_u(y)$ is a perturbed generalised Pareto distribution (GPD), where the df of the GPD has the form

$$G_{\beta,\xi}(x) = \begin{cases} 1 - \left(1 + \xi \frac{x}{\beta}\right)^{-1/\xi}, & \xi \neq 0 \\ 1 - e^{-x/\beta}, & \xi = 0 \end{cases}, \quad x \in \begin{cases} [0, \infty), & \xi \geq 0 \\ [0, -\beta/\xi], & \xi < 0 \end{cases}.$$

This means that for large values of $u_n$, $F_u(y) \approx G_{\beta(u_n),\xi}(y)$ in the sense that

$$\lim_{u_n \uparrow y_F} \sup_{0 < y < y_F - u_n} |F_u(y) - G_{\beta(u_n),\xi}(y)| = 0,$$

where $y_F$ is the right end point of $F$ and $\beta$ is some positive function, see Theorem 3.4.13 in Embrechts et al. (1997).

By definition $\bar{F}(u_n + y) = \bar{F}(u_n)\bar{F}_u(y)$. And, for $N = N_n = |\{i : X_i > u_n\}|$, the number of $X_i$'s which exceed $u_n$, we estimate the probability $p_n = \bar{F}(u_n)$ by

$$\hat{p} = \widehat{\bar{F}}(u_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i > u_n\}} = \frac{N}{n}.$$

For large values of $u_n$, use

$$\widehat{\bar{F}}_u(y) = \bar{G}_{\hat{\beta},\hat{\xi}}(y),$$

for appropriate estimates $\hat{\xi} = \hat{\xi}_n$ and $\hat{\beta}_n = \hat{\beta}(u_n)$. Note that $\beta$ will be estimated separately, i.e. $\beta = \xi u_n$ will not be used. The reason for this is to achieve greater flexibility in the parameter fitting, compensating for the underlying distribution not being an exact GPD.

We have now arrived at an alternative estimate, $\hat{M}$, of $\mathsf{E}[X]$,

$$\begin{aligned}\hat{M} &:= \int_0^{u_n} x dF_n(x) + \int_{u_n}^{\infty} x d\hat{F}(x) \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i \mathbf{1}_{\{X_i \le u_n\}} + \int_{u_n}^{\infty} x \frac{N}{n\hat{\beta}}\Big(1+\hat{\xi}\frac{x-u_n}{\hat{\beta}}\Big)^{-1-1/\hat{\xi}} dx \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i \mathbf{1}_{\{X_i \le u_n\}} + \hat{p}\Big(u_n + \frac{\hat{\beta}}{1-\hat{\xi}}\Big), \text{ for } \hat{\xi} \in (0,1),\end{aligned}$$

where $\hat{p} = N/n$. It will be seen later that, if $\xi \in (0,1)$ then $\mathsf{P}(\hat{\xi}_n \in (0,1)) \to 1$, so the convergence of the integral will asymptotically not be a problem.

If $\hat{\xi} \ge 1$, then $\hat{M}$ should be set to $\infty$ since this would indicate that the first moment of the distribution in (1) is infinite.

If the precise nature of the dependence between the random variables $X_k$ is known, it might be possible to use the maximum likelihood estimates of $\beta$ and $\xi$. Naturally, this would be the preferable situation, but in practise this may be impossible.

Instead $\beta$ and $\xi$ will be estimated in the following way. Let the excesses over the threshold $u_n$ be $\{Y_k\}$ and assume that $Y_k \sim F_u$, where $F_u$ is given in (2). Further, let the pseudo-loglikelihood function be

$$l(\beta,\xi) = \sum_{k=1}^{n} \ln\Big(\frac{1}{\beta}\Big(1+\xi\frac{X_k-u_n}{\beta}\Big)^{-1-1/\xi}\Big)\mathbf{1}_{\{X_k>u_n\}}, \tag{3}$$

i.e. the likelihood function we would get if the exceedances over the threshold were independent. The estimates will then be $(\hat{\beta},\hat{\xi}) = \arg\max l(\beta,\xi)$.

## 2.1 Properties of the estimator

We now examine the properties of the proposed estimator and how the tail-parameter estimates $\hat{\beta}$ and $\hat{\xi}$ are distributed. After that, we state a few lemmas that will be needed in the process of calculating the asymptotic distribution of $\hat{M}$, which is given in Theorem 2.1.

The log-likelihood function in (3) gives us the score function (as in Smith (1987))

$$U(\beta_n,\xi_n) = \begin{bmatrix} U_{\beta_n} \\ U_{\xi_n}\end{bmatrix} = \sum_{k=1}^{n} \begin{bmatrix} \frac{1}{\beta\xi} - \frac{1}{\beta}\Big(\frac{1}{\xi}+1\Big)\Big(1+\xi\frac{X_k-u_n}{\beta}\Big)^{-1} \\ \frac{1}{\xi^2}\ln\Big(1+\xi\frac{X_k-u_n}{\beta}\Big) - \frac{\xi+1}{\xi^2}\Big(1-\Big(1+\xi\frac{X_k-u_n}{\beta}\Big)^{-1}\Big)\end{bmatrix}\mathbf{1}_{\{X_k>u_n\}} \tag{4}$$

and the information matrix

$$J_n = \begin{bmatrix} -\frac{\partial^2 l}{\partial\beta^2} & -\frac{\partial^2 l}{\partial\beta\partial\xi} \\ -\frac{\partial^2 l}{\partial\beta\partial\xi} & -\frac{\partial^2 l}{\partial\xi^2}\end{bmatrix} = \begin{bmatrix} J_{\beta\beta} & J_{\beta\xi} \\ J_{\beta\xi} & J_{\xi\xi}\end{bmatrix},$$

where

$$
\begin{aligned}
J_{\beta\beta} &= \sum_{k=1}^{n} \Big\{ \frac{1}{\beta^2\xi} - \frac{1+\xi}{\beta^2\xi}\Big(1+\xi\frac{X_k-u_n}{\beta}\Big)^{-2} \Big\} \mathbf{1}_{\{X_k>u_n\}} \\
J_{\beta\xi} &= \sum_{k=1}^{n} \Big\{ \frac{1}{\beta\xi^2} + \frac{1}{\beta\xi}\left(1+\frac{1}{\xi}\right)\Big(1+\xi\frac{X_k-u_n}{\beta}\Big)^{-2} \\
&\qquad - \frac{1}{\beta\xi}\left(1+\frac{2}{\xi}\right)\Big(1+\xi\frac{X_k-u_n}{\beta}\Big)^{-1} \Big\} \mathbf{1}_{\{X_k>u_n\}} \qquad (5) \\
J_{\xi\xi} &= \sum_{k=1}^{n} \Big\{ \frac{2}{\xi^2}\left(1+\frac{2}{\xi}\right)\Big(1+\xi\frac{X_k-u_n}{\beta}\Big)^{-1} + \frac{2}{\xi^3}\ln\Big(1+\xi\frac{X_k-u_n}{\beta}\Big) \\
&\qquad - \frac{1}{\xi^2}\left(1+\frac{1}{\xi}\right)\Big(1+\xi\frac{X_k-u_n}{\beta}\Big)^{-2} - \frac{1}{\xi^2}\left(1+\frac{3}{\xi}\right) \Big\} \mathbf{1}_{\{X_k>u_n\}}.
\end{aligned}
$$

The proof now proceeds with first looking at the distribution of the tail-parameters $(\hat{\beta}_n, \hat{\xi}_n)$. Then the other two parameters $\hat{\mu}$ and $\hat{p}$ are investigated. Finally, the joint distribution of all parameters will be needed for arriving at the main result in Theorem 2.1. But first a note about asymptotics.

The proposed model in (1) specifies the tail behaviour for the distribution of the observations $X_k$. Since the tail behaviour, by definition, is an asymptotic property of $F$ we want $p_n = \bar{F}(u_n) \propto u_n^{-1/\xi} \to 0$ (where $a_n \propto b_n$ means that $a_n/b_n \to$ constant). Since $\xi \in (0,1)$ this means that $u_n \to \infty$, and for ease of analysis we specifically assume $u_n \propto n^\theta$ for some $\theta > 0$. This restriction means that we let the tail speak for itself and not be influenced by the bulk of the distribution.

Further, we want $np_n \to \infty$, i.e. we want the number of observations in the tail to grow as the sample grows. This is equivalent to requiring $n^{1-\theta/\xi} \to \infty$ or $\theta < \xi$. This is achieved by selecting $u_n = O_+(n^{\alpha\xi})$ for some $\alpha \in (0,1)$, where $a_n = O_+(n^\alpha)$ denotes a positive sequence such that $a_n/n^\alpha$ is bounded away from zero and infinity.

This leads to the following assumption, which will be assumed to hold throughout the rest of the paper.

**Assumption 2.1 (Basic assumptions)**
The random variables $X_1, \ldots, X_n$ are stationary and $m$-dependent with distribution function $F$ as in (1). Further, the threshold $u_n = O_+(n^{\alpha\xi})$ for some $\alpha \in (0,1)$. □

Further restrictions will have to be placed on $\alpha$ later on, which will be apparent from the proofs.

In order to find the asymptotic distribution of $(\beta_n, \xi_n)$, we need the following lemma, where $a_n \sim b_n$ means that $a_n/b_n \to 1$ as $n$ (and hence also $u_n$ and $\beta_n$) tend to infinity.

**Lemma 2.1 (Expected values)**
Let $Y \sim F_u$, where $F_u(y) = \mathsf{P}(Y \leq y) = \mathsf{P}(X_1 - u_n \leq y | X_1 > u_n)$ is given in (2). Then

$$\begin{aligned}
\mathsf{E}\Big[\Big(1+\frac{Y}{u}\Big)^{-r}\Big] &\sim \frac{1}{1+r\xi} + \frac{1}{1+r\xi}\cdot\frac{r\delta\xi^2}{(r+\delta)\xi+1}u^{-\delta}L(u)\\
\mathsf{E}\Big[\ln\Big(1+\frac{Y}{u}\Big)\Big] &\sim \xi - \xi\frac{\delta\xi}{1+\delta\xi}u^{-\delta}L(u)\\
\mathsf{E}\Big[\ln\Big(1+\frac{Y}{u}\Big)^2\Big] &\sim 2\xi^2 - 2\xi^2\frac{\delta\xi(2+\delta\xi)}{(1+\delta\xi)^2}u^{-\delta}L(u)\\
\mathsf{E}\Big[\Big(1+\frac{Y}{u}\Big)^{-r}\ln\Big(1+\frac{Y}{u}\Big)\Big] &\sim \frac{\xi}{(1+r\xi)^2} + \frac{\xi}{(1+r\xi)^2}\cdot\frac{\delta\xi(r^2\xi^2-\delta\xi-1)}{((r+\delta)\xi+1)^2}u^{-\delta}L(u).
\end{aligned}$$

Proof: This follows from straight-forward calculations using Karamata's Theorem (see e.g. Bingham et al. (1987)). See Appendix A.1 for details. □

**Lemma 2.2 (Score function and information matrix)**
Let the log-likelihood function be as in Equation (3) and let $p_n = \mathsf{P}(X_1 > u_n)$. Then, for the score function $U(\beta,\xi)$ defined in Equation (4),

$$\begin{aligned}
\mathsf{E}[U_\beta] &\sim \frac{-\delta\xi}{\beta((1+\delta)\xi+1)}np_nu^{-\delta}L(u)\\
\mathsf{E}[U_\xi] &\sim \frac{-\delta\xi}{(1+\delta\xi)((1+\delta)\xi+1)}np_nu^{-\delta}L(u).
\end{aligned}$$

Further, the information matrix, $J_n$ in (5) satisfies

$$\begin{aligned}
\mathsf{E}[J_{\beta\beta}] &\sim \frac{np_n}{\beta^2(1+2\xi)} + \frac{np_n}{\beta^2(1+2\xi)}\cdot\frac{2\delta\xi(\xi+1)}{(2+\delta)\xi+1}u^{-\delta}L(u)\\
\mathsf{E}[J_{\beta\xi}] &\sim \frac{np_n}{\beta(1+\xi)(1+2\xi)} + \frac{np_n}{\beta(1+\xi)(1+2\xi)}\cdot\frac{\delta\xi(3+(6+\delta)\xi+2\xi^2)}{((1+\delta)\xi+1)((2+\delta)\xi+1)}u^{-\delta}L(u)\\
\mathsf{E}[J_{\xi\xi}] &\sim \frac{2np_n}{(1+\xi)(1+2\xi)}\\
&\quad + \frac{2\delta}{\xi}\Big(\frac{1}{1+\delta\xi} + \frac{\xi+2}{(1+\xi)((\delta+1)\xi+1)} - \frac{\xi+1}{(1+2\xi)(1+(2+\delta)\xi)}\Big)np_nu^{-\delta}L(u).
\end{aligned}$$

Finally, let $C = \mathrm{diag}(\beta_n, 1)$. Then

$$C\frac{1}{np_n}J_nC \to_p G^{-1} = \frac{1}{1+2\xi}\begin{pmatrix} 1 & \frac{1}{1+\xi}\\ \frac{1}{1+\xi} & \frac{2}{1+\xi}\end{pmatrix}, \quad \text{for } n\to\infty.$$

Proof: See Appendix A.2. □

**Lemma 2.3 (Distribution of the score function)**
Assume that $\alpha \in ((1+2\delta\xi)^{-1}, 1)$, where $\xi$ and $\delta$ are parameters in the definition of $F$ in (1). Let $(U_\beta, U_\xi)^t$ be the score function for the GPD defined in (4) and assume that

$$\mathsf{Var}(\beta U_\beta) \to \infty, \quad \mathsf{Var}(U_\xi) \to \infty$$

and that these variances are of the same order. Then

$$\frac{1}{\sqrt{np_n}}\Sigma_n^{-1/2}\begin{pmatrix}\beta U_\beta\\ U_\xi\end{pmatrix} \to_d \mathcal{N}(0,I),$$

where the covariance matrix $\Sigma_n$ is best estimated numerically since it is rather cumbersome and uninformative to write out explicitly. Refer to Section A.3.1 below for a discussion of how this could be done.

Proof: See Appendix A.3. □

**Lemma 2.4 (Distribution of $(\hat{\beta}_n, \hat{\xi}_n)$)**
Assume that $\alpha \in ((1+2\delta\xi)^{-1}, 1)$ and let

$$G = (1+\xi)\begin{pmatrix}2 & -1\\ -1 & 1+\xi\end{pmatrix}.$$

Then

$$\sqrt{np_n}(G\Sigma_n G^t)^{-1/2}\begin{pmatrix}\hat{\beta}_n/\beta - 1\\ \hat{\xi}_n - \xi\end{pmatrix} \to_d \mathcal{N}(0,I),$$

where $\Sigma_n$ is the matrix in Lemma 2.3.

Proof: A sketch of the proof follows below. Taylor expansion of the score function $U$ around the ML-estimate $\hat{\Psi}_n = (\hat{\beta}_n, \hat{\xi}_n)^t$ gives us the result

$$U(\hat{\Psi}_n) \approx U(\Psi) + \frac{\partial U(\Psi)}{\partial \Psi^t}(\hat{\Psi}_n - \Psi),$$

where $U(\hat{\Psi}_n) = 0$ (since $\hat{\Psi}_n$ is the ML-estimate) and $\approx$ means that higher order terms have been neglected. Then

$$\sqrt{np_n}(\hat{\Psi}_n - \Psi) \approx (np_n J_n^{-1})\frac{1}{\sqrt{np_n}}U(\Psi).$$

Let $C = \mathrm{diag}(\beta, 1)$, then

$$\begin{aligned}\sqrt{np_n}\begin{pmatrix}\hat{\beta}_n/\beta - 1\\ \hat{\xi}_n - \xi\end{pmatrix} &\approx C^{-1}np_nJ_n^{-1}\frac{1}{\sqrt{np_n}}\begin{pmatrix}U_\beta\\ U_\xi\end{pmatrix}\\ &= C^{-1}np_nJ_n^{-1}C^{-1}C\frac{1}{\sqrt{np_n}}\begin{pmatrix}U_\beta\\ U_\xi\end{pmatrix}\\ &= C^{-1}np_nJ_n^{-1}C^{-1}\frac{1}{\sqrt{np_n}}\begin{pmatrix}\beta U_\beta\\ U_\xi\end{pmatrix}\\ &\to_d \mathcal{N}(0, G\Sigma G),\end{aligned}$$

by Lemmas 2.2 and 2.3. The proof implicitly assumes that there exists a solution $\hat{\Psi}$ to the maximum likelihood equations, see e.g. Smith (1985, 1987) for a discussion of this. □

Now turn to the distribution of $\hat{\mu}$ and $\hat{p}$.

**Lemma 2.5 (Distribution of $\hat{\mu}$)**
Let $\hat{\mu} = n^{-1}\sum_{k=1}^{n} X_k I_k$, where $I_k = \mathbf{1}_{\{X_k \leq u_n\}}$. Further, let $\mu_n = \mathsf{E}[X_1 I_1]$, $\sigma_\mu^2 = \mathsf{Var}(\hat{\mu})$ and $\gamma_{n,k}^2 = \mathsf{Cov}(X_1 I_1, X_{1+k} I_{1+k})$, $k = 0, 1, \ldots, m$. Then

$$\frac{1}{\sigma_\mu}(\hat{\mu} - \mu_n) \to_d \mathcal{N}(0,1),$$

where we assume that $\sigma_\mu^2 \sim \gamma_{n,0}^2 + 2\gamma_{n,1}^2 + \ldots + 2\gamma_{n,m}^2 \propto n^\rho$, with $\rho > 2\alpha\xi - 1$.

Proof: See Appendix A.4. □

Note that $\sigma_\mu^2 \propto n^{\alpha(2\xi-1)}$ if the $X_k$'s are independent. Hence Lemma 2.5 holds in this case.

**Lemma 2.6 (Distribution of $\hat{p}$)**
Let $\hat{p} = n^{-1}\sum_{k=1}^{n} \mathbf{1}_{\{X_k > u_n\}}$, $p_n = \mathsf{P}(X_1 > u_n)$ and $\sigma_p^2 = \mathsf{Var}(\hat{p})$. Then

$$\frac{1}{\sigma_p}(\hat{p} - p_n) \to_d \mathcal{N}(0,1), \quad \text{as } n \to \infty.$$

Proof: See Appendix A.5. □

In order to be able to prove the main theorem, the joint distribution of the parameters is needed. This is given by the following lemma.

**Lemma 2.7 (Joint distribution)**
Let

$$\Psi_n = \begin{bmatrix} (\hat{\mu} - \mu_n)/\sigma_\mu \\ (\hat{p} - p_n)/\sigma_p \\ \beta U_{\beta_n}/\sqrt{np_n} \\ U_{\xi_n}/\sqrt{np_n} \end{bmatrix} \quad \text{and} \quad F_n = \begin{bmatrix} \sigma_\mu & 0 & 0 & 0 \\ 0 & \sigma_p & 0 & 0 \\ 0 & 0 & 2(1+\xi)/\sqrt{np_n} & -(1+\xi)/\sqrt{np_n} \\ 0 & 0 & -(1+\xi)/\sqrt{np_n} & (1+\xi)^2/\sqrt{np_n} \end{bmatrix}.$$

Assume that $\mathsf{Var}(A\Psi_n) = A\Sigma_n A^t = O_+(n^\tau)$ for some $\tau \geq 0$ and all $A \in \mathbb{R}^4 \setminus \{0\}$. Then

$$(F_n \Sigma_n F_n^t)^{-1/2} \begin{bmatrix} \hat{\mu} - \mu_n \\ \hat{p} - p_n \\ \hat{\beta}_n/\beta - 1 \\ \hat{\xi}_n - \xi \end{bmatrix} \to_d \mathcal{N}(0, I).$$

Proof: See Appendix A.6. □

Finally, let

$$B_n = \left[1; \quad u_n + \beta/(1-\xi); \quad p_n\beta/(1-\xi); \quad p_n\beta/(1-\xi)^2\right]$$

and $\sigma_\mu^2 = \mathsf{Var}(\hat{\mu})$. Then we can formulate the main theorem as follows.

**Theorem 2.1 (Distribution of $\hat{M}$)**
Assume that $\sigma_\mu^2 = O_+(n^{\rho-1})$ with $\rho = \alpha(2\xi - 1)$ and $\alpha \in ((1+2\delta\xi)^{-1}, 1)$. Then, under the assumptions in Lemma 2.7,

$$(B_n K_n B_n^t)^{-1/2}(\hat{M} - M) \to_d \mathcal{N}(0,1)$$

where $K_n = \mathsf{Cov}([\hat{\mu} - \mu, \hat{p} - p_n, \hat{\beta}_n/\beta - 1, \hat{\xi}_n - \xi]^t)$.

Proof: See Appendix A.7. □

# 3 A simulation study

In order to compare the estimate $\hat{M}$ of the expected value to the standard method using the mean $\bar{X}$, some simulations were made. First the mean was estimated based on a sample of independent r.v.'s (a similar simulation was done in Johansson (2001)). The reason for this was to see how the methodology behaved in a standard situation. In the second simulation data was generated by an $MA(3)$-process and the mean was again estimated using $\hat{M}$.

One of the main reasons for contemplating $m$-dependence was that it would open a way to estimate covariances. Ultimately, such a method could be used as a diagnostics tool for determining the order of $ARMA$-processes, see Brockwell and Davis (1987). An example of how this may be done is shown in the last simulation.

The tail-parameters $\beta$ and $\xi$ were estimated using the S-plus program package EVIS, Version 3, by Alexander J McNeil. Depending on the value of $\hat{\xi}$, different estimates of the mean were made, as illustrated in Figure 1. $\hat{\xi} > 1$ would indicate that the mean is infinite, and so the estimate should be infinity in this case. Further, if $\hat{\xi} < 0$ this would mean a distribution with finite tail and then the sample mean was used. Except in degenerate cases, the sample mean tends to a normally distributed random variable also for $m$-dependent variables, provided that the variance is finite.

## 3.1 Simulation based on independent data

In simulation one, 500 samples of independent random variables were drawn from a distribution with pdf as in Figure 2 with parameters $\xi = 0.7$, $w = 10$ and $\delta = 1$. The results are displayed in Figure 3. For the case $w = 10$ the tail starts at approximately $p = \mathsf{P}(X > u) = 0.4$, which explains the bias for $p = 0.45$ and $p = 0.50$. In the

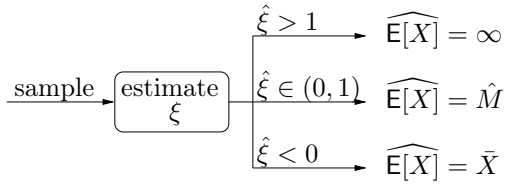


Figure 1: The figure shows how $\hat{M}$ is used in the simulations.

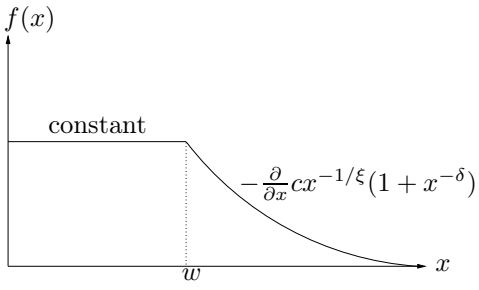


Figure 2: The distribution used in the simulations.

light-tailed case with $\xi = 0.3$ and $w = 10$, the tail starts at $p = 0.23$, with similar implications for the bottom graph in Figure 3. (More simulations of this kind can be found in Johansson (2001)).

There appears to be less of a median bias in the heavy tailed cases using $\hat{M}$ compared to using $\bar{X}$. In the case of light-tailed data, there is not much of a difference between the two methods.

### 3.2 Data from an MA(3)-process

In simulation two, independent samples were fed to the MA(3) process $Y_k = 0.25(X_k + X_{k-1} + X_{k-2} + X_{k-3})$, where the $X_k$ have a distribution as in Figure 2 with parameters $\xi = 0.7$, $\delta = 1$ and $w = 10$. The mean was estimated using $\hat{M}$ and $\bar{X}$ as in simulation one and the results are displayed in Figure 4.

Noticeable in simulation two is the larger median bias for $\hat{M}$ compared to the case with independent observations. The reason for this lies in the fact that the tail behaviour for $Y_k = (X_k + \ldots + X_{k-3})/4$ sets in later than for the $X_k$. This is illustrated in Figure 5 where the pdfs for sums of iid GPD random variables with parameters $\xi = 0.7$ and $\beta = 10$ were plotted. This in turn means that $\hat{\xi}$ was underestimated, but that the bias in the $\xi$ estimates decreased as $p$ decreased/ the threshold $u$ increased.

### 3.3 Covariance estimation

In the case of an MA(1)-process, $Y_k = (X_k + X_{k-1})/2$, the covariance $\mathsf{Cov}(Y_1, Y_2) = \mathsf{Var}(X_1)/4$. In simulation three, iid GPD-distributed innovations $X_k$ with parameters $\xi = 0.35$ and $\beta = 10$ were fed to the above MA-process and the covariance was estimated in the following way:

1. The estimate of $\mathsf{E}[Y_k]$, denoted $\hat{M}_1$, was calculated as described in Figure 1.
2. Then $\mathsf{E}[Y_k Y_{k-1}]$ was estimated by $\hat{M}_2$ in a similar fashion.
3. The estimate of $\mathsf{Cov}(Y_k, Y_{k-1})$ was then set to $\hat{M}_2 - \hat{M}_1^2$.

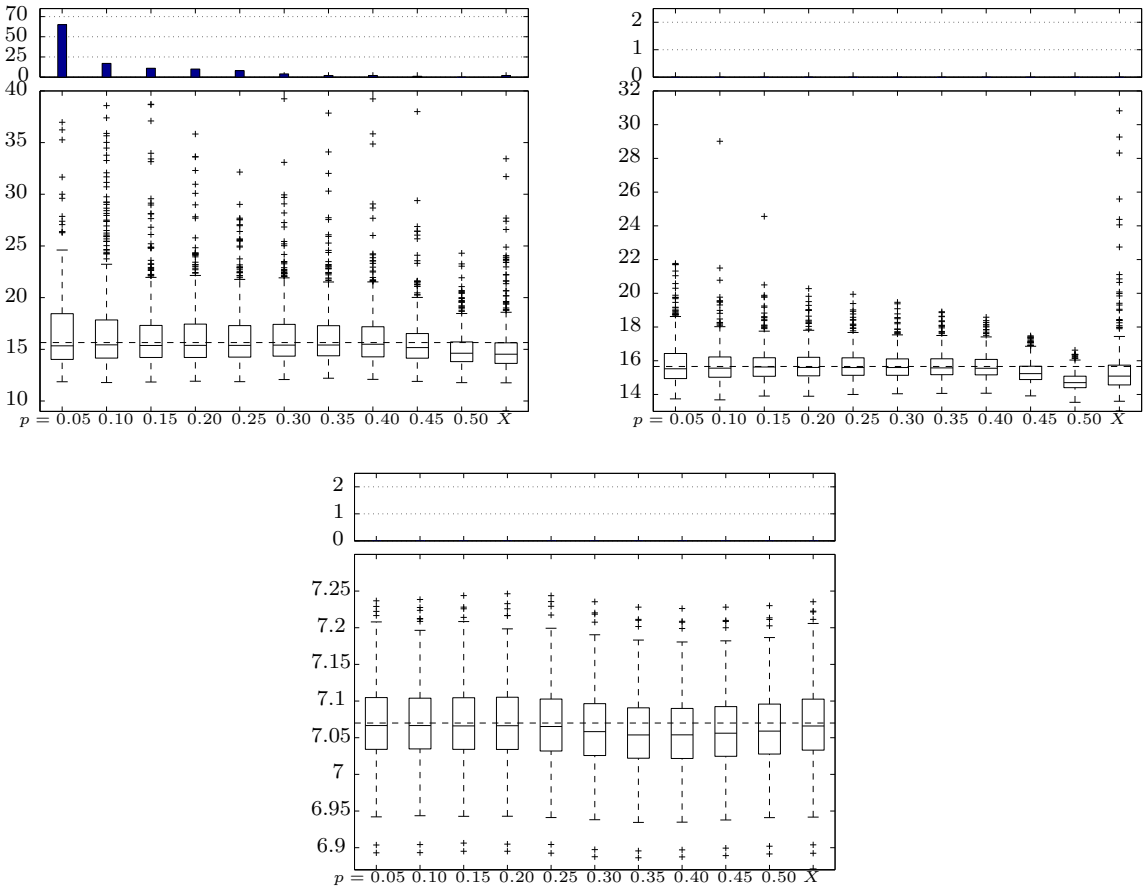


Figure 3: Simulation one, independent r.v.'s. *Top left:* $\xi = 0.7$, $\delta = 1$, $w = 10$ and sample size 1000. *Top right:* Same as top left but with sample size 10000. *Bottom:* $\xi = 0.3$, $\delta = 1$, $w = 10$ and sample size 10000. The bar graphs indicate how many observations were too large to be displayed in the graphs. This includes the cases where the mean was estimated to infinity using $\hat{M}$. The dashed lines indicate the true values of the mean.

This was then compared to the standard method for estimating the covariance

$$\hat{\rho}(1) = \frac{1}{n-1}\sum_{k=1}^{n-1}(X_k - \bar{X})(X_{k+1} - \bar{X}).$$

The results are displayed in Figure 6.

There appears to be a smaller median bias for the estimate $\hat{M}_2 - \hat{M}_1^2$ compared to $\hat{\rho}(1)$. The median bias also decreases as $p$ decreases, that is when we move further out on the tail.

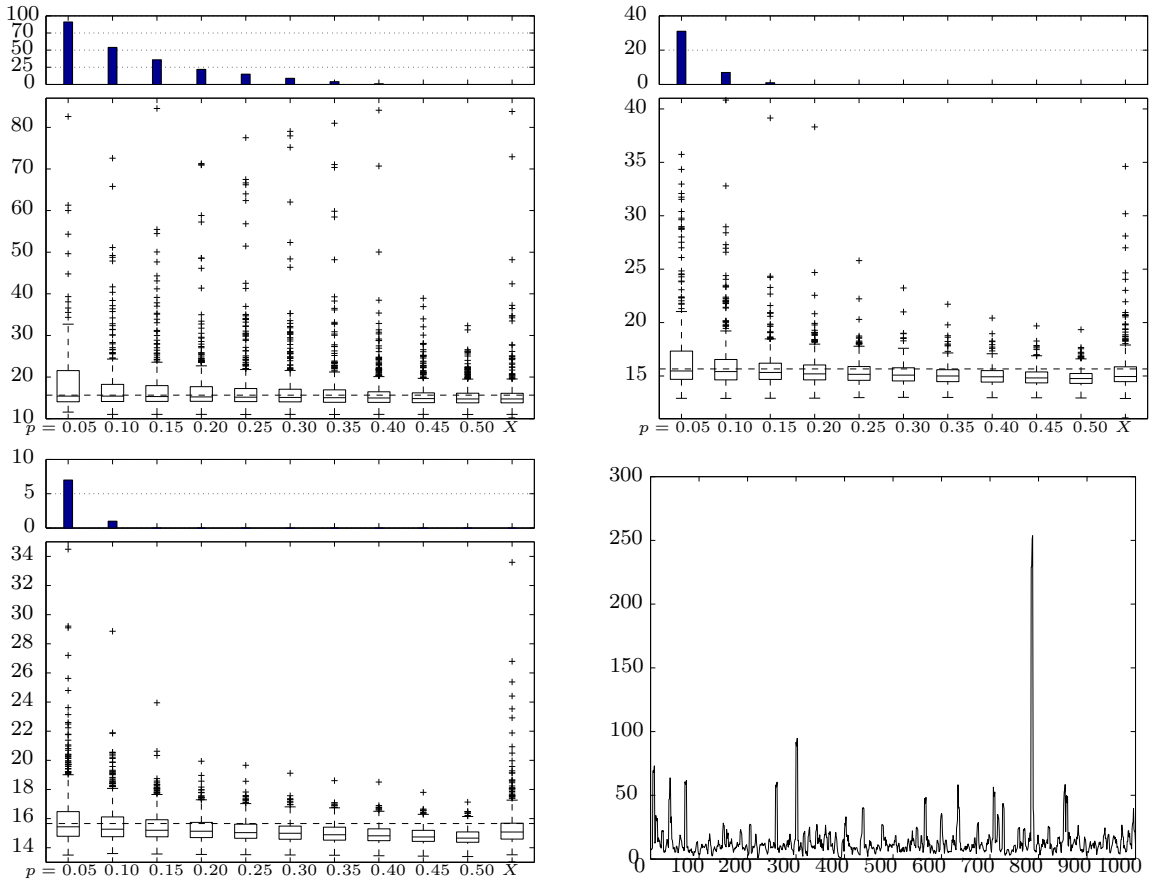


Figure 4: Simulation two, data from an $MA(3)$-process. 500 samples with parameters $\xi = 0.7$, $\delta = 1$ and $w = 10$ and sample sizes *top left:* 1000, *top right:* 5000 and *bottom left:* 10000. The bar graphs show how many observations were too large to be displayed in the boxplots and the dashed lines indicate the true value of the mean. *Bottom right:* a typical sample from the MA(3) process.

# 4 Concluding remarks and suggestions for further study

Judging from the simulations, the suggested method appears useful for sample sizes over 1000. However, it could also be noted that the simulations do not completely mimic how the estimate would be used for a single sample. In the one sample case, the threshold would be set so that good estimates of the tail could be made. In the simulations, no such considerations were made. The probability of exceeding the threshold ($p$) was fixed for all samples, regardless of their individual characteristics.

Finally, note that even though the calculations in the paper were made for the case of $m$-dependence, it should be fairly easy to apply the same techniques of proof for other kinds of dependence. The pseudo maximum likelihood estimates of $\beta$ and $\xi$ could also be replaced by more advanced ML estimates if more was known about the dependence structure. See for instance Drees (2000).

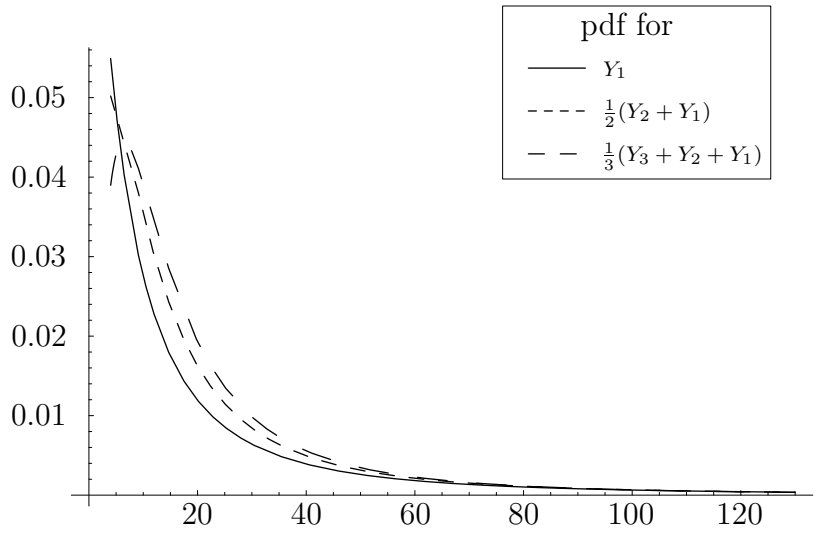


Figure 5: Note the different behaviours for the tails and how this would affect the estimation of $\xi$ so that smaller values of $\hat{\xi}$ would be expected for large values of $p$, i.e. when we are far from the asymptotic tail-behaviour.

### 4.1 Further applications

A common diagnostics tool for fitting the order of an ARMA-process is to examine the behaviour of

$$\hat{\rho}_n(h) = \frac{\frac{1}{n}\sum_{k=1}^{n-h} X_k X_{k+h}}{\frac{1}{n}\sum_{k=1}^{n} X_k^2}, \quad h = 0, 1, 2, \ldots.$$

In the heavy-tailed case, this might not be a good idea since the distribution of $\hat{\rho}_n(h)$ may sometimes be random, see Resnick et al. (1999).

Since the numerator in $\hat{\rho}_n(h)$ is an estimate of $\mathsf{E}[X_1 X_{1+h}]$ and the denominator is an estimate of $\mathsf{E}[X_1^2]$, an idea would be to replace these by estimates of the kind described in Section 2. Let $\hat{M}(h)$ be such an estimate of $\mathsf{E}[X_1 X_{1+h}]$ for $h = 0, 1, 2, \ldots$ and approximate $\hat{\rho}_n(h)$ by $\hat{\rho}_M(h) = \hat{M}(h)/\hat{M}(0)$. Further, in the heavy-tailed case, $\hat{\rho}_n(h)$ is expected to be close to zero in case of independent variables.

For $\hat{\rho}_M(h)$ to be useful as a diagnostic tool, we need to know its asymptotic variance. Let $\mu_h = \mathsf{E}[\hat{M}(h)]$, $\sigma_h^2 = \mathsf{Var}(\hat{M}(h))$ and $\sigma_{kh} = \mathsf{Cov}(\hat{M}(k), \hat{M}(h))$ for $h, k = 0, 1, 2, \ldots$. Using the $\delta$-method we find that

$$\mathsf{Var}(\hat{\rho}_M(h)) \approx \frac{\sigma_h^2}{\mu_0^2} + \frac{\mu_h}{\mu_0^3}\Big[\frac{\mu_h}{\mu_0}\sigma_0^2 - \sigma_{0h}\Big].$$

How estimation of $\sigma_{0h}$ should be made is not clear at this time and further research is required. One possible way around this problem might be to use a bootstrapping scheme for $\hat{M}$.

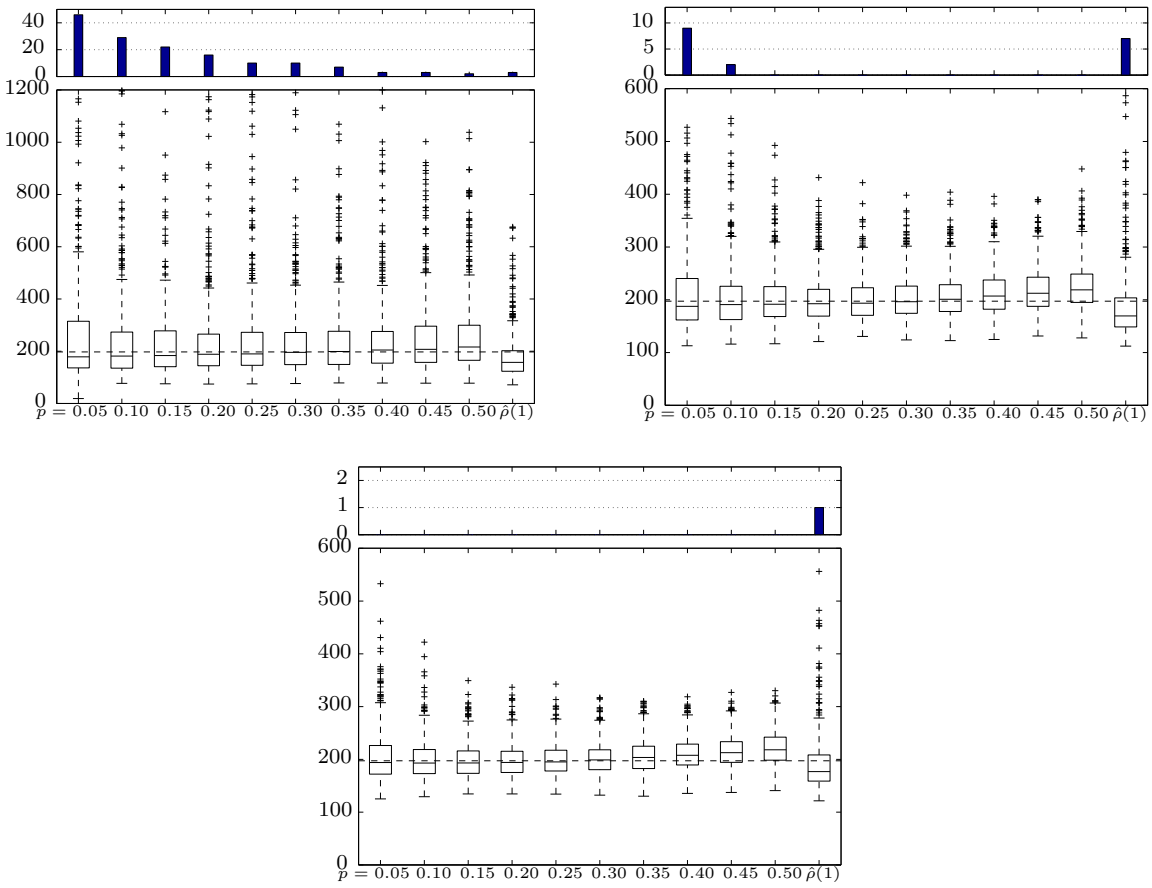


Figure 6: Simulation three. 500 samples of the MA(1) process with GPD($\xi = 0.35$,$\beta = 10$) innovations and with sample sizes *Top left:* 1000, *top right:* 5000 and *bottom:* 10000. The bar graphs show how many observations were too large to fit into the boxplots and the dashed lines indicate the true value of the covariance.

# A Proofs

Repeated use will be made of the following theorem which follows directly from Utev (1990).

**Theorem A.1**
Let $\xi_{n,1}, \ldots, \xi_{n,N_n}$, $n \geq 1$ be a triangular array of $m$-dependent random variables with zero mean and finite variances. Put

$$S_{n,N_n} = \sum_{i=1}^{N_n} \xi_{n,i} \quad \text{and} \quad \sigma^2_{n,N_n} = \mathsf{E}[S^2_{n,N_n}]$$

and assume that

$$\frac{1}{\sigma^2_{n,N_n}} \sum_{i=1}^{N_n} \mathsf{E}[\xi^2_{n,i}; |\xi_{n,i}| \geq \epsilon \sigma_{n,N_n}] \to 0, \quad \text{as } n \to \infty,$$

for each $\epsilon > 0$. Then

$$\frac{S_{n,N_n}}{\sigma_{n,N_n}} \to_d \mathcal{N}(0,1).$$

□

**Corollary A.1**
Let $\xi^{(i)}_{n,1}, \ldots, \xi^{(i)}_{n,N_n}$, $n \geq 1$, $i = 1, \ldots, r$ be triangular arrays of $m$-dependent r.v.'s with zero mean and finite variances. That is, the r.v.'s $\{\xi^{(i)}_{n,k}, \ldots, \xi^{(i)}_{n,k+m}, i = 1, \ldots, r\}$ are dependent, $k = 1, \ldots, N_n - m$. Let $S_i = \sum_{k=1}^{N_n} \xi^{(i)}_{n,k}$, $i = 1, \ldots, r$ and $\sigma_i^2 = \mathsf{E}[S_i^2]$.

In order to conclude that

$$\frac{1}{\sigma_1} \begin{bmatrix} S_1 \\ \vdots \\ S_r \end{bmatrix} \to_d \mathcal{N}(0, T),$$

for some matrix $T$, it is enough that all the $\sigma_i^2$ are of the same order, i.e.

$$(i) \qquad \frac{\sigma_i^2}{\sigma_j^2} = O_+(1), \quad i \neq j$$

and that the row-sums satisfy a Lindeberg condition

$$(ii) \qquad \frac{1}{\sigma_i^2} \sum_{k=1}^{N_n} \mathsf{E}[(\xi^{(i)}_{n,k})^2; |\xi^{(i)}_{n,k}| \geq \epsilon \sigma_i] \to 0, \quad \text{for all } \epsilon > 0$$

Proof: The proof is based on the Lindeberg-Feller theorem and the Cramér-Wold device. Let $\theta = (\theta_1, \ldots, \theta_r) \in \mathbb{R}^r$ and define $S_\theta = \theta \cdot (S_1, \ldots, S_r) = \sum_{k=1}^{N_n} \{\theta_1 \xi^{(1)}_{n,k} + \ldots + \theta_r \xi^{(r)}_{n,k}\}$.

By theorem A.1 we have that $S_\theta/\sigma_1 \to_d \mathcal{N}(0,\sigma_\theta)$, for some constant $\sigma_\theta$, if the following tends to zero for all $\epsilon > 0$

$$\frac{1}{\sigma_1^2}\sum_{k=1}^{N_n} \mathsf{E}[(\theta_1\xi_{n,k}^{(1)} + \ldots + \theta_r\xi_{n,k}^{(r)})^2; |\theta_1\xi_{n,k}^{(1)} + \ldots + \theta_r\xi_{n,k}^{(r)}| \geq \epsilon\sigma_1] = (*).$$

If this holds for all $\theta$ then, by the Cramér-Wold device, the vector $\sigma_1^{-1}(S_1,\ldots,S_r)^t$ also converges to an $r$-dimensional normal distribution.

Now apply the inequality

$$\mathsf{E}[(\xi_1 + \ldots + \xi_r)^2; |\xi_1 + \ldots + \xi_r| \geq \epsilon] \leq r^2\mathsf{E}[\xi_1^2; |\xi_1| \geq \epsilon/r] + \ldots + r^2\mathsf{E}[\xi_r^2; |\xi_r| \geq \epsilon/r]$$

to $(*)$. Then

$$(*) \leq \frac{\theta_1^2 r^2}{\sigma_1^2}\sum_{k=1}^{N_n} \mathsf{E}[(\xi_{n,k}^{(1)})^2; |\theta_1\xi_{n,k}^{(1)}| \geq \epsilon\sigma_1/r] + \ldots$$
$$\ldots + \theta_r^2 r^2 \frac{\sigma_r^2}{\sigma_1^2}\frac{1}{\sigma_r^2}\sum_{k=1}^{N_n} \mathsf{E}[(\xi_{n,k}^{(r)})^2; |\theta_r\xi_{n,k}^{(r)}| \geq \epsilon\sigma_r(\sigma_1/\sigma_r)/r] \to 0$$

by assumptions $(i)$ and $(ii)$ and the result follows. □

## A.1 Proof of Lemma 2.1

Sketching the proof, let $Y \sim F_u$. Then, following Smith (1987)

$$\begin{aligned}
\mathsf{E}\Big[\Big(1+\frac{Y}{u}\Big)^{-r}\Big] &= \int_{y=0}^{\infty}\Big(1+\frac{y}{u}\Big)^{-r} f_u(y)dy = \int_{t=1}^{\infty} t^{-r} f_u(tu-u)udt \\
&= 1 - \int_{t=1}^{\infty} rt^{-r-1-1/\xi}dt - \Big(\int_{t=1}^{\infty} rt^{-r-1-1/\xi-\delta}\frac{L(tu)}{L(u)}dt \\
&\quad - \int_{t=1}^{\infty} rt^{-r-1-1/\xi}dt\Big)\frac{u^{-\delta}L(u)}{1+u^{-\delta}L(u)} \\
&= \{\text{Use Karamata's Theorem, assume that } u \text{ is large}\} \\
&\sim \frac{1}{1+r\xi} + \frac{1}{1+r\xi}\cdot\frac{r\delta\xi^2}{(r+\delta)\xi+1}u^{-\delta}L(u).
\end{aligned}$$

The other results follow in a similar fashion.

## A.2 Proof of Lemma 2.2

Let $Y_1, \ldots, Y_n \sim F_u$ where $F_u$ is defined in (2). Then, using Lemma 2.1, we find that

$$\begin{aligned}
\mathsf{E}[U_\beta] &= \sum_{k=1}^{n} \mathsf{E}\Big[\frac{1}{\beta\xi} - \frac{1}{\beta}\Big(\frac{1}{\xi}-1\Big)\Big(1+\xi\frac{X_k-u_n}{\beta}\Big)^{-1}\Big| X_k > u_n\Big]\mathsf{P}(X_k > u_n) \\
&= \sum_{k=1}^{n} \mathsf{E}\Big[\frac{1}{\beta\xi} - \frac{1}{\beta}\Big(\frac{1}{\xi}-1\Big)\Big(1+\xi\frac{Y_k}{\beta}\Big)^{-1}\Big]p_n \\
&\sim \sum_{k=1}^{n} \Big\{\frac{1}{\beta\xi} - \frac{1}{\beta}\Big(\frac{1}{\xi}+1\Big)\Big[\frac{1}{1+\xi} + \frac{1}{1+\xi}\cdot\frac{\delta\xi^2}{(1+\delta)\xi+1}u^{-\delta}L(u)\Big]\Big\}p_n \\
&= \frac{-\delta\xi}{\beta((1+\delta)\xi+1)} n p_n u^{-\delta} L(u).
\end{aligned}$$

$\mathsf{E}[U_\xi]$, $\mathsf{E}[J_{\beta\beta}]$, $\mathsf{E}[J_{\beta\xi}]$ and $\mathsf{E}[J_{\xi\xi}]$ are calculated in a similar fashion.

Let $C = \mathrm{diag}(\beta_n, 1)$. To prove that $C(np_n)^{-1}J_nC \to_p G^{-1}$, we examine each of the components of the matrix individually. Let $r_n$ be a sequence of numbers and recall that $\beta = O_+(u_n)$. Further, let $\beta^2 J_{\beta\beta}$ be the $(1,1)$ component in the $C(np_n)^{-1}J_nC$ matrix, see Equation (5). Let $\sigma_{\beta\beta} = \mathsf{Var}(\beta^2 J_{\beta\beta})$ and select $r_n$ so that $\sigma_{\beta\beta}/r_n^2 \to 0$. It then follows that $r_n^{-1}(\beta^2 J_{\beta\beta} - \mathsf{E}[\beta^2 J_{\beta\beta}]) \to_p 0$.

From (5) and using the $m$-dependence between the random variables $X_k$, Cauchy-Schwarz inequality and Lemma 2.1, we find that

$$\begin{aligned}
\sigma^2_{\beta\beta} &= \mathsf{Var}\Big(\sum_{k=1}^{n}\Big\{\frac{1}{\xi} - \frac{1+\xi}{\xi}\Big(1+\xi\frac{X_k-u_n}{\beta}\Big)^{-2}\Big\}\mathbf{1}_{\{X_k>u_n\}}\Big) \\
&\le n(2m+1)\mathsf{E}\Big[\Big\{\frac{1}{\xi} - \frac{1+\xi}{\xi}\Big(1+\xi\frac{X_1-u_n}{\beta}\Big)^{-2}\Big\}^2\mathbf{1}_{\{X_1>u_n\}}\Big] \\
&= \{Y \sim F_u, \text{ where } F_u \text{ is defined in (2)}\} \\
&= np_n(2m+1)\mathsf{E}\Big[\Big\{\frac{1}{\xi} - \frac{1+\xi}{\xi}\Big(1+\xi\frac{Y}{\beta}\Big)^{-2}\Big\}^2\Big] \sim np_n(2m+1)(a+bu^{-\delta}L(u)),
\end{aligned}$$

for constants $a$ and $b$ which can be found from Lemma 2.1. Hence, selecting $r_n = np_n$ we obtain $(np_n)^{-1}\beta_n^2 J_{\beta\beta} - \mathsf{E}[(np_n)^{-1}\beta_n^2 J_{\beta\beta}] \to_p 0$. Similar calculations are done for the other components. From the earlier results it also follows that $(np_n)^{-1}\mathsf{E}[CJ_nC] - G^{-1} \to 0$, which proves the last assertion of the lemma.

## A.3 Proof of Lemma 2.3

This is a direct application of Theorem A.1 and Corollary A.1. The aim is to show convergence of $(np_n)^{-1/2}U(\beta,\xi)$, where $U(\beta,\xi) = (U_\beta, U_\xi)^t$ is the score function from Equation (4). Let

$$\begin{aligned}
\xi_{n,k} &= \Big[\frac{1}{\beta\xi} - \frac{1}{\beta}\Big(\frac{1}{\xi}+1\Big)\Big(1+\xi\frac{X_k-u_n}{\beta}\Big)^{-1}\Big]\mathbf{1}_{\{X_k>u_n\}} \qquad \text{and} \\
\eta_{n,k} &= \Big[\frac{1}{\xi^2}\ln\Big(1+\xi\frac{X_k-u_n}{\beta}\Big) - \frac{\xi+1}{\xi^2} + \frac{\xi+1}{\xi^2}\Big(1+\xi\frac{X_k-u_n}{\beta}\Big)^{-1}\Big]\mathbf{1}_{\{X_k>u_n\}}.
\end{aligned}$$

Further, let

$$T_{1,n} = \frac{\beta}{\sqrt{np_n}}(U_\beta - \mathsf{E}[U_\beta]) = \sum_{k=1}^{n} \frac{\beta}{\sqrt{np_n}}(\xi_{n,k} - \mathsf{E}[\xi_{n,k}]) \quad \text{and}$$

$$T_{2,n} = \sqrt{np_n}(U_\xi - \mathsf{E}[U_\xi]) = \sum_{k=1}^{n} \frac{1}{\sqrt{np_n}}(\eta_{n,k} - \mathsf{E}[\eta_{n,k}]).$$

We now apply Corollary A.1. First look at

$$\sigma_{1,n}^2 = \mathsf{Var}(T_{1,n}) = \frac{\beta^2}{np_n}\Big[n\mathsf{Var}(\xi_{n,1}) + 2(n-1)\mathsf{Cov}(\xi_{n,1},\xi_{n,2}) + \\ \ldots + 2(n-m)\mathsf{Cov}(\xi_{n,1},\xi_{n,m+1})\Big]$$

A similar calculation holds for $\sigma_{2,n}^2 = \mathsf{Var}(T_{2,n})$. Had the variables $\xi_{n,k}$ been independent, $\sigma_{1,n}^2 = O_+(1)$. For the case with $m$-dependent r.v.'s this can be modified a bit and we assume that $\mathsf{Var}(\beta U_\beta) = np_n\sigma_{1,n}^2 \to \infty$, $\mathsf{Var}(U_\xi) = np_n\sigma_{2,n}^2 \to \infty$ and $\sigma_{1,n}^2/\sigma_{2,n}^2 = O_+(1)$. This means that $\sigma_{k,n}^2$, $k = 1, 2$, can be allowed to tend to zero, as long as the rate is not too fast.

It is straightforward to check convergence for $T_{1,n}$ using Theorem A.1:

$$\begin{aligned}
&\frac{1}{\sigma_{1,n}^2}\sum_{k=1}^{n}\mathsf{E}\Big[\Big(\frac{\beta}{\sqrt{np_n}}(\xi_{n,k} - \mathsf{E}[\xi_{n,k}])\Big)^2; \Big|\frac{\beta}{\sqrt{np_n}}(\xi_{n,k} - \mathsf{E}[\xi_{n,k}])\Big| \ge \epsilon\sigma_{1,n}\Big] \\
&= \frac{1}{p_n\sigma_{1,n}^2}\mathsf{E}\Big[\Big(\Big\{\frac{1}{\xi} - \Big(\frac{1}{\xi}+1\Big)\Big(1+\xi\frac{X_1-u}{\beta}\Big)^{-1}\Big\}\mathbf{1}_{\{X_1>u\}} - \frac{\mathsf{E}[\xi_{n,1}]}{\beta}\Big)^2; \\
&\quad \Big|\Big\{\frac{1}{\xi} - \Big(\frac{1}{\xi}+1\Big)\Big(1+\xi\frac{X_1-u}{\beta}\Big)^{-1}\Big\}\mathbf{1}_{\{X_1>u\}} - \frac{\mathsf{E}[\xi_{n,1}]}{\beta}\Big| \ge \epsilon\sqrt{np_n}\sigma_{1,n}\Big] \\
&= \{\text{Let } Y \sim F_u, \text{ where } F_u \text{ is given in (2), condition on } \{X_k > u\} \\
&\quad \text{and use Cauchy-Schwarz inequality }\} \\
&\le \frac{1}{\sigma_{1,n}^2}\mathsf{E}\Big[\Big\{\frac{1}{\xi} - \Big(\frac{1}{\xi}+1\Big)\Big(1+\xi\frac{Y_1}{\beta}\Big)^{-1} - p_n\mathsf{E}\Big[\frac{1}{\xi} - \Big(\frac{1}{\xi}+1\Big)\Big(1+\xi\frac{Y_1}{\beta}\Big)^{-1}\Big]\Big\}^4\Big]^{1/2} \\
&\quad \cdot \mathsf{P}\Big(\Big|\frac{1}{\xi} - \Big(\frac{1}{\xi}+1\Big)\Big(1+\xi\frac{Y_1}{\beta}\Big)^{-1} - p_n\mathsf{E}\Big[\frac{1}{\xi} - \Big(\frac{1}{\xi}+1\Big)\Big(1+\xi\frac{Y_1}{\beta}\Big)^{-1}\Big]\Big| \ge \epsilon\sqrt{np_n}\sigma_{1,n}\Big)^{1/2} \\
&\quad \to 0
\end{aligned}$$

since the expected value tends to a finite constant by Lemma A.1 and the probability tends to zero since $np_n\sigma_{1,n}^2 \to \infty$ by assumption and the term within $|\cdot|$ is smaller than a constant. The second sum, $T_{2,n}$, is treated similarly although the calculations for that case are slightly longer.

Further, by Lemma 2.1 and for any $a$ and $b$,

$$\begin{aligned}
&\frac{a\beta}{\sqrt{np_n}}\mathsf{E}[U_\beta] + \frac{b}{\sqrt{np_n}}\mathsf{E}[U_\xi] \\
&\qquad \sim -\delta\xi\Big(\frac{a}{(1+\delta)\xi+1} + \frac{b}{(1+\delta\xi)((1+\delta)\xi+1)}\Big)\sqrt{np_n}u^{-\delta}L(u) \to 0
\end{aligned}$$

only if $\sqrt{np_n}u^{-\delta} \propto n^{1/2-\alpha(1/2+\delta\xi)} \to 0$, i.e. if $\alpha > (1+2\delta\xi)^{-1}$. Then, under this assumption and using Corollary A.1

$$\frac{1}{\sqrt{np_n}} \begin{bmatrix} \beta U_\beta \\ U_\xi \end{bmatrix} \to_d \mathcal{N}(0, \Sigma), \quad \text{as } n \to \infty,$$

where $\Sigma$ is the covariance matrix. This concludes the proof.

### A.3.1 Estimating $\Sigma$

The covariance matrix $\Sigma$ is most easily estimated numerically. This can be done, for instance, by first substituting estimates $\hat{\beta}$ and $\hat{\xi}$ for $\beta$ and $\xi$ in the expressions for $\xi_{n,k}$ and $\eta_{n,k}$ above. Then calculate the sample equivalents of $\sigma^2_{1,n}$ and $\sigma^2_{2,n}$ as indicated above. The off-diagonal term in $\Sigma$ is treated similarly.

The difficulty here is how to know how many of the covariance terms $\mathsf{Cov}(\xi_{n,1}, \xi_{n,1+k})$ to include in the estimates of $\sigma^2_{1,n}$. One possibility is to make an autocorrelation plot and try to estimate how far the $m$-dependence goes. Another possibility might be to make a plot of $\sigma^2_{1,n}$ for different choices of $m$ and select an $m$ where the plot seems to flatten out.

## A.4 Proof of Lemma 2.5

The proof is a direct application of Theorem A.1. Let

$$\hat{\mu} - \mu_n = \sum_{k=1}^{n} \frac{X_k I_k - \mu_n}{n} =: \sum_{k=1}^{n} \zeta_{n,k} =: S_n.$$

Then $\mathsf{E}[\zeta_{n,k}] = 0$ and we denote

$$\begin{aligned}
\sigma_n^2 &= \mathsf{E}[S_n^2] = \{m\text{-dependence and stationarity}\} \\
&= \frac{1}{n}\mathsf{Var}(X_1 I_1) + \frac{2(n-1)}{n^2}\mathsf{Cov}(X_1 I_1, X_2 I_2) + \ldots + \frac{2(n-m)}{n^2}\mathsf{Cov}(X_1 I_1, X_{m+1} I_{m+1}).
\end{aligned}$$

Assume that $\sigma_n^2 \propto n^{\rho-1}$ for some $\rho$. Since, for positive r.v.'s $\eta$

$$\mathsf{E}[\eta^k] = \int_0^\infty k x^{k-1} \mathsf{P}(\eta > x) dx,$$

we note that if $X_1, \ldots, X_n$ were independent, then $\sigma_n^2 = O_+(u_n^{2-1/\xi})$. Since $u_n = O_+(n^{\alpha\xi})$ for $\alpha \in (0,1)$, this corresponds to $\rho = \alpha(2\xi - 1)$ for the independent case.

We now turn to the Lindeberg condition in Theorem A.1:

$$\begin{aligned}
&\frac{1}{\sigma_n^2}\sum_{k=1}^{n} \mathsf{E}[\zeta_{n,k}^2; |\zeta_{n,k}| \geq \sigma_n] = \frac{1}{n\sigma_n^2}\mathsf{E}[(X_1 I_1 - \mu_n)^2; |X_1 I_1 - \mu_n| \geq n\sigma_n] \\
&\quad = \{\mathsf{E}[|XY|] \leq \mathsf{E}[X^2]^{1/2}\mathsf{E}[Y^2]^{1/2}, \text{ and } \mathsf{E}[(X_1 I_1 - \mu_n)^4] = O_+(u^{4-1/\xi})\} \\
&\quad \leq \frac{u_n^{2-1/2\xi}}{n^\rho} O_+(1)\Big\{\mathsf{P}(X_1 I_1 \leq -n\sigma_n + \mu_n) + \mathsf{P}(X_1 I_1 \geq n\sigma_n + \mu_n)\Big\}^{1/2} = (*).
\end{aligned}$$

Since $\xi \in (0,1)$ it follows that $\mu_n = \mathsf{E}[X_1 I_1]$ tends to a finite, positive constant. Further, $X_1 I_1 > 0$ and $n\sigma_n = O_+(n^{(\rho+1)/2}) \to \infty$ if $\rho > -1$, so $\mathsf{P}(X_1 I_1 \leq -n\sigma_n + \mu_n) = 0$ for large enough $n$.

Finally, $X_1 I_1 \leq u_n$, so if $n\sigma_n = O_+(n^{(\rho+1)/2})$ grows faster than $u_n$, then $\mathsf{P}(X_1 I_1 \geq n\sigma_n + \mu_n) = 0$ for large enough $n$. This corresponds to $\rho > 2\alpha\xi - 1$.

Hence $(*) = 0$ for large values of $n$ if $\rho > 2\alpha\xi - 1$. The result now follows from Theorem A.1.

## A.5 Proof of Lemma 2.6

Let $H_k = \mathbf{1}_{\{X_k > u\}}$, where $X_1, \ldots, X_n \sim F$ and are $m$-dependent. Then

$$\hat{p} - p_n = \sum_{k=1}^{n} \frac{H_k - p_n}{n} = \sum_{k=1}^{n} \frac{\zeta_{n,k}}{n} = S_n,$$

where $p_n = \mathsf{E}[H_1]$. Further, let

$$\begin{aligned}
\sigma_p^2 = \mathsf{E}[S_n^2] &= \{m\text{-dependence and stationarity}\} \\
&= \frac{1}{n}\Big[\mathsf{Var}(H_1) + 2\mathsf{Cov}(H_1, H_2) + \ldots + 2\mathsf{Cov}(H_1, H_{1+m})\Big] \\
&\quad - \frac{2}{n^2}\Big[\mathsf{Cov}(H_1, H_2) + 2\mathsf{Cov}(H_1, H_3) + \ldots + m\mathsf{Cov}(H_1, H_{1+m})\Big]
\end{aligned}$$

where $\mathsf{Var}(H_1) = p_n(1-p_n)$ and $-p_n^2 \leq \mathsf{Cov}(H_1, H_k) \leq p_n(1-p_n)$ for $k = 1, \ldots, m$. It follows that $\sigma_p^2 = O_+(p_n/n)$. We now use Theorem A.1 and, for each $\epsilon > 0$, examine

$$\begin{aligned}
\frac{1}{\sigma_p^2}\sum_{k=1}^{n} \mathsf{E}\Big[\Big(\frac{\zeta_{n,k}}{n}\Big)^2; \Big|\frac{\zeta_{n,k}}{n}\Big| \geq \epsilon\sigma_p\Big] &= \frac{1}{n\sigma_p^2}\mathsf{E}[\underbrace{(H_1 - p_n)^2}_{\leq 1}; |H_1 - p_n| \geq \epsilon n\sigma_p] \\
&\leq \frac{1}{n\sigma_p^2}\mathsf{P}(|H_1 - p_n| \geq \epsilon n\sigma_p) \to 0,
\end{aligned}$$

since $n\sigma_p \to \infty$. Hence, by Theorem A.1

$$\frac{1}{\sigma_p}(\hat{p} - p_n) \to_d \mathcal{N}(0,1), \quad \text{as } n \to \infty.$$

## A.6 Proof of Lemma 2.7

Let

$$\begin{aligned}
\xi_{n,k} &= \frac{1}{\sqrt{np_n}}\Big\{\frac{1}{\xi} - \Big(\frac{1}{\xi} + 1\Big)\Big(1 + \xi\frac{X_k - u_n}{\beta}\Big)^{-1}\Big\}\mathbf{1}_{\{X_k > u_n\}} \qquad \text{and} \\
\eta_{n,k} &= \frac{1}{\sqrt{np_n}}\Big\{\frac{1}{\xi^2}\ln\Big(1 + \xi\frac{X_k - u_n}{\beta}\Big) - \frac{\xi+1}{\xi^2}\Big(1 - \Big(1 + \xi\frac{X_k - u_n}{\beta}\Big)^{-1}\Big)\Big\}\mathbf{1}_{\{X_k > u_n\}}.
\end{aligned}$$

Then the score function $U(\beta, \xi) = (U_\beta, U_\xi)^t$ defined in (4) can be written as

$$\beta U_\beta = \sum_{k=1}^{n} \xi_{n,k} \quad \text{and} \quad U_\xi = \sum_{k=1}^{n} \eta_{n,k}.$$

Let

$$\begin{aligned}
S_1 &= \frac{1}{\sigma_\mu}(\hat{\mu} - \mu_n) = \frac{1}{\sigma_\mu} \sum_{k=1}^{n} (X_k \mathbf{1}_{\{X_k \le u_n\}} - \mu_n) \quad \text{as in Lemma A.4,} \\
S_2 &= \frac{1}{\sigma_p}(\hat{p} - p_n) = \frac{1}{\sigma_p} \sum_{k=1}^{n} (\mathbf{1}_{\{X_k > u_n\}} - p_n) \qquad \text{as in Lemma A.5,} \\
S_3 &= \frac{1}{\sqrt{np_n}}(U_\beta - \mathsf{E}[U_\beta]) = \frac{1}{\sqrt{np_n}} \sum_{k=1}^{n} (\xi_{n,k} - \mathsf{E}[\xi_{n,k}]) \quad \text{and} \\
S_4 &= \sqrt{np_n}(U_\xi - \mathsf{E}[U_\xi]) = \frac{1}{\sqrt{np_n}} \sum_{k=1}^{n} (\eta_{n,k} - \mathsf{E}[\eta_{n,k}]) \quad \text{as in Lemma A.3.}
\end{aligned}$$

Then, by Corollary A.1 and the proofs of Lemmas A.3 – A.5

$$\Sigma_n^{-1/2} \begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \end{bmatrix} \to_d \mathcal{N}(0, I), \quad \text{where } \Sigma_n = \mathsf{Cov}((S_1, S_2, S_3, S_4)^t),$$

since the $S_k$ all have variances of the same order. By Lemma 2.3

$$\Sigma_n^{-1/2} \begin{bmatrix} 0 \\ 0 \\ \beta \mathsf{E}[U_\beta]/\sqrt{np_n} \\ \mathsf{E}[U_\xi]/\sqrt{np_n} \end{bmatrix} \to 0,$$

which means that

$$\Sigma_n^{-1/2} \Psi_n = \Sigma_n^{-1/2} \begin{bmatrix} (\hat{\mu} - \mu_n)/\sigma_\mu \\ (\hat{p} - p_n)/\sigma_p \\ \beta U_\beta/\sqrt{np_n} \\ U_\xi/\sqrt{np_n} \end{bmatrix} \to_d \mathcal{N}(0, I).$$

Another way to express this is to let

$$F = \begin{bmatrix} \sigma_\mu & 0 & 0 & 0 \\ 0 & \sigma_p & 0 & 0 \\ 0 & 0 & 2(1+\xi)/\sqrt{np_n} & -(1+\xi)/\sqrt{np_n} \\ 0 & 0 & -(1+\xi)/\sqrt{np_n} & (1+\xi)^2/\sqrt{np_n} \end{bmatrix}, \text{ so that } F\Psi_n \approx \begin{bmatrix} \hat{\mu} - \mu_n \\ \hat{p} - p_n \\ \hat{\beta}_n/\beta - 1 \\ \hat{\xi}_n - \xi \end{bmatrix}.$$

This latter form will be used in the proof of the following theorem. This means that $(F\Sigma_n F^t)^{-1/2}(F\Psi_n) \to_d \mathcal{N}(0, I)$.

## A.7 Proof of Theorem 2.1

We will begin by Taylor expanding the expression for $\hat{M} - M$. After that we identify which components are of order $o_p(1)$ and finally apply the Continuous Mapping Theorem and Lemma 2.7 in order to arrive at the asymptotic distribution. So, let

$$\begin{aligned}\hat{M} - M &= \hat{\mu} - \mu_n + \hat{p}\Big(u_n + \frac{\hat{\beta}_n}{1-\hat{\xi}_n}\Big) - p_n\Big(u_n + \frac{\beta}{1-\xi}\Big) \\ &= \hat{\mu} - \mu_n + (\hat{p} - p_n)\Big(u_n + \frac{\hat{\beta}_n}{1-\hat{\xi}_n}\Big) + p_n\Big(\frac{\hat{\beta}_n}{1-\hat{\xi}_n} - \frac{\beta}{1-\xi}\Big).\end{aligned}$$

Taylor expansion and dividing by $\sigma_\mu$ leaves us with

$$\begin{aligned}\frac{1}{\sigma_\mu}(\hat{M} - M) &= \frac{1}{\sigma_\mu}(\hat{\mu} - \mu_n) + \frac{1}{\sigma_\mu}(\hat{p} - p_n)\Big(u_n + \frac{\beta}{1-\xi}\Big) \\ &\quad + \frac{1}{\sigma_\mu(1-\xi)}(\hat{p} - p_n)(\hat{\beta}_n - \beta)(1 + o_p(1)) \\ &\quad + \frac{\beta}{\sigma_\mu(1-\xi)^2}(\hat{p} - p_n)(\hat{\xi}_n - \xi)(1 + o_p(1)) \\ &\quad + \frac{p_n}{\sigma_\mu(1-\xi)}(\hat{\beta}_n - \beta)(1 + o_p(1)) + \frac{\beta p_n}{\sigma_\mu}\frac{\hat{\xi}_n - \xi}{(1-\xi)^2}(1 + o_p(1)) \\ &= I + II + III + IV + V + VI.\end{aligned}$$

We look at these components one at a time. Assume that $\rho = \alpha(2\xi - 1)$. Then $\sigma_p/\sigma_\mu \propto u_n^{-1}$ so, using $\beta = O_+(u_n)$,

$$III = \underbrace{\frac{\hat{p} - p_n}{\sigma_p}}_{\to_d \mathcal{N}(0,1)} \underbrace{\Big(\frac{\sqrt{np_n}}{\beta}(\hat{\beta}_n - \beta)\Big)}_{\to_d \mathcal{N}(\cdot,\cdot)} \underbrace{\frac{\sigma_p \beta}{\sigma_\mu \sqrt{np_n}(1-\xi)}}_{\propto (np_n)^{-1/2} \to 0}(1 + o_p(1)) \to_p 0$$

by Lemmas 2.6 and 2.4.

$$IV = \underbrace{\frac{\hat{p} - p_n}{\sigma_p}}_{\to_d \mathcal{N}(0,1)} \underbrace{\frac{\sigma_p \beta}{\sigma_\mu(1-\xi)^2}}_{=O_+(1)} \underbrace{\sqrt{np_n}(\hat{\xi}_n - \xi)}_{\to_d \mathcal{N}(\cdot,\cdot)} \underbrace{\frac{1}{\sqrt{np_n}}}_{\to 0}(1 + o_p(1)) \to_p 0,$$

by Lemma 2.6. Further

$$\begin{aligned}V &= \frac{p_n}{\sigma_\mu(1-\xi)}(\hat{\beta}_n - \beta)(1 + o_p(1)) \\ &= \underbrace{\frac{\sqrt{np_n}}{\beta}(\hat{\beta}_n - \beta)}_{\to_d \mathcal{N}(\cdot,\cdot)} \underbrace{\frac{p_n \beta}{\sqrt{np_n}\sigma_\mu}(1-\xi)}_{=O_+(1) \text{ if } \rho = \alpha(2\xi-1)} (1 + o_p(1)).\end{aligned}$$

Finally

$$VI = \frac{\beta p_n}{\sigma_\mu}\frac{\hat{\xi}_n - \xi}{(1-\xi)^2} = \underbrace{\frac{\beta p_n}{\sigma_\mu \sqrt{np_n}(1-\xi)^2}}_{=O_+(1)} \underbrace{\sqrt{np_n}(\hat{\xi}_n - \xi)}_{\to_d \mathcal{N}(\cdot,\cdot)}(1 + o_p(1)).$$

Summing up, we've shown that

$$\frac{1}{\sigma_\mu}(\hat{M} - M) = \frac{1}{\sigma_\mu}\left[1, \quad u_n + \tfrac{\beta}{1-\xi}, \quad \tfrac{p_n\beta}{1-\xi}, \quad \tfrac{p_n\beta}{(1-\xi)^2}\right]\begin{bmatrix} \hat{\mu} - \mu_n \\ \hat{p} - p_n \\ \hat{\beta}_n/\beta - 1 \\ \hat{\xi}_n - \xi \end{bmatrix} + o_p(1)$$

$$\to_d \mathcal{N}(0, \sigma_M^2),$$

for some constant $\sigma_M^2$, by Lemma 2.7 and using the Continuous Mapping Theorem. More specifically

$$\sigma_M^2 = \lim_{n\to\infty} \frac{1}{\sigma_\mu^2} AF\Sigma_n F A^t,$$

where

$$A = \left[1, \quad u + \tfrac{\beta}{1-\xi}, \quad \tfrac{p_n\beta}{1-\xi}, \quad \tfrac{p_n\beta}{(1-\xi)^2}\right]$$

and $F$ and $\Sigma_n$ are the matrices from the proof of Lemma 2.7.

# References


Athreya, K. B. (1987), "Bootstrap of the mean in the infinite variance case," *The annals of statistics*, 15, 724–731.

Bingham, N., Goldie, C., and Teugels, J. (1987), *Regular Variation*, no. 27 in Encyclopedia of Mathematics and its Applications, Cambridge University Press.

Brockwell, P. and Davis, R. (1987), *Time Series: Theory and Methods*, Springer, 2nd ed.

Coles, S. (2001), *An introduction to statistical modeling of extreme values*, Springer.

Crovella, M. E. and Taqqu, M. S. (1999), "Estimating the heavy tail index from scaling properties," *Methodol. Comput. Appl. Probab.*, 1, 55–79.

Crovella, M. E., Taqqu, M. S., and Bestavros, A. (1998), "Heavy-Tailed Probability Distributions in the World Wide Web," in *A Practical Guide to Heavy Tails*, eds. Adler, R. J., Feldman, R. E., and Taqqu, M. S., Birkhäuser, pp. 3–25.

Diebold, F., Schuermann, T., and Stroughair, J. D. (1997), "Pitfalls and opportunities in the use of extreme value theory in risk management," *Advances in computational management science*, 2, 3 – 12.

Drees, H. (2000), "Weighted approximations of tail processes for $\beta$-mixing random variables," *Ann. Appl. Probab.*, 10, 1274–1301.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997), *Modelling Extremal Events*, Springer.

Hall, P. and Jing, B.-Y. (1998), "Comparison of bootstrap and asymptotic approximations to the distribution of a heavy-tailed mean," *Statist. Sinica*, 8, 887–906.

Johansson, N. C. J. (2001), "A semiparametric estimator of the mean of heavytailed distributions," Licentiate thesis, Chalmers University of Technology, Sweden.

Longin, F. M. (2000), "From value at risk to stress testing: the extreme value approach," *Journal of banking and finance*, 24, 1097–1130.

Nolan, J. P. (1999), "Fitting Data and Assessing Goodness-of-fit with Stable Distributions," Available from http://academic2.american.edu/ jpnolan/stable/stable.html.

— (2001), "Maximum likelihood estimation and diagnostics for stable distributions," in *Lévy processes*, Boston, MA: Birkhäuser Boston, pp. 379–400.

Peng, L. (2001), "Estimating the mean of a heavy tailed distribution," *Statistics and probability letters*, 52, 255–264.

Pickands, III, J. (1975), "Statistical inference using extreme order statistics," *Ann. Statist.*, 3, 119–131.

Politis, D. N. and Romano, J. P. (1994), "Large sample confidence regions based on subsamples under minimal assumptions," *Ann. Statist.*, 22, 2031–2050.

Politis, D. N., Romano, J. P., and Wolf, M. (1999), *Subsampling*, Springer Series in Statistics, New York: Springer-Verlag.

Reiss, R.-D. and Thomas, M. (2001), *Statistical analysis of extreme values*, Basel: Birkhäuser Verlag, 2nd ed., from insurance, finance, hydrology and other fields, With 1 CD-ROM (Windows).

Resnick, S., Samorodnitsky, G., and Xue, F. (1999), "How misleading can sample ACFs of stable MAs be? (Very!)," *Ann. Appl. Probab.*, 9, 797–817.

Romano, J. P. and Wolf, M. (1999), "Subsampling inference for the mean in the heavy-tailed case," *Metrika*, 50, 55–69.

Smith, R. (1987), "Estimating Tails of Probability Distributions," *The Annals of Statistics*, 15, 1174–1207.

Smith, R. L. (1985), "Maximum likelihood estimation in a class of nonregular cases," *Biometrika*, 72, 67–90.

Tajvidi, N. (1996), "Characterisation and Some Statistical Aspects of Univariate and Multivariate Generalised Pareto Distributions," Ph.D. thesis, Chalmers University of Technology.

Utev, S. (1990), "The central limit theorem for $\varphi$-mixing arrays of random variables," *Theory of probability and its applications*, 35, 131–139.

# Paper C

# An extreme value approach to regression through the origin with an application to superpopulation sampling

Joachim Johansson*

May 6, 2003

**Abstract**

Consider the model $Y_k = aX_k + \epsilon_k\sigma(X_k)$ for regression through the origin with heteroscedastic errors. In the model, $a$ is a constant and the $X_k$ are assumed known and bounded away from zero and infinity. The errors $\epsilon_k$ have one or two polynomially decreasing tails and lack finite variance. Finally, $\sigma(\cdot)$ is an unknown function such that $\sigma(X_k)/X_k$ is bounded away from zero and infinity.

In the paper, an extreme value based estimate, $\hat{a}_M$, of the unknown constant $a$ is suggested. It is shown that $\hat{a}_M$ is asymptotically unbiased and normally distributed and also robust against changes in the underlying variance structure.

In a small simulation study, $\hat{a}_M$ is compared to the standard least squares estimate of $a$. The estimation technique is also applied in a superpopulation sampling framework.

Keywords: Regression, extreme value theory, peaks over threshold, sampling, superpopulation model.

MSC2000 classification: primary – 62G32
secondary – 62F10, 62F12, 62P30

## 1 Background

Consider the model

$$Y_k = aX_k + \epsilon_k\sigma(X_k) \tag{1}$$

for regression through the origin. The parameter $a$ is unknown and the constants $X_k$ and the function $\sigma(\cdot)$ are bounded away from zero and infinity. The $\epsilon_k$ are iid random variables with (at least one) polynomially decreasing tail and $\mathsf{E}[\epsilon_k] = 0$.

*Chalmers University of Technology, Göteborg, Sweden. E-mail: joachimj@math.chalmers.se

We are interested in estimating the regression coefficient $a$. The common method is to use the least squares estimate of $a$:

$$\hat{a}_{LS} = \Big( \sum_{k=1}^{n} X_k Y_k / \sigma(X_k)^2 \Big) \Big/ \Big( \sum_{k=1}^{n} X_k^2 / \sigma(X_k)^2 \Big),$$

where $n$ is the sample size. However, $\hat{a}_{LS}$ has some undesirable properties: it is sensitive to outliers (Chambers (1986)), which makes it unsuitable for the model in Equation (1) if the $\epsilon_k$ have heavy tails. Further, $\hat{a}_{LS}$ will not be asymptotically normally distributed, unless the $\epsilon_k$ have finite variances. Finally, it requires knowledge about the function $\sigma$.

The aim of this paper is to suggest a different approach for estimating $a$ when large samples are available and outliers are considered representative. In Section 2, such an alternative estimate of $a$ is suggested and in Section 3 the method is evaluated through simulations. The proofs have been gathered in an appendix, in order to facilitate reading.

## 2 Estimating $a$

In this section we suggest an estimate for the unknown constant $a$ in (1) based on extreme value theory. First the estimate is motivated and then its properties are stated in Theorem 2.1. All proofs can be found in the appendix.

First note that

$$\frac{Y_k}{X_k} = a + \epsilon_k \frac{\sigma(X_k)}{X_k},$$

which can be written as $Z_k = a + \epsilon_k \sigma_k$, where $Z_k = Y_k/X_k$ and $\sigma_k = \sigma(X_k)/X_k$. Throughout this paper we will work under the following assumption.

**Assumption 2.1**
$\sigma_k = \sigma(X_k)/X_k$ is bounded away from zero and infinity. Further,

$$\begin{aligned} \mathsf{P}(\epsilon_k > x) &= c_u x^{-1/\xi_u}(1 + x^{-\delta_u} L_u(x)), x > 0 \quad \text{and} \\ \mathsf{P}(\epsilon_k < x) &= c_l |x|^{-1/\xi_l}(1 + |x|^{-\delta_l} L_l(|x|)), x < 0, \end{aligned}$$

where $c_u$ and $c_l$ are real constants, $\xi_u, \xi_l \in (0,1)$, $\delta_u, \delta_l > 0$ and the functions $L_u$ and $L_l$ are slowly varying at infinity, i.e. $L_u(tx)/L_u(x) \to 1$ for every $t > 0$ as $x \to \infty$. □

For further properties of slowly varying functions, see Bingham et al. (1987). In the calculations below, more restrictions will be placed on $\delta_u$ and $\delta_l$.

The aim is now to find parametric models for the tails of $Z_k$. We first look at the upper tail, where

$$\mathsf{P}(Z_k > x) = \mathsf{P}\Big(\epsilon_k > \frac{x-a}{\sigma_k}\Big) = c_u\Big(\frac{x-a}{\sigma_k}\Big)^{-1/\xi_u} \cdot \Big[1 + \Big(\frac{x-a}{\sigma_k}\Big)^{-\delta_u} L_u\Big(\frac{x-a}{\sigma_k}\Big)\Big]$$

and thus

$$\begin{aligned}\bar{F}_{u,k}(x) &= \mathsf{P}(Z_k > x + u_u | Z_k > u_u) = \Big(1 + \frac{x}{u_u - a}\Big)^{-1/\xi_u} R_{u,k} \\ &= \Big(1 + \xi_u \frac{x}{\beta_u}\Big)^{-1/\xi_u} R_{u,k}(x), \end{aligned} \tag{2}$$

where $\beta_u = \xi_u(u_u - a)$ and

$$R_{u,k}(x) = \frac{1 + \Big(\frac{x+u_u-a}{\sigma_k}\Big)^{-\delta_u} L_u\Big(\frac{x+u_u-a}{\sigma_k}\Big)}{1 + \Big(\frac{u_u-a}{\sigma_k}\Big)^{-\delta_u} L_u\Big(\frac{u_u-a}{\sigma_k}\Big)}. \tag{3}$$

Hence, the exceedances over the upper threshold $u_u$ follow a distribution that looks like a perturbed generalised Pareto distribution, where only the perturbation $R_{u,k}$ depends on the unknown $\sigma_k$. This method of approximating the exceedances over a threshold goes under the name Peaks Over Threshold (POT), see Leadbetter (1991), Embrechts et al. (1997) or Coles (2001) for an introduction.

A similar result holds for the lower tail, for which the perturbation is called $R_{l,k}$ for the $k$th term; let $y > 0$ and $u_l \to -\infty$, then

$$\bar{F}_{l,k}(y) = \mathsf{P}(-(Z_k - u_l) > y | Z_k < u_l) = \Big(1 + \xi_l \frac{y}{\beta_l}\Big)^{-1/\xi_l} R_{l,k}(y) \tag{4}$$

where

$$R_{l,k}(y) = \frac{1 + \Big(\frac{u_l-y-a}{\sigma_k}\Big)^{-\delta_l} L_l\Big(\frac{u_l-y-a}{\sigma_k}\Big)}{1 + \Big(\frac{u_l-a}{\sigma_k}\Big)^{-\delta_l} L_l\Big(\frac{u_l-a}{\sigma_k}\Big)}. \tag{5}$$

Note that, had it not been for the perturbations $R_{u,k}$, the exceedances over the upper threshold would be iid, and similarly for the lower tail. Still, the observations of the exceedances can be used for calculating maximum likelihood estimates of the GPD parameters $(\beta_u, \xi_u)$ and $(\beta_l, \xi_l)$ for the upper and lower tails, respectively. The likelihood function is then based on the exact GPD, but the underlying distribution is that in (2) and (4). Thus we can estimate the upper and lower tails as

$$\hat{F}_{u,k}(x) = 1 - \Big(1 + \hat{\xi}_u \frac{x}{\hat{\beta}_u}\Big)^{-1/\hat{\xi}_u} \quad \text{and} \quad \hat{F}_{l,k}(x) = 1 - \Big(1 + \hat{\xi}_l \frac{x}{\hat{\beta}_l}\Big)^{-1/\hat{\xi}_l}. \tag{6}$$

These models for the excesses over the upper and lower thresholds, $u_u$ and $u_l$, will be used in the estimation of the parameter $a$ below.

Let $F_k(x) = \mathsf{P}(Z_k \le x)$ and note that we may express $a$ in the following way:

$$a = \mathsf{E}\Big[\frac{1}{n}\sum_{k=1}^{n} Z_k\Big] = \frac{1}{n}\sum_{k=1}^{n}\Big\{\int_{-\infty}^{u_l} z dF_{Z_k}(z) + \int_{u_l}^{u_u} z dF_{Z_k}(z) + \int_{u_u}^{\infty} z dF_{Z_k}(z)\Big\}$$
$$= \frac{1}{n}\sum_{k=1}^{n}\Big\{\int_{-\infty}^{0} (x+u_l) dF_{Z_k}(x+u_l) + \int_{u_l}^{u_u} z dF_{Z_k}(z) + \int_{0}^{\infty} (x+u_u) dF_{Z_k}(x+u_u)\Big\},$$

where $n$ is the sample size. Let $F_{n,k}(x)$ be the empirical cdf for $Z_k$ and assume that $\hat{\xi}_l, \hat{\xi}_u \in (0,1)$. Then $a$ can be estimated by

$$\hat{a}_M = \frac{1}{n}\sum_{k=1}^{n}\Big\{\int_{-\infty}^{0} (x+u_l) d\hat{F}_{l,k}(-x) p_{l,k} + \int_{u_l}^{u_u} x dF_{n,k}(x) + \int_{0}^{\infty} (x+u_u) d\hat{F}_{u,k}(x) p_{u,k}\Big\}$$
$$= \hat{p}_l\Big(u_l - \frac{\hat{\beta}_l}{1-\hat{\xi}_l}\Big) + \frac{1}{n}\sum_{k=1}^{n} Z_k \mathbf{1}_{\{u_l < Z_k < u_u\}} + \hat{p}_u\Big(u_u + \frac{\hat{\beta}_u}{1-\hat{\xi}_u}\Big)$$
$$= \hat{p}_l\Big(u_l - \frac{\hat{\beta}_l}{1-\hat{\xi}_l}\Big) + \hat{\mu} + \hat{p}_u\Big(u_u + \frac{\hat{\beta}_u}{1-\hat{\xi}_u}\Big)$$

where $\hat{p}_l$ and $\hat{p}_u$ are estimates of

$$p_l = \frac{1}{n}\sum_{k=1}^{n} \mathsf{P}(Z_k < u_l) = \frac{1}{n}\sum_{k=1}^{n} p_{l,k} \text{ and } p_u = \frac{1}{n}\sum_{k=1}^{n} \mathsf{P}(Z_k > u_u) = \frac{1}{n}\sum_{k=1}^{n} p_{u,k}. \quad (7)$$

So $\hat{p}_u$ is the fraction of observations that exceed the upper threshold $u_u$ and similarly for $\hat{p}_l$.

Let $b_n = O_+(a_n)$ denote a sequence such that $b_n/a_n$ is bounded away from zero and infinity. The following result is proved in the appendix and describes the asymptotic distribution of $\hat{a}_M$.

**Theorem 2.1 (Distribution of $\hat{a}_M$)**
Suppose Assumption 2.1 holds and assume that $u_l = -O_+(n^{\alpha\xi_l})$ and $u_u = O_+(n^{\alpha\xi_u})$ for some $\alpha \in (\frac{1}{2\delta_l\xi_l} \vee \frac{1}{2\delta_u\xi_u}, 1)$ and let $\gamma_n^2 = \mathsf{Var}(\hat{\mu})$. Then

$$\frac{1}{\gamma_n k_n}(\hat{a}_M - a) \to_d \mathcal{N}(0,1), \quad \text{as } n \to \infty,$$

where

$$k_n^2 = 1 + \Big[\frac{p_l(1-p_l)}{\gamma_n^2 n}\Big(u_l - \frac{\beta_l}{1-\xi_l}\Big)^2 + \frac{p_l\beta_l^2(1+\xi_l)^2}{n\gamma_n^2(1-\xi_l)^4}\Big]\mathbf{1}_{\{\xi_u \le \xi_l\}}$$
$$+ \Big[\frac{p_u(1-p_u)}{\gamma_n^2 n}\Big(u_u + \frac{\beta_u}{1-\xi_u}\Big)^2 + \frac{p_u\beta_u^2(1+\xi_u)^2}{n\gamma_n^2(1-\xi_u)^4}\Big]\mathbf{1}_{\{\xi_l \le \xi_u\}} = O_+(1).$$

□

Note that, if the lower tail of the errors $\{\epsilon_k\}$ decays faster than polynomially, then we may set the lower threshold $u_l = -\infty$. This means only fitting a model for the upper tail. The theorem will be valid also for this case, with obvious changes. Naturally, similar reasoning is applicable when the upper tail decays faster than polynomially, this time with $u_u = \infty$.

The proof of Theorem 2.1 makes use of a result by Smith (1987). It is possible to use a more direct argument as that in Johansson (2002) and show that the condition on $\alpha$ can be replaced by $\alpha \in ((1 + 2\delta_l\xi_l)^{-1} \vee (1 + 2\delta_u\xi_u)^{-1}, 1)$.

The conditions $u_u = O_+(n^{\alpha\xi_u})$ and $u_l = -O_+(n^{\alpha\xi_l})$ means that $\mathsf{P}(Z_k > u_u)$ and $\mathsf{P}(Z_k < u_l)$ are both proportional to $n^{-\alpha} \to 0$. This means that, as the sample size increases, we use observations farther out in the tails for estimating the tail parameters. This is sometimes referred to as letting the tails speak for themselves. The condition also means that the number of observations on which the tail estimates are based increases with $n$.

# 3 A simulation study

In order to assess the value of the proposed estimation method, two simulations were made based on the model $Y_k = aX_k + \epsilon_k\sigma(X_k)$ in (1) with $a = 1.2$ and different selections of $\epsilon_k$, $X_k$ and $\sigma(\cdot)$.

The parameters $\xi_u$, $\xi_l$, $\beta_u$ and $\beta_l$ were estimated for thresholds $u_l$ and $u_u$ such that $p_u = p_l = 0.03, 0.05, \ldots, 0.15$, where $p_u$ is the probability of an observation exceeding $u_u$ and $p_l$ is ditto for $u_l$. Selecting $p_u = p_l$ is possibly a weakness in the simulations since the underlying distributions of $Y_k/X_k$ were not selected to be symmetric and hence, perhaps these probabilities should differ from one-another. The automatic selection of the best combination of values is difficult however. As a result, the estimates in the simulations might not be of the same quality as those which would be made in a single sample case where the individual sample characteristics could be examined more closely.

The regression coefficient $\hat{a}_M$ was calculated in the following way; if either of $\hat{\xi}_u$ or $\hat{\xi}_l$ were $> 1$, $\hat{a}_M$ was set to infinity since such a tail-index would indicate that the underlying distribution lacks finite mean. Further, if $\hat{\xi}_u < 0$ and $\hat{\xi}_l \in (0, 1)$, we set $u_u = \infty$ and hence only fit a Pareto distribution to the lower tail. This is done since this would indicate a finite upper tail. The case $\hat{\xi}_l < 0$ and $\hat{\xi}_u \in (0, 1)$ is treated analogously. Finally, if both $\hat{\xi}_u$ and $\hat{\xi}_l < 0$, $\hat{a}_M$ was simply the sample mean since $u_u = \infty$ and $u_l = -\infty$ in that case.

The estimate $\hat{a}_M$ was compared to the least-squares estimate $\hat{a}_{LS}$ mentioned in Section 1. Since $\hat{a}_{LS}$ requires knowledge about the shape of $\sigma$, the alternative estimate

$$\hat{a}_{LS}(r) = \frac{\sum_{k=1}^{n} X_k Y_k / X_k^r}{\sum_{k=1}^{n} X_k^2 / X_k^r}, \quad r = 0, 0.2, 0.4, \ldots, 2$$

was used. This made it possible to assess how sensitive the least-squares estimate was to deviations from the assumptions placed on $\sigma$.

A closer examination of the estimator makes it clear that

$$\hat{a}_M \approx \hat{\tau}_l + \hat{\tau}_u + a + \frac{1}{n} \sum_{k=1}^{n} \epsilon_k \frac{\sigma(X_k)}{X_k},$$

where $\hat{\tau}_l$ and $\hat{\tau}_u$ are the lower and upper tail-parts of the estimate respectively and $n$ is the sample size. This means that, in cases where the tail-parts are negligible, $\hat{a}_M$ is expected to behave in approximately the same manner as $\hat{a}_{LS}(2.0)$, since

$$\hat{a}_{LS}(2) = a + \frac{1}{n} \sum_{k=1}^{n} \epsilon_k \frac{\sigma(X_k)}{X_k}.$$

### 3.1 Application to superpopulation sampling

For a population of size $N$, the superpopulation model $Y_k = aX_k + \epsilon_k \sigma(X_k)$ might for instance describe how an individuals actual income, $Y_k$, is related to the declared income $X_k$. In such a case the $\epsilon_k$ might be skewed to the left with a finite left endpoint and a polynomially decreasing right tail. For an overview (and more) of model assisted survey sampling, refer to Särndal et al. (1992).

In this framework, we are interested in estimating the regression coefficient $a$ and the finite population total $T(Y) = \sum_{k \in U} Y_k$, where $U = \{Y_k : k = 1, \ldots, N\}$ denotes the population consisting of $N$ individuals. We will do this by drawing a sample, $S$, of size $n < N$ from $U$ and observing $\{(X_k, Y_k), k \in S\}$. This sample will be used for estimating the unknown parameter $a$ with $\hat{a}$. The estimate of $T(Y)$ is then

$$\hat{T}(Y) = \sum_{k \in S} Y_k + \hat{a} \sum_{k \notin S} X_k.$$

It is desirable to find estimators of $\hat{T}(Y)$ which are robust with respect to changes in the variance structure, $\sigma$, and outlier robust. As discussed earlier, an estimator based on $\hat{a}_M$ may have these properties. We examine the following estimates of the population total

$$\hat{T}_M = \sum_{k \in S} Y_k + \hat{a}_M \sum_{k \notin S} X_k \quad \text{and} \quad \hat{T}_{LS}(r) = \sum_{k \in S} Y_k + \hat{a}_{LS}(r) \sum_{k \notin S} X_k.$$

These were also compared to the simpler estimate

$$\hat{T}_{mean} = \frac{N}{n} \sum_{k \in S} Y_k,$$

where $N$ is the population size and $n$ the size of the sample.

There are other, robust, methods available for estimating the population total $T(Y)$. An overview of the literature can be found in Karlberg (1999), see also Iachan (1984), de Bragança Pereira and Rodrigues (1983) and Chambers (1986). But, as Karlberg (1999) states, these methods are most suitable for situations in which the outliers encountered are non-representative. That is, when the observations are in some way unique or incorrectly recorded (Chambers (1986)). The model (1) above is suitable for populations where outliers are to be expected.

It is important to note that $\hat{a}_M$ does not take the sampling variability into account in Theorem 2.1. Rather, the results are conditional on the sample drawn. This means that the result presently can not be used for calculating confidence intervals in a sampling application. Still, the sampling framework is interesting and the simulations were made in order to assess wether or not this is a useful method for estimating the population total.

In both simulation one and two below, populations of size $N = 100,000$ were created. From these, 50 samples of size 1000 or $10,000$ were drawn without replacement (simple random sampling). This procedure was repeated for ten different populations. No special simulation, more adapted to the regression case, was made. The reason is that those results would be very similar to the ones displayed below.

### 3.2 Simulation one

In simulation one, the $X_k$'s were distributed as $10+1000W_k$, where $W_k \sim$Beta(2,10), which conforms to the assumption above that $\{X_k\} = O_+(1)$. The errors were selected as

$$\epsilon_k =_d \begin{cases} GPD(\xi = 0.7, \beta = 0.1) & \text{with prob } 1/3 \\ -GPD(\xi = 0.4, \beta = 0.1) & \text{with prob } 2/3 \end{cases}$$

making $\mathsf{E}[\epsilon_k] = 0$. The four different variance structures $\sigma(x) = 1$, $\sqrt{x}$, $x$ and $x^2$, were all investigated. The results are found in Figures 1 - 4, The boxplots were made using MatLab's default settings. In order to emphasize the boxes themselves, not all outliers are displayed in the graphs.

For the case $\sigma(x) = 1$ the estimates of $\xi_l$ and $\xi_u$ fluctuated quite a bit for sample size 1000, which contributed largely to the variance of $\hat{a}_M$. In this setting, the errors $\epsilon_k$ are very small compared to the $X_k$'s. Still, both estimators are quite accurate with errors in the third decimal place. For both sample sizes, $\hat{T}_{mean}$ displayed a considerably larger variance than was the case for $\hat{T}_M$ and $\hat{T}_{LS}$.

For $\sigma(x) = \sqrt{x}$ the errors have a larger influence. It is possible that the heavier right tail accounts for $\hat{a}_M$ having less of a median bias in the large sample case. In this case $\hat{T}_{mean}$ varies substantially more than the other methods, this may also be due to the heavy right tail.

With $\sigma(x) = x$, it is expected that $\hat{a}_M$ should perform very well. This is confirmed in Figure 3 below. Finally, in the case $\sigma(x) = x^2$, the $X_k$'s are drowned by the very large errors $\epsilon_k X_k^2$, which is noted in the large variation of the results. Here $\hat{a}_M$ is a competitive method, even though the results are generally quite far off the mark.

### 3.3 Simulation two

In the second simulation $\epsilon_k \sim \mathcal{N}(0,1)$ and $X_k \sim Exp(2)$. The same four variance structures as in simulation one were used, namely $\sigma(x) = 1$, $\sqrt{x}$, $x$ and $x^2$. The reason this model was selected is that the different $\sigma$'s will illustrate some qualitatively different situations.

A comment about the distributions of $\epsilon_k$ and $X_k$ is in order. In the previous section the $\epsilon_k$ were assumed to have polynomially decaying tails and $\sigma(X_k)/X_k$ was assumed to be bounded away from zero and infinity. This is at odds with the model above. However, the important thing is that $\epsilon_k \sigma(X_k)/X_k$ has a distribution which is heavy-tailed. This will be the case for $\sigma(x) = 1$ and $\sqrt{x}$. The cases $\sigma(x) = x$ and $x^2$ are included here just to see what happens when these assumptions are not valid.

The results from the simulation are shown in Figures 5–8.

It is clear from Figure 5 that estimation of $a$ using $\hat{a}_M$ is not applicable in the case $\sigma(x) = 1$ since both $\xi_l$ and $\xi_u$ are mainly estimated to be $> 1$, indicating an infinite mean. In this case $\hat{a}_{LS}(r)$ was the preferable method.

Figure 6 shows the results of the simulation using $\sigma(x) = \sqrt{x}$. With estimates of $\xi_u$ and $\xi_l$ in $(0,1)$, $\hat{a}_M$ is an applicable method. However, in this case $\hat{a}_{LS}(r)$ outperforms $\hat{a}_M$ across the whole range of $r$ values, $\hat{a}_M$ being comparable only to $\hat{a}_{LS}(2.0)$. It is interesting to note how stable the least-squares estimate is with regards to changes in assumptions about $\sigma$. It is also clear that the simpler estimate $\hat{T}_{mean}$ does very well in this case.

In the simulation using $\sigma(x) = x$, it is quite clear from the outset that $\hat{a}_M$ is not the method of choice since the assumptions on the errors are not met. Studying the results, note that $\hat{\xi}_l$ and $\hat{\xi}_u$ are both estimated at $< 0$. Hence, $\hat{a}_M$ is simply the sample mean of the $Y_k/X_k$. Here, the sample mean outperforms the maximum-likelihood estimates over the entire range of $r$ values, except for $r = 2$, as expected. Further, $\hat{T}_M$ outperforms $\hat{T}_{mean}$ in this case.

The results of the simulation using $\sigma(x) = x^2$, are presented in Figure 8. In this case the tails are not very heavy, as seen from $\hat{\xi}_l$ and $\hat{\xi}_u$. It is clear that $\hat{a}_M$ is preferable over using $\hat{a}_{LS}(r)$, except possibly for the optimal choice $\hat{a}_{LS}(2.0)$. This carries over to estimation of the population total where $\hat{T}_M$ also gives better results than $\hat{T}_{mean}$.

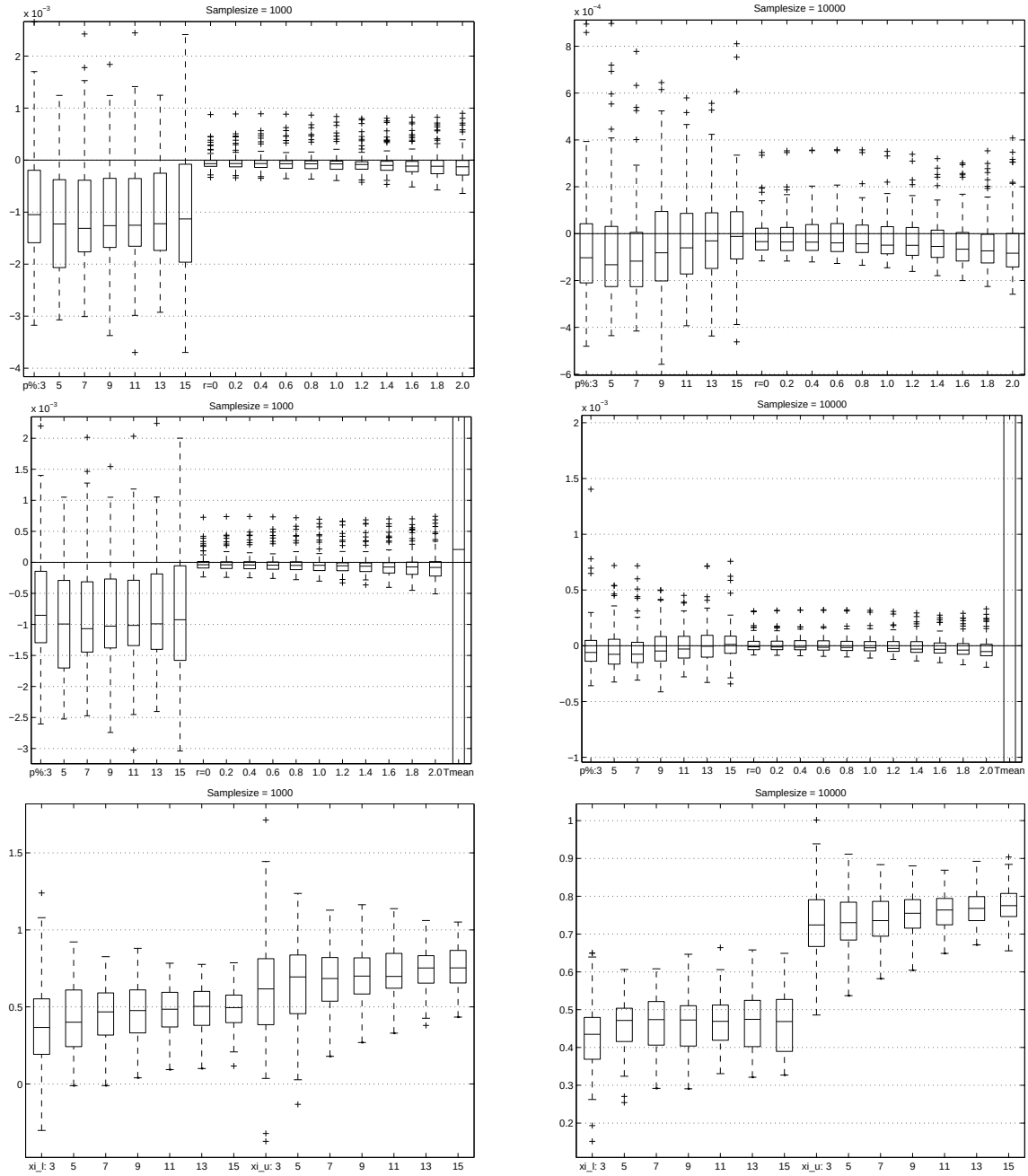


Figure 1: Simulation one with $\sigma(x) = 1$. Left: sample size 1000. Right: sample size 10 000. The upper graphs display $\hat{a} - a$ where $a = 1.2$. The seven leftmost columns are $\hat{a}_M(p)$ for $p = p_u = p_l = 0.03, \ldots, 0.15$. The remaining columns show $\hat{a}_{LS}(r)$ for $r = 0, 0.2, \ldots, 2$. The middle graphs show $(\hat{T} - T)/T$, where $T$ is the true population total. The columns correspond to the ones in the top graphs except the rightmost one which displays $\hat{T}_{mean}$. The bottom graphs are the estimates of $\xi_l$ (left) and $\xi_u$ (right) for different values of $p = p_l = p_u$ (%). $T_{mean}$ had an inter quartile range of 0.08 for sample size 1000 and 0.02 for sample size 10000.

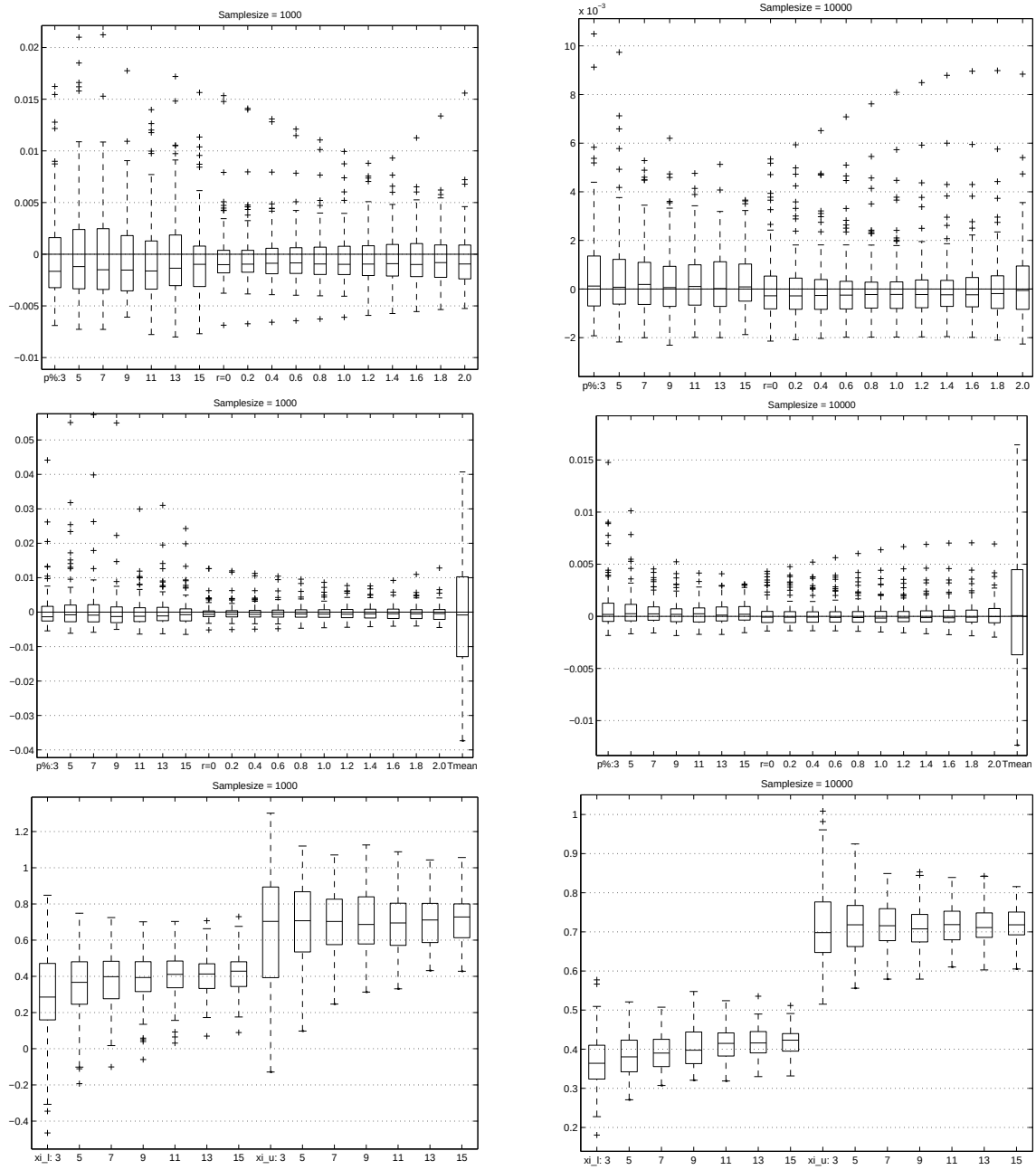


Figure 2: Simulation one with $\sigma(x) = \sqrt{x}$. The graphs are arranged in the same way as in Figure 1.

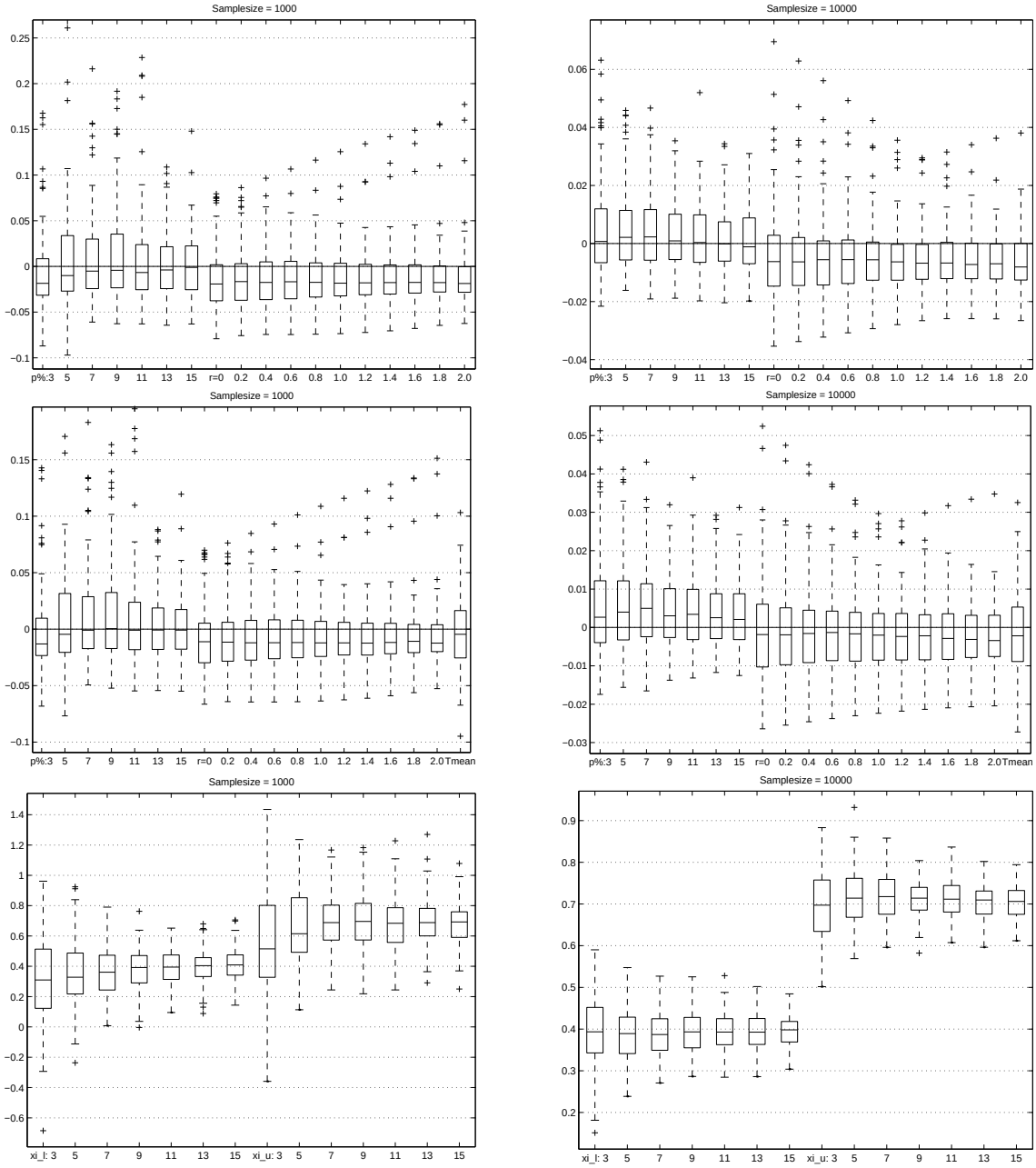


Figure 3: Simulation one with $\sigma(x) = x$. The graphs are arranged in the same way as in Figure 1.

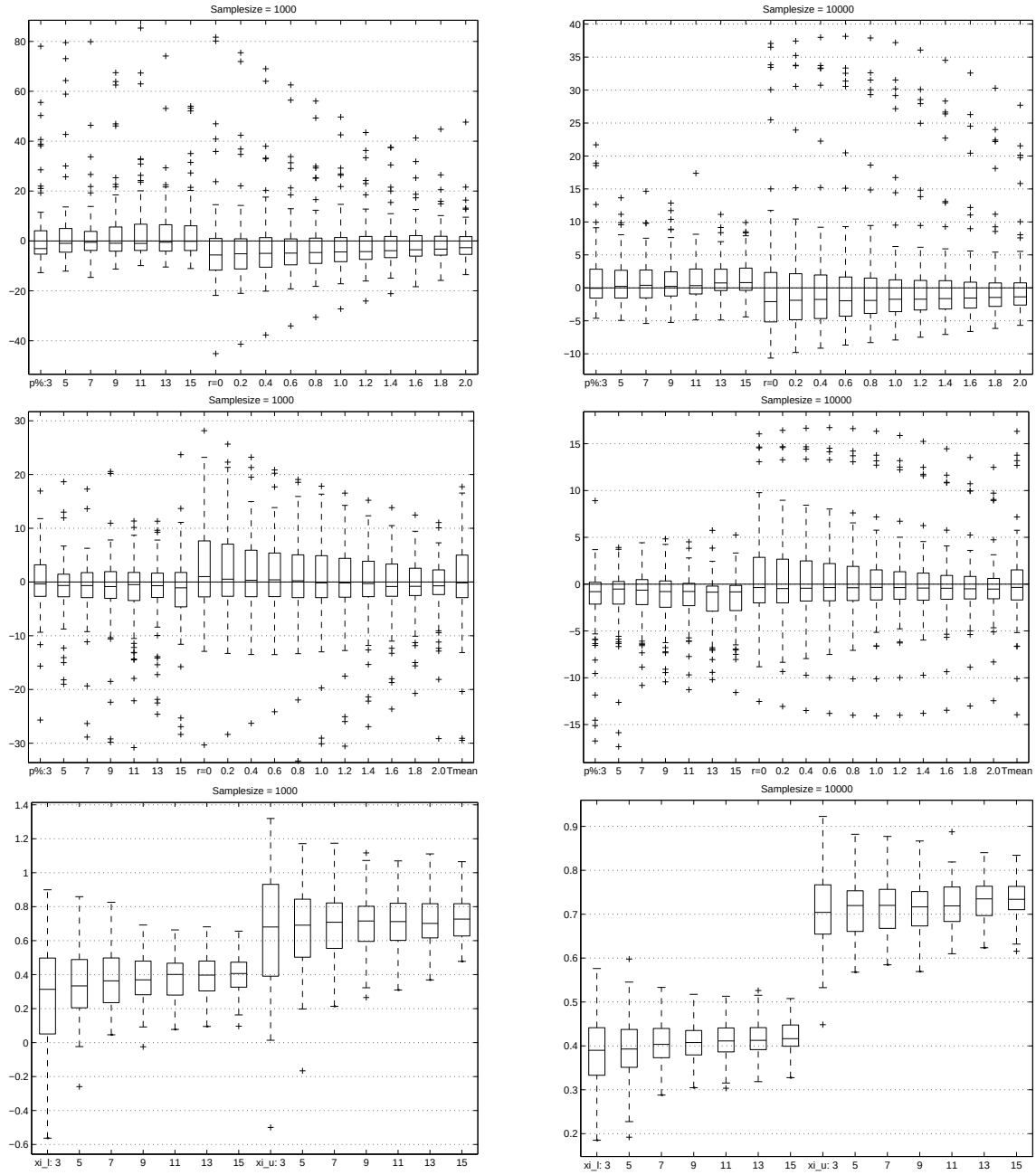


Figure 4: Simulation one with $\sigma(x) = x^2$. The graphs are arranged in the same way as in Figure 1. There were outliers of the order $\pm 500$ for all estimates of $a$ with sample size 1000. For sample size 10 000, the outliers were larger for $\hat{a}_{LS}$ than for $\hat{a}_M$.

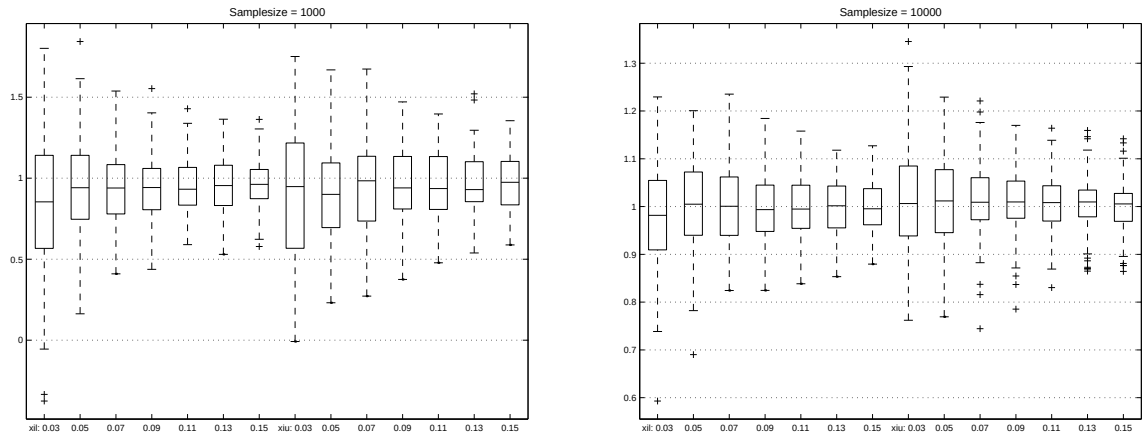


Figure 5: Simulation two with $\sigma(x) = 1$. The seven leftmost columns are $\xi_l(p_l)$ for $p_l = 0.03, \ldots, 0.15$, starting with $\xi_l(0.03)$ to the far left. The following boxplots are $\xi_u(p_u)$ for $p_u = 0.03, \ldots, 0.15$ with $\xi_u(0.15)$ being the rightmost column. The left graph shows the results for sample size 1000 and the right graph shows the results for sample size 10 000.

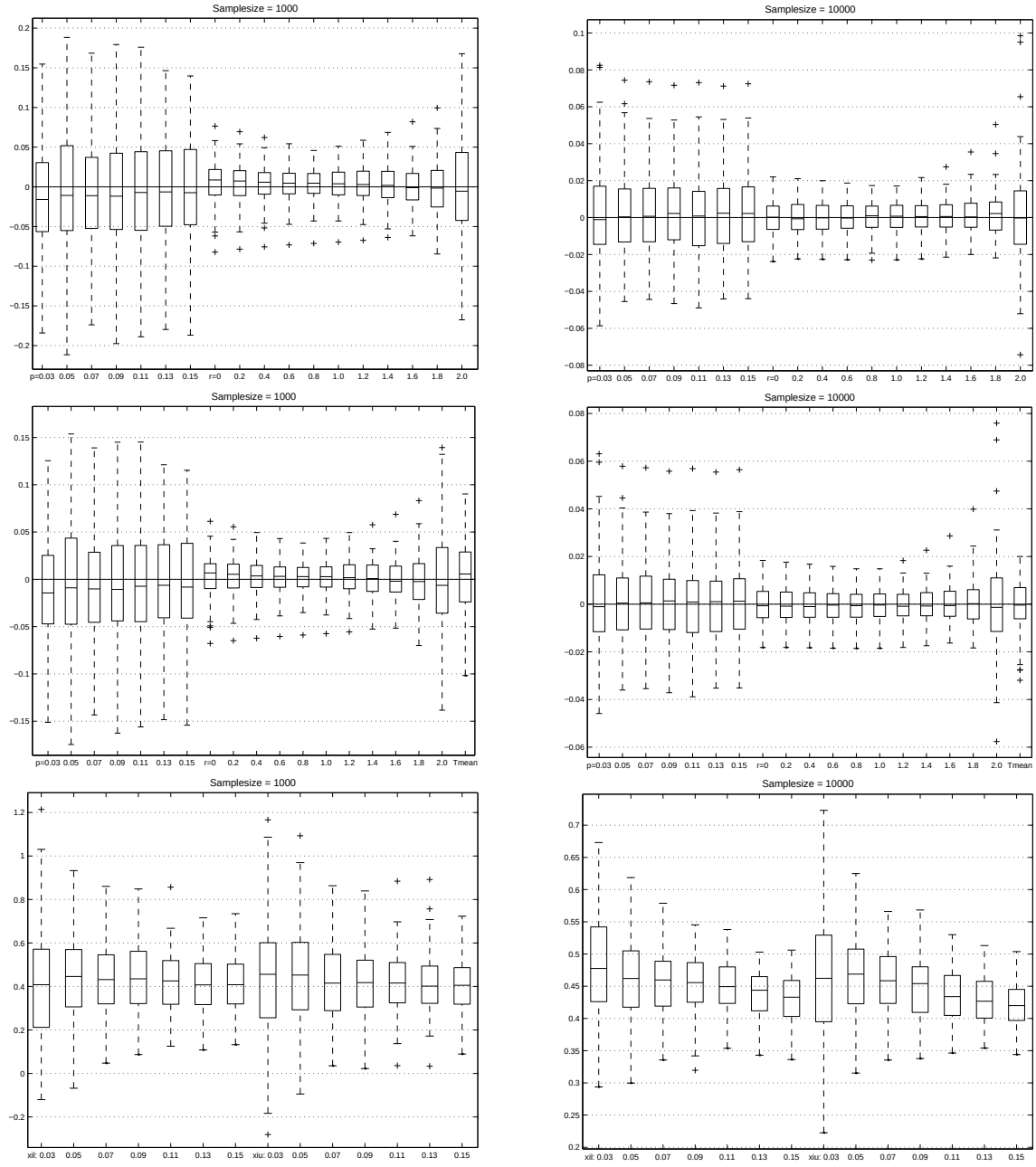


Figure 6: Simulation two with $\sigma(x) = \sqrt{x}$. Left: sample size 1000. Right: sample size 10 000. The upper graphs display $\hat{a} - a$ where $a = 1.2$. The seven leftmost columns are $\hat{a}_M(p)$ for $p = p_u = p_l = 0.03, \ldots, 0.15$. The following columns show $\hat{a}_{LS}(r)$ for $r = 0, 0.2, \ldots, 2$. The middle graphs show $(\hat{T} - T)/T$, where $T$ is the true population total. The columns correspond to the ones in the top graphs except the rightmost one which displays $\hat{T}_{mean}$. The bottom graphs are the estimates of $\xi_l$ (left) and $\xi_u$ (right) as in Figure 5.

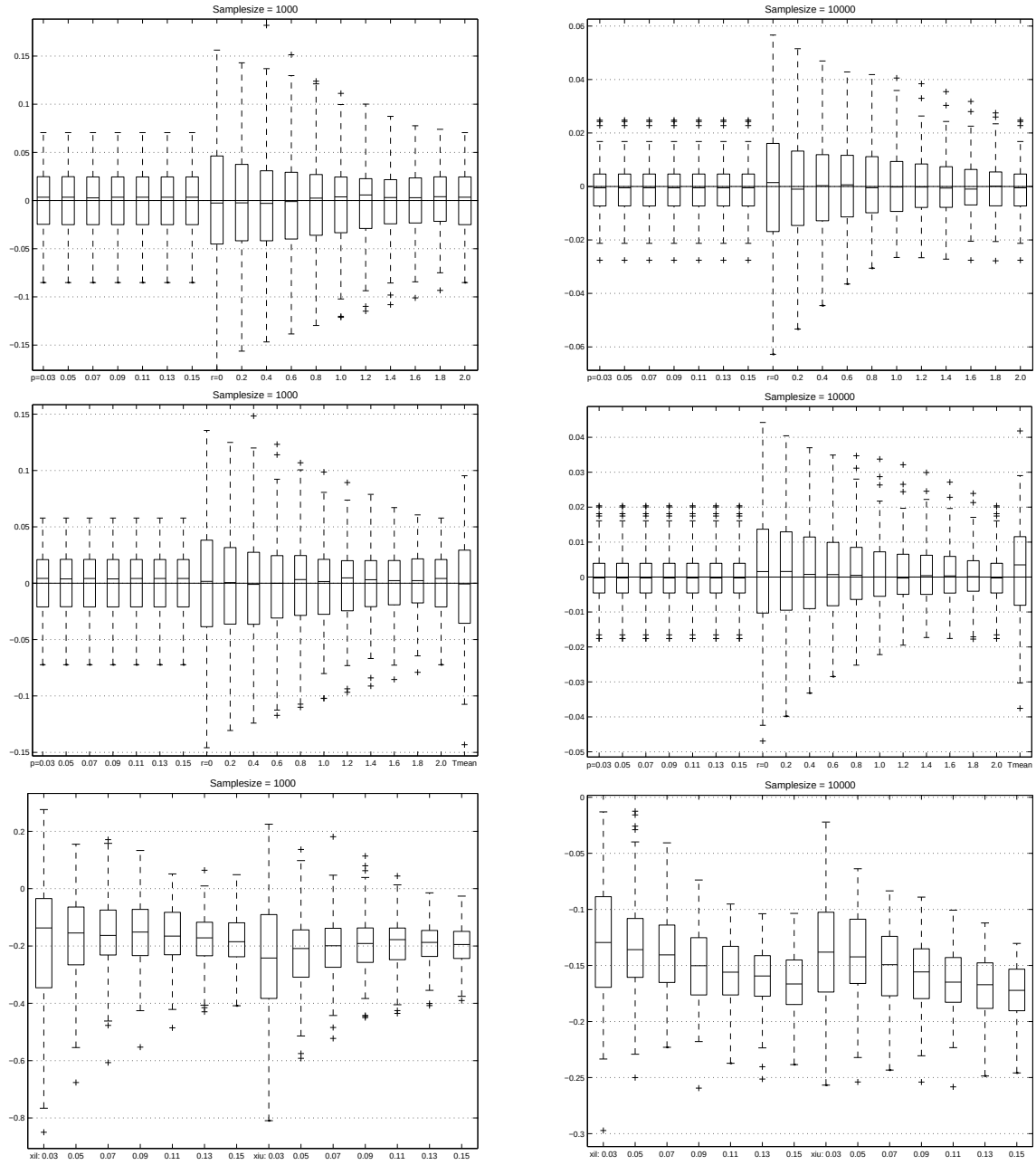


Figure 7: Simulation two with $\sigma(x) = x$. The graphs are arranged in the same way as in Figure 6.

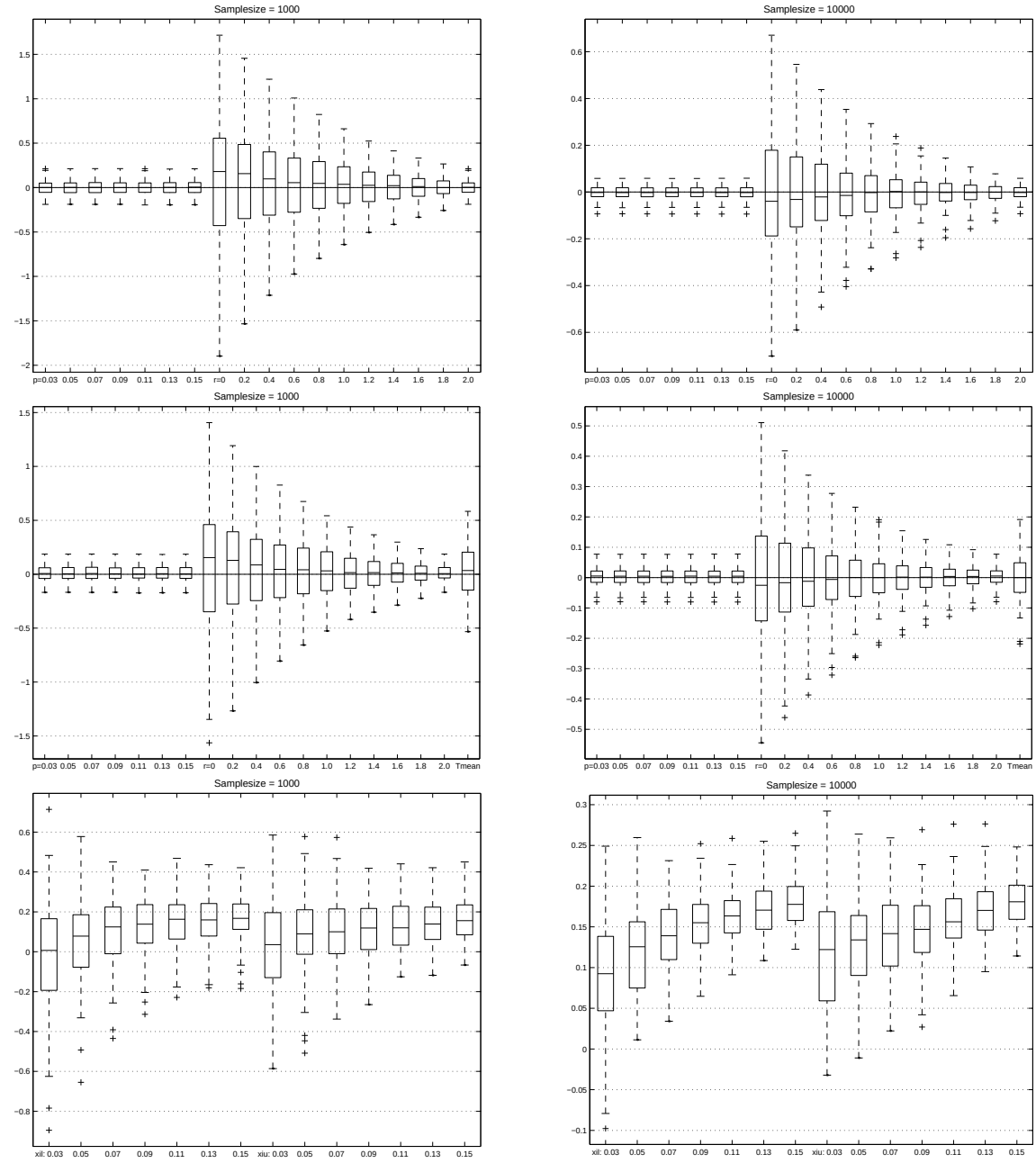


Figure 8: Simulation two with $\sigma(x) = x^2$. The graphs are arranged in the same way as in Figure 6.

# 4 Discussion and conclusions

There are some difficulties judging the behaviour of the alternative estimate $\hat{a}_M$ from simulations. The first is that it is not easy to automate the selection of $p_l$ and $p_u$ so as to generate the best possible estimates of the tail parameters. A second problem is in the selection of distribution of $\epsilon_k$ and $X_k$. Here, an infinite number of possibilities exist and the behaviour of one such combination in comparison with another may be quite different.

In the sampling application, a third difficulty is which alternative estimates to compare with. There are a number of different outlier-robust estimates on the market, for instance the one- and two-sided windsorization-based estimators discussed in Kokic and Smith (1998b,a) and Chambers and Kokic (1993). Some other methods are discussed in Karlberg (1999).

As is often the case, no one method consistently outperforms another. However, as could be seen from the simulations, there are cases where the alternative estimators $\hat{a}_M$ and $\hat{T}_M$ perform well in comparison to the competitors $\hat{a}_{LS}(r)$ and $\hat{T}_{LS}(r)$ and to $\hat{T}_{mean}$. The sensitivity to the selection of $X_k$'s and $\sigma(\cdot)$ is also well illustrated by the simulations, although the least-squares estimate performs quite well even when $\sigma$ is slightly misspecified. One of the advantages of using $\hat{a}_M$ is that no knowledge about $\sigma(\cdot)$ is required.

Another advantage of the proposed method is that the asymptotic distribution of $\hat{a}_M$ is known, which can be used for constructing confidence intervals. For the least squares estimate $\hat{a}_{LS}$, on the other hand, the asymptotic distribution would be more difficult to find. There are subsampling schemes available which generate consistent estimators of $a$. However, these are believed to be less efficient than asymptotic methods such as these investigated in this paper (Hall and Jing, 1998).

Judging from the above results, a pragmatic course of action suggests itself. If the distribution of $Y_k/X_k$ looks heavy-tailed and seems to have a finite mean, then $\hat{a}_M$ might be a good alternative to $\hat{a}_{LS}$, especially if nothing is known about the variance structure.

## 4.1 Possible extensions

Some improvements could be made on the estimate $\hat{a}_M$. For instance one concern is how useful the expression for the asymptotic variance is. Assume that $\xi_u = \xi_l - \epsilon$ for some small $\epsilon > 0$. Then the tails are almost balanced so both the contribution to the variance from the upper and lower tails should be taken into consideration, especially for small sample sizes. Hence, there might be a need to find a more suitable method for estimating the variance for finite samples. One way of doing this is to replace the indicator functions in the expression for $k_n$ in Theorem 2.1 with the corresponding $o_p(1)$-terms from the proof.

It might also be possible to obtain a better estimate by replacing $\hat{\mu}$ with a weighted average like

$$\hat{\mu}_w = \sum_{k=1}^{n} w_k Z_k \mathbf{1}_{\{u_l < Z_k < u_u\}},$$

where the weights $w_k$ are proportional to the inverse of the variances. The drawback of this is that such a method would require knowledge about the characteristics of $\sigma$.

A third possible extension is to replace the condition $\{\sigma_k\} = O_+(1)$ by $\{\sigma_k\} = O_+(n^r)$ for $r$ in a suitable interval. This would add generality to the model and could probably be done without affecting the proofs very much.

For the sampling application outlined in the simulations, a natural extension is to include the sampling variability into the estimate. How this should be done is not clear at the present time. Another interesting extension is to try to apply similar methodology to the more general regression case $Y_k = a_1 + a_2 X_k + \epsilon_k \sigma(X_k)$, where $a_1$ and $a_2$ are unknown constants.

# A Proof of Theorem 2.1

**Lemma A.1 (Rates of convergence)**
Let $p_{u,k} = \mathsf{P}(Z_k > u_u)$, $p_{l,k} = \mathsf{P}(Z_k < u_l)$ and $\gamma_k = \mathsf{Var}(Z_k \mathbf{1}_{\{u_l < Z_k < u_u\}})$. Then

$$p_{u,k} = O_+(u_u^{-1/\xi_u}), \quad p_{l,k} = O_+(|u_l|^{-1/\xi_l}) \quad \text{and} \quad \gamma_k = O_+(|u_l|^{2-1/\xi_l} \vee u_u^{2-1/\xi_u}).$$

Proof: By Assumption 2.1 $\{\sigma_k\} = O_+(1)$, and the results follow from $\mathsf{E}[\eta^k] = \int_0^\infty k x^{k-1} \mathsf{P}(\eta > x) dx$ for a positive random variable $\eta$ and $k = 1, 2, 3, \ldots$. □

**Lemma A.2 (Distribution of $\hat{\mu}$)**
Suppose that Assumption 2.1 holds and that the thresholds $u_l = -O_+(n^{\alpha\xi_l})$ and $u_u = O_+(n^{\alpha\xi_u})$ for some $\alpha \in (0, 1)$. Let

$$\hat{\mu} = \frac{1}{n} \sum_{k=1}^{n} Z_k \mathbf{1}_{\{u_l < Z_k < u_u\}}, \quad \mu_n = \mathsf{E}[\hat{\mu}], \quad \text{and} \quad \gamma_n^2 = \mathsf{Var}(\hat{\mu}).$$

Then

$$(\hat{\mu} - \mu_n)/\gamma_n \to_d \mathcal{N}(0, 1) \quad \text{as } n \to \infty.$$

Proof: The proof is a direct application of the Lindeberg-Feller theorem. Hence, it suffices to check that the following tends to zero for every $\epsilon > 0$:

$$\begin{aligned}
&\frac{1}{\gamma_n^2} \sum_{k=1}^{n} \mathsf{E}\Big[\Big(\frac{Z_k \mathbf{1}_{\{u_l < Z_k < u_u\}} - \mu_n}{n}\Big)^2; \Big|\frac{Z_k \mathbf{1}_{\{u_l < Z_k < u_u\}} - \mu_n}{n}\Big| > \epsilon \gamma_n\Big] \\
&\quad \le \frac{1}{\gamma_n^2} \sum_{k=1}^{n} \mathsf{E}\Big[\Big(\frac{Z_k \mathbf{1}_{\{u_l < Z_k < u_u\}} - \mu_n}{n}\Big)^4\Big]^{1/2} \mathsf{P}(|Z_k \mathbf{1}_{\{u_l < Z_k < u_u\}} - \mu_n| > \epsilon n \gamma_n)^{1/2} = (*).
\end{aligned}$$

Since $|Z_k \mathbf{1}_{\{u_l < Z_k < u_u\}} - \mu_n| < C(n^{\alpha\xi_l} \vee n^{\alpha\xi_u})$ a.s. for some constant $C$ and, by Lemma A.1, $n\gamma_n \propto n^{1-\alpha/2}(n^{\alpha\xi_l} \vee n^{\alpha\xi_u})$, it follows that there exists a constant $n_0$ such that $(*) = 0$ for all $n > n_0$. □

**Lemma A.3 (Expected values, tail components)**
For the upper tail, let $Y \sim F_{u,k}$, where $F_{u,k}(y)$ is given by equations (2) and (3) above. Further, let $w_k = (u_u - a)/\sigma_k$. Then for $u_u \to \infty$

$$\begin{aligned}
\mathsf{E}\Big[\Big(1 + \frac{Y}{u_u - a}\Big)^{-r}\Big] &\sim \frac{1}{1 + r\xi_u} + \frac{1}{1 + r\xi_u} \cdot \frac{r\delta_u \xi_u^2}{(r + \delta_u)\xi_u + 1} w_k^{-\delta_u} L_u(w_k) \\
\mathsf{E}\Big[\ln\Big(1 + \frac{Y}{u_u - a}\Big)\Big] &\sim \xi_u - \xi_u \frac{\delta_u \xi_u}{1 + \delta_u \xi_u} w_k^{-\delta_u} L_u(w_k)
\end{aligned}$$

Similarly, for the lower tail, let $Y \sim F_{l,k}$, where $F_{l,k}(y)$ is given by equations (4) and (5) above. Further, let $w_k = (a - u_l)/\sigma_k$, then for small enough $u_l$ such that $a - u_l > 0$ ($u_l \to -\infty$),

$$\begin{aligned}
\mathsf{E}\Big[\Big(1 + \frac{Y}{a - u_l}\Big)^{-r}\Big] &\sim \frac{1}{1 + r\xi_l} + \frac{1}{1 + r\xi_l} \cdot \frac{r\delta_l \xi_l^2}{(r + \delta_l)\xi_l + 1} w_k^{-\delta_l} L_l(w_k) \\
\mathsf{E}\Big[\ln\Big(1 + \frac{Y}{a - u_l}\Big)\Big] &\sim \xi_l - \xi_l \frac{\delta_l \xi_l}{1 + \delta_l \xi_l} w_k^{-\delta_l} L_l(w_k).
\end{aligned}$$

Proof: This follows from straight-forward calculations using Karamata's Theorem (see e.g. Bingham et al. (1987)). Similar calculations have been done in Smith (1987) and Johansson (2002). □

**Lemma A.4 (Distribution of $(\hat{\beta}_l, \hat{\xi}_l)$ and $(\hat{\beta}_u, \hat{\xi}_u)$)**
Let $r \in \{u, l\}$ denote either the upper-tail parameters ($u$) or the lower-tail ($l$). Then, for maximum likelihood estimates $(\hat{\beta}_{u_r}, \hat{\xi}_{u_r})$ of the parameters in (6),

$$\sqrt{n_r} Q_r^{1/2} \begin{pmatrix} \hat{\beta}_{n_r} - \beta_r \\ \hat{\xi}_{n_r} - \xi_r \end{pmatrix} \to_d \mathcal{N}(0, I), \quad \text{as } n_r \to \infty,$$

where

$$Q_r^{-1} = (1 + \xi_r) \begin{pmatrix} 2\beta_r^2 & -\beta_r \\ -\beta_r & 1 + \xi_r \end{pmatrix},$$

under the condition that $\sqrt{n_r} w_r^{-\delta_r} L_r(w_r) \to 0$, where $w_u = (u_u - a)/\sigma_k$, $w_l = (a - u_l)/\sigma_k$ and $x^{-\delta_r} L_r(x)$ is non-increasing. This is true, for instance, when $u_r = O_+(n^{\alpha_r \xi_r})$, $\alpha_r \in (1/2\delta_r \xi_r, 1)$.

Proof: If the perturbations $R_{l,k} = R_{u,k}$ (in Equations (5) and (3)) for all $k$ then this is Theorem 3.2 in Smith (1987). Otherwise, it follows by extension of this theorem, using Lemma A.3. Observe that, as defined here, the exceedances below $u_l$ are positive random variables as in Equation (4). □

**Lemma A.5 (Distribution of $(\hat{\beta}_{N_l}, \hat{\xi}_{N_l})$ and $(\hat{\beta}_{N_u}, \hat{\xi}_{N_u})$)**
Let $N_u = |\{i : Z_i > u_u\}|$ and $N_l = |\{i : Z_i < u_l\}|$, then

$$\sqrt{np_r} Q_r^{1/2} \begin{pmatrix} \hat{\beta}_{N_r} - \beta_r \\ \hat{\xi}_{N_r} - \xi_r \end{pmatrix} \to_d \mathcal{N}(0, I), \quad \text{as } n \to \infty,$$

where $r \in \{l, u\}$ and $Q$ is the matrix in Lemma A.4.

Proof: Let

$$\phi_{u|N_u}(t_1, t_2) = \mathsf{E}[\exp\{i\sqrt{np_u} Q_u^{1/2}(t_1, t_2) \begin{pmatrix} \hat{\beta}_{N_u} - \beta_u \\ \hat{\xi}_{N_u} - \xi_u \end{pmatrix}\} | N_u].$$

Then Lemma A.4 tells us that $\phi_u(t_1, t_2) \to \exp\{-(t_1^2 + t_2^2)/2\}$ as $n \to \infty$. This means that, for some constant $n_0$ and $\delta > 0$, we have

$$\begin{aligned} \mathsf{P}(|\phi_{u|N_u} - \phi_u| < \delta) &= \mathsf{P}(|\phi_{u|N_u} - \phi_u| < \delta, N_u > n_0) \\ &\quad + \mathsf{P}(|\phi_{u|N_u} - \phi_u| < \delta, N_u \le n_0) \\ &= \mathsf{P}(N_u > n_0) + \mathsf{P}(|\phi_{u|N_u} - \phi_u| < \delta, N_u \le n_0) \to 1, \end{aligned}$$

as $n \to \infty$, since $N_u \to_p \infty$. Hence $\phi_{u|N_u}(t_1, t_2) \to_p \exp\{-(t_1^2 + t_2^2)/2\}$ and the claim follows by taking expectations. The calculations are analogous for the lower tail parameters. □

We need to estimate $p_u$ and $p_l$. The natural estimates are $\hat{p}_u = N_u/n$, where $N_u$ is the number of observations exceeding $u_u$. Similarly $\hat{p}_l = N_l/n$ is used for estimating $p_l$, where $N_l$ is the number of observations less than $u_l$.

**Lemma A.6 (Distribution of $\hat{p}_u$ and $\hat{p}_l$)**
Let $p_r$, $r \in \{u, l\}$ be defined as in (7). Then

$$\Big(\frac{\hat{p}_l - p_l}{\sqrt{\frac{1}{n}p_l(1-p_l)}}, \frac{\hat{p}_u - p_u}{\sqrt{\frac{1}{n}p_u(1-p_u)}}\Big)^t \to_d \mathcal{N}(0, I),$$

where $I$ is the identity matrix.

Proof: We sketch the proof for the upper tail, $p_u$. Let

$$\xi_{k,n} = \frac{1}{n}(\mathbf{1}_{\{\epsilon_k > \frac{u_u - a}{\sigma_k}\}} - p_{k,u}), \text{ where } p_{k,u} = \mathsf{P}\Big(\epsilon_k > \frac{u_u - a}{\sigma_k}\Big)$$

and the distribution of $\epsilon_k$ is given in Assumption 2.1. Then $\mathsf{E}[\xi_{k,n}] = 0$ and $\mathsf{E}[\xi_{k,n}^2] = n^{-2}p_{k,n}(1-p_{k,n})$. Note that

$$\frac{\mathsf{Var}\Big(\sum_{k=1}^n \xi_{k,n}\Big)}{n^{-1}p_u(1-p_u)} = \frac{1 - n^{-1}\sum_{k=1}^n p_{k,u}^2/p_u}{1-p_u} \to 1,$$

if

$$n^{-1}\sum_{k=1}^n p_{k,u}^2/p_u \le \max_{1\le k\le n} p_{k,u} = \max_{1\le k\le n} \mathsf{P}(\epsilon_k > (u_u - a)/\sigma_k) \to 0,$$

which it does since the $\sigma_k$ are bounded by assumption. Now applying the Lindeberg-Feller theorem to

$$\frac{\sum_{k=1}^n \xi_{k,n}}{\sqrt{n^{-1}p_u(1-p_u)}},$$

it turns out that

$$\begin{aligned}
&\sum_{k=1}^n \mathsf{E}\Big[\Big(\frac{\xi_{k,n}}{\sqrt{n^{-1}p_u(1-p_u)}}\Big)^2; \Big|\frac{\xi_{k,n}}{\sqrt{n^{-1}p_u(1-p_u)}}\Big| > \epsilon\Big] \\
&\quad \le \frac{1}{np_u(1-p_u)}\sum_{k=1}^n \mathsf{P}(|\mathbf{1}_{\{\epsilon_k > \frac{u_u-a}{\sigma_k}\}} - p_{k,u}| > \sqrt{np_u(1-p_u)}\epsilon) \to 0
\end{aligned}$$

since $np_u = O_+(n^{1-\alpha_u}) \to \infty$, as $n \to \infty$, where $\alpha_u \in (0,1)$ by assumption. In fact, the sum is exactly zero for $n$ large enough. The computations are similar for the lower tail. The joint convergence follows from a standard Cramér-Wold argument. This completes the proof. □

**Lemma A.7 (Joint distribution)**
Use the above notation and let $u_l = -O_+(n^{\alpha\xi_l})$ and $u_u = n^{\alpha\xi_u}$ for $\alpha \in ((2\delta_l\xi_l)^{-1} \vee (2\delta_u\xi_u)^{-1}, 1)$. Further let $A$ be the block diagonal matrix with diagonal elements

$$\Big[\sqrt{np_u}Q_u^{1/2}, \sqrt{np_l}Q_l^{1/2}, \frac{1}{\gamma}, \sqrt{\frac{n}{p_u(1-p_u)}}, \sqrt{\frac{n}{p_l(1-p_l)}}\Big]$$

and

$$\Theta = \left[\hat{\beta}_{N_u} - \beta_u, \quad \hat{\xi}_{N_u} - \xi_u, \quad \hat{\beta}_{N_l} - \beta_l, \quad \hat{\xi}_{N_l} - \xi_l, \quad \hat{\mu} - \mu, \quad \hat{p}_u - p_u, \quad \hat{p}_l - p_l\right]^t.$$

Then

$$A\Theta \to_d \mathcal{N}(0, I), \quad \text{as } n \to \infty.$$

Proof: Let, for $r \in \{u, l\}$,

$$\begin{aligned}
\phi_{r|N_r}(t_1, t_2) &= \mathsf{E}[\exp\{i\sqrt{np_u}Q_r^{1/2}(t_1, t_2)\begin{pmatrix}\hat{\beta}_{N_r} - \beta_r \\ \hat{\xi}_{N_r} - \xi_r\end{pmatrix}\}|N_r] \\
\phi_{p_r|N_r}(t) &= \mathsf{E}[\exp\{it\frac{\sqrt{n}(\hat{p}_r - p_r)}{\sqrt{p_r(1-p_r)}}\}|N_r] \\
&= \exp\{it\frac{\sqrt{n}(\hat{p}_r - p_r)}{\sqrt{p_r(1-p_r)}}\} = \phi_{p_r}(t) \\
\phi_{\mu|N_u,N_l}(t) &= \mathsf{E}[\exp\{it\frac{\hat{\mu} - \mu}{\gamma}\}|N_u, N_l],
\end{aligned}$$

where $\gamma^2 = \mathsf{Var}(\hat{\mu})$. Then, using independence conditional on $N_u$ and $N_l$, the joint characteristic function is

$$\phi(t|N_u, N_l) = \phi_{u|N_u}(t_1, t_2)\phi_{l|N_l}(t_3, t_4)\phi_{\mu|N_u,N_l}(t_5)\phi_{p_u|N_u}(t_6)\phi_{p_l|N_l}(t_7).$$

By Lemmas A.2 – A.6

$\phi(t|N_u, N_l) \to_d$
$\exp\{-(t_1^2 + t_2^2)/2\}\exp\{-(t_3^2 + t_4^2)/2\}\exp\{-t_5^2/2\}\exp\{it_6\mathcal{N}(0,1)\}\exp\{it_7\mathcal{N}(0,1)\}$,

and the claim follows by taking expectations. □

Now all pieces that are needed for proving Theorem 2.1 are in place.

Proof:(Theorem 2.1) First recall that

$$a = \frac{1}{n}\sum_{k=1}^{n}\mathsf{E}[Z_k]$$

where, using previously introduced notation and writing $F_{Z_k}(z) = \mathsf{P}(Z_k \le z)$ and $\mu_k = \mathsf{E}[Z_k\mathbf{1}_{\{u_l < Z_k < u_u\}}]$,

$$\begin{aligned}
\mathsf{E}[Z_k] &= \int_{-\infty}^{u_l} z dF_{Z_k}(z) + \int_{u_l}^{u_u} z dF_{Z_k}(z) + \int_{u_u}^{\infty} z dF_{Z_k}(z) \\
&= \int_{-\infty}^{u_l} z d\bar{F}_{l,k}(u_l - z)p_{l,k} + \mathsf{E}[Z_k\mathbf{1}_{\{u_l < Z_k < u_u\}}] + \int_{u_u}^{\infty} z dF_{u,k}(z - u_u)p_{u,k} \\
&= p_{l,k}\Big(u_l - \int_0^{\infty}\bar{F}_{l,k}(z)dz\Big) + \mu_k + p_{u,k}\Big(u_u + \int_0^{\infty}\bar{F}_{u,k}(z)dz\Big)
\end{aligned}$$

Using Karamata's theorem, for small $u_l$ and large $u_u$, we find that

$$\int_0^\infty \bar{F}_{r,k}(z)dz \sim \frac{\beta_r}{1-\xi_r} + R_{r,k}\Big[\frac{\beta_r}{1-(1-\delta_r)\xi_r} - \frac{\beta_r}{1-\xi_r}\Big], \quad r \in \{l,u\}$$

where

$$R_{l,k} = \frac{\Big(\frac{a-u_l}{\sigma_k}\Big)^{-\delta_l} L_l\Big(\frac{a-u_l}{\sigma_k}\Big)}{1+\Big(\frac{a-u_l}{\sigma_k}\Big)^{-\delta_l} L_l\Big(\frac{a-u_l}{\sigma_k}\Big)} \quad \text{and} \quad R_{u,k} = \frac{\Big(\frac{u_u-a}{\sigma_k}\Big)^{-\delta_u} L_u\Big(\frac{u_u-a}{\sigma_k}\Big)}{1+\Big(\frac{u_u-a}{\sigma_k}\Big)^{-\delta_u} L_u\Big(\frac{u_u-a}{\sigma_k}\Big)}.$$

Hence

$$a \sim p_l\Big(u_l - \frac{\beta_l}{1-\xi_l}\Big) + \mu + p_u\Big(u_u + \frac{\beta_u}{1-\xi_u}\Big) + R,$$

where

$$R = \Big[\frac{\beta_u}{1-(1-\delta_u)\xi_u} - \frac{\beta_u}{1-\xi_u}\Big]\frac{1}{n}\sum_{k=1}^n p_{u,k}R_{u,k} - \Big[\frac{\beta_l}{1-(1-\delta_l)\xi_l} - \frac{\beta_l}{1-\xi_l}\Big]\frac{1}{n}\sum_{k=1}^n p_{l,k}R_{l,k}.$$

Using Taylor expansion and

$$S_r = \beta_r \sum_{k=2}^\infty \frac{(-1)^k(\hat{\xi}_{N_r}-\xi_r)^k}{(1-\xi_r)^{k+1}} + \sum_{k=1}^\infty \frac{(-1)^k}{(1-\xi_r)^k}(\hat{\beta}_{N_r}-\beta_r)(\hat{\xi}_{N_r}-\xi_r)^{k-1}, \quad r \in \{l,u\},$$

we arrive at

$$\begin{aligned}
\hat{a}_M - a \sim\ & (\hat{p}_l - p_l)\Big(u_l - \frac{\hat{\beta}_{N_l}}{1-\hat{\xi}_{N_l}}\Big) - p_l\Big(\frac{\hat{\beta}_{N_l}}{1-\hat{\xi}_{N_l}} - \frac{\beta_l}{1-\xi_l}\Big) + \hat{\mu} - \mu \\
& + (\hat{p}_u - p_u)\Big(u_u + \frac{\hat{\beta}_{N_u}}{1-\hat{\xi}_{N_u}}\Big) + p_u\Big(\frac{\hat{\beta}_{N_u}}{1-\hat{\xi}_{N_u}} - \frac{\beta_u}{1-\xi_u}\Big) - R \\
=\ & (\hat{p}_l - p_l)\Big(u_l - \frac{\beta_l}{1-\xi_l} + \frac{\beta_l}{(1-\xi_l)^2}(\hat{\xi}_{N_l}-\xi_l) - S_l\Big) - p_l\Big(\frac{-\beta_l}{(1-\xi_l)^2}(\hat{\xi}_{N_l}-\xi_l) + S_l\Big) \\
& + \hat{\mu} - \mu + (\hat{p}_u - p_u)\Big(u_u + \frac{\beta_u}{1-\xi_u} - \frac{\beta_u}{(1-\xi_u)^2}(\hat{\xi}_{N_u}-\xi_u) + S_u\Big) \\
& - p_u\Big(\frac{-\beta_u}{(1-\xi_u)^2}(\hat{\xi}_{N_l}-\xi_l) + S_l\Big) - R.
\end{aligned}$$

Note that, by Lemmas A.1 and A.5, $S_u = O_p(u_u)$, $S_l = O_p(u_l)$ and $R = O_+(u_u^{1-1/\xi_u-\delta_u} + |u_l|^{1-1/\xi_l-\delta_l})$, where $u_u = O_+(n^{\alpha\xi_u})$, $u_l = -O_+(n^{\alpha\xi_l})$ and $\alpha \in ((2\xi_u\delta_u)^{-1} \vee (2\xi_l\delta_l)^{-1})$. Hence $R/\gamma \to 0$ and using Lemmas A.1 and A.7, we can write

$$\begin{aligned}
\frac{\hat{a}_M - a}{k_n\gamma} \sim\ & \frac{1}{k_n}\Big[\frac{\sqrt{p_l(1-p_l)}}{\gamma\sqrt{n}}\Big(u_l - \frac{\beta_l}{1-\xi_l}\Big)\sqrt{\frac{n}{p_l(1-p_l)}}(\hat{p}_l - p_l) \\
& + \frac{\sqrt{p_l}\beta_l}{\sqrt{n}\gamma(1-\xi_l)^2}\sqrt{np_l}(\hat{\xi}_{N_l}-\xi_l)\Big]\mathbf{1}_{\{\xi_u \le \xi_l\}} + \frac{\hat{\mu}-\mu}{k_n\gamma} \\
& + \frac{1}{k_n}\Big[\frac{\sqrt{p_u(1-p_u)}}{\gamma\sqrt{n}}\Big(u_u + \frac{\beta_u}{1-\xi_u}\Big)\sqrt{\frac{n}{p_u(1-p_u)}}(\hat{p}_u - p_u) \\
& + \frac{\sqrt{p_u}\beta_u}{\sqrt{n}\gamma(1-\xi_u)^2}\sqrt{np_u}(\hat{\xi}_{N_u}-\xi_u)\Big]\mathbf{1}_{\{\xi_l \le \xi_u\}} + o_p(1) \to_d \mathcal{N}(0,1),
\end{aligned}$$

where

$$k_n^2 = 1 + \Big[\frac{p_l(1-p_l)}{\gamma^2 n}\Big(u_l - \frac{\beta_l}{1-\xi_l}\Big)^2 + \frac{p_l\beta_l^2(1+\xi_l)^2}{n\gamma^2(1-\xi_l)^4}\Big]\mathbf{1}_{\{\xi_u \le \xi_l\}}$$
$$+ \Big[\frac{p_u(1-p_u)}{\gamma^2 n}\Big(u_u - \frac{\beta_u}{1-\xi_u}\Big)^2 + \frac{p_u\beta_u^2(1+\xi_u)^2}{n\gamma^2(1-\xi_u)^4}\Big]\mathbf{1}_{\{\xi_l \le \xi_u\}} = O_+(1).$$

□

## References


Bingham, N., Goldie, C., and Teugels, J. (1987), *Regular Variation*, no. 27 in Encyclopedia of Mathematics and its Applications, Cambridge University Press.

Chambers, R. (1986), "Outlier robust finite population estimation," *Journal of the American Statistical Association*, 81, 1063–1069.

Chambers, R. and Kokic, P. (1993), "Outlier Robust Sample Survey Inference," in *Proceedings of the 49th Session of the International Statistical Institute*, Firenze, no. 55 in Bulletin of the International Statistical Institute, pp. 55–72.

Coles, S. (2001), *An introduction to statistical modeling of extreme values*, Springer.

de Bragança Pereira, C. A. and Rodrigues, J. (1983), "Robust Linear Prediction in Finite Populations," *International Statistical Review*, 51, 293–300.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997), *Modelling Extremal Events*, Springer.

Hall, P. and Jing, B.-Y. (1998), "Comparison of bootstrap and asymptotic approximations to the distribution of a heavy-tailed mean," *Statist. Sinica*, 8, 887–906.

Iachan, R. (1984), "Sampling Strategies, Robustness and Efficiency: The State of the Art," *International Statistical Review*, 52, 209–218.

Johansson, N. C. J. (2002), "Estimating the mean of heavytailed distributions in the presence of dependence," *To appear*.

Karlberg, F. (1999), "Survey Estimation for Highly Skewed Data," Ph.D. thesis, Department of Statistics, Stockholm University, Sweden.

Kokic, P. and Smith, P. (1998a), "Outlier-robust Estimation in Sample Surveys Using Two-Sided Winsorization," Submitted to Journal of the American Statistical Association.

— (1998b), "Windsorization of Outliers in Business Surveys," Submitted to Journal of the Royal Statistical Society Series D.

Leadbetter, M. R. (1991), "On a basis for "peaks over threshold" modeling," *Statist. Probab. Lett.*, 12, 357–362.

Smith, R. (1987), "Estimating Tails of Probability Distributions," *The Annals of Statistics*, 15, 1174–1207.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer.