Linear regression for heavy tails

Guus Balkema & Paul Embrechts

Universiteit van Amsterdam & ETH Zürich

Abstract

There exist several estimators of the regression line in the simple linear regression

$$Y_i = b + aX_i + Y_i^* \qquad i = 1, \ldots, n.$$

Least Squares, Least Absolute Deviation, Right Median, Theil-Sen, Weighted Balance, Least Trimmed Squares. Their performance for heavy tails is compared below on the basis of a quadratic loss function. The case where $X = 1/U$ for a uniform variable U on $(0, 1)$ and where Y^* has a Cauchy distribution plays a central role, but heavier and lighter tails are also considered. Tables list the empirical sd and bias for ten batches of a hundred thousand simulations when X has a Pareto distribution and Y^* a symmetric Student distribution or a one-sided Pareto distribution for various tail indices. The results in the tables may be used as benchmarks. The sample size is $n = 100$ but results for $n = \infty$ will also be presented. The error in the estimate of the slope need not be asymptotically normal. For symmetric errors the symmetric generalized beta prime densities often give a good fit.

0 Introduction

In linear regression the explanatory variables are often assumed to be equidistant on an interval. If the values are random they may be uniformly distributed over an interval or normal or have some other distribution. In the paper below the explanatory variables are random. The X_i are inverse powers of uniform variables U_i in $(0, 1)$: $X_i = 1/U_i^{\xi}$. The variables X_i have a Pareto distribution with tail index $\xi > 0$. The tails become heavier as the index increases. For $\xi \geq 1$ the expectation is infinite. We assume that the error variables Y_i^* have heavy tails too, with tail index $\eta > 0$. The aim of this paper is threefold:

- The paper compares a number of estimators E for the regression line in the case of heavy tails. The distribution of the error is Student or Pareto. The errors are scaled to have interquartile distance IQD $= 1$. The tail index ξ of the Pareto distribution of the explanatory variable varies between zero and three; the tail index η of the error varies between zero and four. The performance of an estimator E is measured by the loss function $L(u) = u^2$ applied to the difference between the slope a of the regression line and its estimate $\hat{a}_E$. Our approach is unorthodox. For various values of the tail indices ξ and η we compute the average loss for ten batches of a hundred thousand simulations of a sample of size hundred. Theorems and proofs are replaced by tables and programs. If the error has a symmetric distribution the square root of the average loss is the empirical sd. From the tables in Section 6 it may be seen that for good estimators this sd depends on the tail index of the explanatory variables rather than the tail index of the error. As a rule of thumb the sd is of the order of

$$1/10^{\xi+1} \qquad 0 \leq \xi \leq 3, \quad 0 \leq \eta \leq 4, \qquad n = 100. \tag{0.1}$$

  This crude approximation is also valid for errors with a Pareto distribution. It may be used to determine whether an estimator of the regression line performs well for heavy tails.

- The paper introduces a new class of non-linear estimators. A *weighted balance estimator* of the regression line is a bisector of the sample. For even sample size half the points lie below the bisector, half above. There are many bisectors. A weight sequence is used to select a bisector which yields a good estimate of the regression line. Weighted balance estimators for linear regression may be likened to the median for univariate samples. The LAD (Least Absolute Deviation) estimator is a weighted balance estimator. However there exist balance estimators which perform better when the explanatory variable has heavy tails.

- The paper gives bounds on the tails of $\hat{a}_E - a$ for some of the estimators E. Occasionally there is a discrepancy between the theoretical bounds on the tail and the empirical sd's listed in the tables in Section 6. The empirical sd may be small with only slight fluctuations in the outcomes for the ten batches of a hundred thousand simulations even when the estimate $\hat{a}_E$ does not have a finite first moment. The implications of this discrepancy for risk analysis will be discussed. The difference $\hat{a}_E - a$ is not asymptotically Gaussian when the explanatory variable has infinite second moment. We shall present empirical evidence which suggests that EGBP (Exponential Generalized Beta Prime) distributions may give a good fit for the distribution of the log of the absolute value of the difference $\hat{a}_E - a$ when the error has a symmetric distribution with heavy tails.

The results of our paper are exemplary rather than analytical. They describe the outcomes of an initial exploration on estimators for linear regression with heavy tails. The numerical results in the tables in Section 6 may be regarded as benchmarks. They may be used as a measure of the performance of alternative estimators. Insight in the performance of estimators of the regression line for samples of size a hundred where the explanatory variable has a Pareto distribution and the error a Student or Pareto distribution may help to select a good estimator in the case of heavy tails.

The literature on the LAD estimator is extensive, see [4]. The theory for the Theil-Sen (TS) estimator is less well developed, even though TS is widely used for data which may have heavy tails, as is apparent from a search on the internet. A comparison of the performance of these two estimators is overdue.

When the tail indices ξ or η are positive outliers occur naturally. Their effect on estimates has been studied in many papers. A major concern is whether an outlier should be accepted as a sample point. In simulations contamination does not play a role. In this paper outliers do not receive special attention. Robust statistics does not apply here. If a good fairy were to delete all outliers that would incommodate us. It is precisely the outliers which allow us to position the underlying distribution in the (ξ, η)-domain and select the appropriate estimator. The formula (0.1) makes no sense in robust regression. Our procedure for comparing estimators by computing the average loss over several batches of a large number of simulations relies on uncontaminated samples. This does not mean that we ignore the literature on robust regression. Robust regression estimates may serve as initial estimates. (This approach does not do justice to the special nature of robust regression, which aims at providing good estimates of the regression line when working with contaminated data.) In our paper we have chosen a small number of geometric estimators of the regression line, whose performance is then compared for a symmetric and an asymmetric error distribution at various points in the ξ, η-domain, see Figure 1a. In robust regression one distinguishes M, R and L-estimators. We shall treat the M-estimators LS and LAD. These minimize the l^p distance of the residuals for $p = 2$ and $p = 1$ respectively. We have not looked at other values of $p \in [1, \infty)$. Tukey's biweight and Huber's Method are non-geometric M-estimators since the estimate depends on the scaling on the vertical axis. The R-estimators of Jaeckel and Jurečková are variations on the LAD estimator. They are less sensitive to the behaviour of the density at the median as we shall see in Section 2. They are related to the weighted balance estimators WB40, and will be discussed in Section 3. Least Trimmed Squares (LTS) was introduced by Rousseeuw in [21]. It is a robust version of least squares. It is a geometric L estimator. Least Median Squares (LMS) introduced in the same paper yields the central line of a closed strip containing fifty of the hundred sample points. It selects the strip with minimal vertical width. If the error has a symmetric unimodal density one may add the extra condition that there are twenty five sample points on either side of the strip. This Least Central Strip (LCS) estimator was investigated in a recent paper [20].

Maximum Likelihood may be used if the error distribution is known. We are interested in estimators which do not depend on the error distribution, even though one has to specify a distribution for the error in order to measure the performance. Nolan and Ojeda-Revah in [19] use Maximum Likelihood to estimate the regression line when the errors have a stable distribution and the explanatory variable (design matrix) is deterministic. Their paper contains many references to applications. They write: "In these applications, outliers are not mistakes, but an essential part of the error distribution. We are interested in both estimating the regression coefficients and in fitting the error distribution." These words also give a good description of the aim of our paper.

There is one recent paper which deserves special mention. It uses the same framework as our paper. In [24] the authors determine limit distributions for the difference $\hat{a}_E - a$ for certain linear estimators for the linear regression $Y_i = aX_i + Y_i^*$. The error Y^* is assumed to have a symmetric distribution with power tails, and the absolute value of the explanatory variable also has a power tail. The tail indices are positive. The estimators are linear expressions in the error terms and functions of the absolute value of the explanatory variables (which in our paper are assumed to be positive):

$$\hat{a}_E - a = \sum |X_i|^{1/(\theta-1)} Y_i^* / \sum |X_i|^{\theta/(1-\theta)}.$$

The estimator $E = E_\theta$ depends on a parameter $\theta > 1$. The value $\theta = 2$ yields LS. The paper distinguishes seven subregions in the positive (ξ, η)-quadrant with different rates of convergence. The paper is theoretical and focuses on the limit behaviour of the distribution of the estimator when the sample size tends to infinity. We look at the same range of values for the tail indices ξ and η, but our approach is empirical. We focus on estimators which perform well in terms of the quadratic loss function $L(u) = u^2$. Such estimators are non-linear. We allow non-symmetric error distributions, and our regression line may have a non-zero abscissa. We only consider two classes of dfs for the error term, Student and Pareto, and our explanatory variables have a Pareto distribution. We restrict attention to the rectangle, $(\xi, \eta) \in [0,3] \times [0,4]$. In our approach the horizontal line $\eta = 1/2$ and the vertical line $\xi = 1/2$ turn out to be critical, but for $\xi, \eta \geq 1/2$ the performance of the estimators depends continuously on the tail indices. There are no sharply defined subregions where some estimator is optimal. Our treatment of the behaviour for $n \to \infty$ is cursory. The two papers present complementary descriptions of linear regression for heavy tails.

Let us give a brief overview of the contents. The exposition in Section 1 gives some background and supplies the technical details for understanding the results announced above. The next four sections describe the estimators which will be investigated in our paper. The first describes the three well-known estimators LS, LAD and RMP. Least Squares performs well for $0 \leq \eta < 1/2$ when the error has finite variance. Least Absolute

Deviation performs well when ξ is small. The estimator RMP (RightMost Point) selects the bisector which passes through the rightmost sample point. Its performance is poor, but its structure is simple. The next section treats the weighted balance estimators. The third Theil's estimator which selects the line such that Kendall's tau vanishes for the residuals, rather than the covariance as for Least Squares. It also introduces a weighted version of the Theil-Sen estimator. The last of these four sections introduces four more estimators: the Least Trimmed Squares estimator, LTS, described above, and three estimators which select a bisector for which a certain state function is minimal when the 25 furthest points above the bisector are trimmed and the furthest 25 below.

The heart of our paper is the set of tables in Section 6 where for $\xi = 0, 1/2, 1, 3/2, 2, 3$ we compare the performance of different estimators. The errors have a Student or Pareto distribution. The tail index of these distributions varies over $0, 1/3, 1/2, 2/3, 1, 3/2, 2, 3, 4$. In order to make the results for different values of the tail index η comparable the errors are standardized so that their df F^* satisfies

$$F^*(-1/2) = 1/4 \qquad F^*(1/2) = 3/4. \tag{0.2}$$

The interquartile distance is IQD $= 1$. There are three sets of six tables, each set containing six tables comparing the performance of various estimators for the slope of the regression line. The six tables correspond to six values of the tail index of the explanatory variable, $\xi = 0, 1/2, 1, 3/2, 2, 3$.

- The first set of tables lists the empirical sd for LS, LAD, LADPC, TS and three versions of LCS for errors with a Student distribution. The estimators do not depend on the tail index η of the error.

- The second set of tables lists the empirical sd for LADGC, RM, HB40 and HB0, WTS, Weighted Theil-Sen, and LTS for errors with a Student distribution. The estimators contain parameters which depend on the value of both tail indices, ξ and η. Thus the Right Median RM depends on an odd positive integer which tells us how many of the rightmost points are used for the median. Theil-Sen, TS, computes the median of the $\binom{n}{2}$ slopes of the lines passing through two sample points; the weighted version introduces a weight which depends on the indices of the two points. The weight increases as the two indices are further apart and as they become smaller. The rate of increase is determined by a parameter. The optimal value of this parameter, the value which yields the smallest empirical sd for a million simulations, depends on the tail indices ξ and η of the explanatory variable and the error. In LTS there are three parameters, the number of sample points which are trimmed, and two positive real valued parameters which determine a penalty for deleting sample points which lie far to the right.

- The third set of tables does the same as the second, but here the errors have a Pareto distribution. Both the empirical sd and bias of the estimates are listed.

The estimators yielding the tables are simple functions of the $2n$ real numbers which determine the sample. Apart from a choice of the parameters they do not depend on the form of the distribution of the error or the explanatory variable. The results show a continuous dependence of the empirical sd on the tail indices ξ and η both for Student and for Pareto errors. The sd and the value of the parameters are different in the third set of tables (Pareto errors) and the second (Student errors) but the similarity of the performance of the estimators for these two error distributions suggests that the empirical sds in these tables will apply to a wide range of error densities. The fourth table lists the optimal values of the parameters for various estimators. Section 6.5 contains an example which shows how the results of the paper should be applied to a sample of size $n = 231$ if the value of the tail indices ξ and η is not known, nor the distribution of the error.

The results in the three tables are for sample size $n = 100$. The explanatory variables are independent observations from a Pareto distribution on $(1, \infty)$, with tail index $\xi > 0$, arranged in decreasing order. One may replace these by the hundred largest points in a Poisson point process on $(0, \infty)$ with a Pareto mean measure with the same tail index. This will be done in Section 8. A scaling argument shows that the slope of the estimate of the regression line for the Poisson point process is larger by a factor approximately 100^ξ compared to the iid sample. For the Poisson point process the rule of thumb (0.1) for the sd of the slope for good estimators has to be replaced by

$$10^{\xi-1} \qquad 0 \leq \xi \leq 3, \quad 0 \leq \eta \leq 4, \qquad n = 100. \tag{0.3}$$

The performance decreases with ξ since the fluctuations in the rightmost point around the central value $x = 1$ increase and the remaining 99 sample points X_i, $i > 1$, with central value $1/i^{\xi}$, tend to lie closer to the vertical axis as ξ increases and hence give less information about the slope of the regression line.

What happens if one uses more points of the point process in the estimate? For $\xi \leq 1/2$ the full sequence of the points of the Pareto point process together with the independent sequence of errors Y_n^* determines the true regression line almost surely. The full sequence always determines the distribution of the error variable, but for errors with a Student distribution and $\xi > 1/2$ it does not determine the slope of the regression line. For weighted balance estimators the step from $n = 100$ to ∞ is a small one. If ξ is large, say $\xi \geq 3/4$, the crude formula in (0.3) remains valid for sample size $n > 100$.

Conclusions are formulated in Section 7. The Appendix contains a brief exposition of the alternative Poisson point process model and an introduction to the EGBP distributions. EGBP distributions often give a good fit to the distribution of the logarithm of the absolute value of $\hat{a}_E - a$ for symmetric errors. The reason for this good fit is not clear.

1 Background

In the paper below both the explanatory variables X_i and the errors Y_i^* in the linear regression

$$Y_i = b + aX_i + Y_i^* \qquad i = 1, \ldots, n \tag{1.1}$$

have heavy tails. The vectors $(X_1, \ldots, X_n)$ and $(Y_1^*, \ldots, Y_n^*)$ are independent; the Y_i^* are iid and the X_i are a sample from a Pareto distribution on $(1, \infty)$ arranged in decreasing order:

$$X_n < \cdots < X_2 < X_1.$$

The Pareto explanatory variables may be generated from the order statistics $U_1 < \cdots < U_n$ of a sample of uniform variables on $(0,1)$ by setting $X_i = 1/U_i^{\xi}$. The parameter $\xi > 0$ is called the *tail index* of the Pareto distribution. Larger tail indices indicate heavier tails. The variables Y_i^* have tail index η. They typically have a symmetric Student t distribution or a Pareto distribution. For the Student distribution the tail index η is the inverse of the degrees of freedom. At the boundary $\eta = 0$ and the Student distribution becomes Gaussian, the Pareto distribution exponential.

The problem addressed in this paper is simple: What are good estimators of the regression line for a given pair (ξ, η) of positive power indices?

For $\eta < 1/2$ the variable Y^* has finite variance and LS (Least Squares) is a good estimator. For $\xi < 1/2$ the Pareto variable $X = 1/U^{\xi}$ has finite variance. In that case the LAD (Least Absolute Deviation) often is a good estimator of the regression line. Asymptotically it has a (bivariate) normal distribution provided the density of Y^* is positive and continuous at the median, see [9]. What happens for $(\xi, \eta) \in [1/2, \infty)^2$? In the tables in Section 6 we compare the performance of several estimators at selected parameter values (ξ, η) for sample size $n = 100$. See Figure 1a. First we shall give an impression of the geometric structure of the samples which are considered in this paper, and describe how such samples may arise in practice.

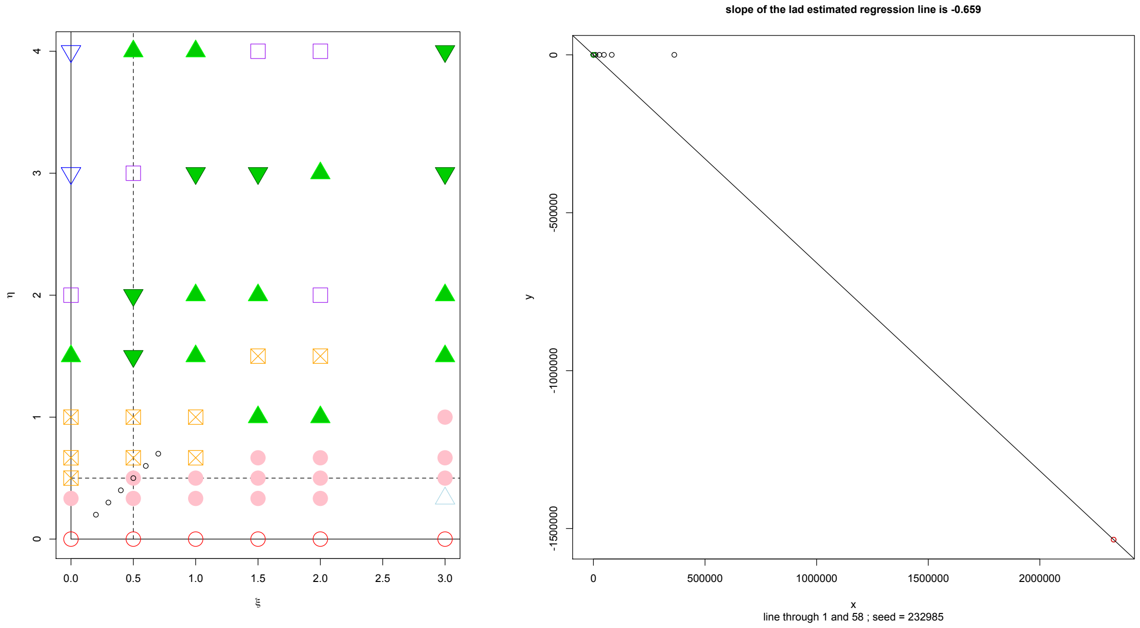
For large ξ the distribution of the points X_i along the positive horizontal axis becomes very skewed. For $\xi = 3$ and a sample of a hundred $\mathbb{P}\{X_1 > 100X_2\} > 1/5$[1]. Exclude the rightmost point. The remaining 99 explanatory variables then all lie in an interval which occupies less than one percent of the range. The point (X_1, Y_1) is a pivot. It yields excellent estimates of the slope if the absolute error is small. The estimator RMP (RightMost Point) may be expected to yield good results. This estimator selects the bisector which passes through (X_1, Y_1).

Definition 1. *A* bisector *of the sample is a line which divides the sample into two equal*

[1] Probabilities for the explanatory variables may be reduced to probabilities for order statistics from the uniform distribution on $(0,1)$. Here we use that given $U_2 = u$ the quotient U_1/U_2 is uniformly distributed on $(0,1)$ and that $5^3 > 100$.

parts. For an even sample of size $2m$ *one may choose the line to pass through two sample points:* $m-1$ *sample points then lie above the line and the same number below.* $\Diamond$

The estimator RMP will perform well most of the time but if Y^* has heavy tails RMP may be far off the mark occasionally, even when $\eta < \xi$. What is particularly frustrating are situations like Figure 1b where RMP so obviously is a poor estimate.



(a) the parameter grid

(b) a bad choice of the regression line

On the left the optimal estimators for Student errors in the points on the grid

$$(\xi, \eta) \in \{0, 1/2, 1, 3/2, 2, 3\} \times \{0, 1/3, 1/2, 2/3, 1, 3/2, 2, 3, 4\}$$

are LS, a red circle, and TS, a pink disk; LAD, a light blue $\triangle$, TB1, a blue $\bigtriangledown$, LADPC, a dark green filled $\bigtriangledown$, and LADHC a light green filled $\triangle$; HB40, an orange $\boxtimes$, and HB0 a purple $\Box$.
In Section 2 the behaviour of LS at the six diagonal points $(i, i)/10$, $i = 2, \ldots, 7$, is investigated.
On the right is a sample for $(\xi, \eta) = (3, 1)$ with Student errors. The true regression line is the horizontal axis. The LAD estimate drawn in the plot is obviously incorrect. This line also is the RMP estimate.

Figure 1a shows the part of (ξ, η)-space to which we restrict attention in the present paper. For practical purposes the square $0 \leq \xi, \eta \leq 2$ is of greatest interest. The results for other values of ξ and η may be regarded as stress tests for the estimators.

Often the variables Y_i^* are interpreted as iid errors. It then is the task of the statistician to recover the linear relation between the vertical and horizontal coordinate from the blurred data for Y. The errors are then usually assumed to have a symmetric distribution, normal or stable.

There is a different point of view. For a bivariate normal vector (X, Y) there exists a line, the regression line $y = b + ax$, such that conditional on $X = x$ the residual $Y^* = Y - (ax + b)$ has a centered normal distribution independent of x. A good estimate of the regression line will help to obtain a good estimate of the distribution of (X, Y). This situation may also occur for heavy-tailed vectors.

Heffernan and Tawn in [11] study the distribution of a vector conditional on the horizontal component being large. In this *conditional extreme value model* they focus on the case where the conditional distribution of Y given $X = x$ is asymptotically independent of x for $x \to \infty$. The vector $\mathbf{Z} = (X, Y)$ conditioned to lie in a half plane $H_t = \{x > t\}$ is a *high risk scenario*, denoted by $\mathbf{Z}^{H_t}$. Properly normalized the high risk scenarios $\mathbf{Z}^{H_t}$ may have a limit distribution for $t \to \infty$. From univariate extreme value theory we know that the horizontal coordinate of the limit vector has a Pareto or exponential distribution. In the *Heffernan-Tawn model* the vertical coordinate of the limit vector is independent of the horizontal coordinate. Heffernan and Tawn in [11] considered vectors with light tails. The results were extended to heavy tails by Heffernan and Resnick in [12]. Other limit distributions are also possible. For instance the quotient of the two coordinates may be independent of the horizontal coordinate. See [2] for a list of all limit distributions.

Given a sample of a few thousand observations from a heavy-tailed bivariate distribution one will select a subsample of say the hundred points for which the horizontal coordinate is maximal. This yields a sequence $x_1 > \cdots > x_{100}$. Choose a vertical coordinate. This yields points $(x_1, y_1), \ldots, (x_{100}, y_{100})$. In the Heffernan-Tawn model the vertical coordinate may be chosen to be asymptotically independent of x for $x \to \infty$. In order to find this preferred vertical coordinate one has to solve the linear regression equation (1.1). The *residuals*, $\hat{y}_i = y_i - (\hat{b} + \hat{a}x_i)$, allow one to estimate the distribution of the error $Y_i^* = Y_i - (b + aX_i)$ and the tail index η.

The estimators of the regression line studied below often contain parameters which depend on the value of the tail indices (ξ, η). The tail index ξ may be estimated by techniques from univariate extreme value theory, see for instance [10]. The tail index η of the vertical coordinate is another matter. Large values of $|Y_i|$ may be due to large values of X_i if the regression line is steep. This leads to a fix. We need the regression line to determine the tail index of Y^*, and we need this tail index η to obtain a good estimate of the regression line. We shall take a two-step approach. Use an estimator E of the regression line which is not strongly dependent on η to obtain an initial estimate of the regression line. Now estimate η from the residuals $Y_i - (\hat{b}_E + \hat{a}_E X_i)$, $i \geq 20$, and use this estimate to improve the initial estimate. Section 6.5 gives the details.

The interpretation of the data should not effect the statistical analysis. Our interest in the Heffernan-Tawn model accounts for the Pareto distribution of the explanatory variable and the assumption of heavy tails for the error term. It also accounts for our focus on estimates of the slope.

We restrict attention to *geometric* estimators of the regression line. Such estimators are called *contravariant* in [17]. A transformation of the coordinates has no effect on the estimate of the regression line L. It is only the coordinates which are affected.

Definition 2. *The group $\mathcal{G}$ of affine transformations of the plane which preserve orientation and map right vertical half planes into right vertical half planes consists of the transformations*

$$(x', y') = (px + q, ax + b + cy) \qquad p > 0, c > 0. \tag{1.2}$$

An estimator of the regression line is geometric *if the estimate is not affected by coordinate transformations in $\mathcal{G}$.* ♢

Simulations are used to compare the performance of different estimators. For geometric estimators one may assume that the true regression line is the horizontal axis, that the Pareto distribution of the explanatory variables X_i is the standard Pareto distribution on $(1, \infty)$ with tail $\mathbb{P}\{X > x\} = 1/x^{1/\xi}$, and that the errors are scaled to have IQD $= 1$. Scaling by the InterQuartile Distance IQD allows us to compare the performance of an estimator for different error distributions.

The aim of this paper is to compare various estimators of the regression line for heavy tails. The heart of the paper is the set of tables in section 6. In order to measure the performance of an estimator E we use the loss function $L(a) = a^2$. We focus on the slope of the estimated regression line $y = \hat{b}_E + \hat{a}_E x$. For given tail indices (ξ, η) we choose $X = 1/U^\xi$ Pareto and errors Y^* with a Student or Pareto distribution with tail index η, scaled to have IQD=1. We then compute the average loss L_r of the slope $\hat{a}_E$ for r simulations of a sample of size $n = 100$ from this distribution. We choose $r = 10^5$. The square root $\gamma = \sqrt{L_r}$ is our measure of performance. It is known as RMSE (Root Mean Square Error). We shall not use this term. It is ambiguous. The mean may indicate the average, but also the expected value. Moreover in this paper the error is the variable Y^* rather than the difference $\hat{a}_E - a$. If the df F^* of Y^* is symmetric the square root $\gamma = \sqrt{L_r}$ is the empirical sd of the sequence of r outcomes of $\hat{a}_E$. The quantity γ is random. If one starts with a different seed one obtains a different value γ. Since r is large one may hope that the fluctuations in γ for different batches of r simulations is small. The fluctuations depend on the distribution of γ, and this distribution is determined by the tail of the random variable $\hat{a}_E$. The average loss L_r is asymptotically normal if L has a finite second moment. For this the estimate $\hat{a}_E$ has to have a finite fourth moment. In Section 2 the distribution of γ is analyzed for $\xi = \eta = i/10$, $i = 2, \ldots, 7$, for the estimator $E =$ LS and Student errors. We will see how the distribution of γ changes on passing the critical value $\eta = 1/2$.

In order to quantify the fluctuations in γ we perform ten batches of 10^5 simulations. This yields ten values γ_i for the empirical sd. Compute the average μ and the sd $\delta = \sqrt{(\gamma_1 - \mu)^2 + \cdots + (\gamma_{10} - \mu)^2}/3$. The two quantities μ and δ describe the performance of the estimator. We shall reduce δ to a digit, 1,2 or 5, and μ to a decimal fraction $m * 10^k$ according to a simple recipe:

Notation: Set $d = 1, 2$ or 5 according as $\delta/10^k$ lies in the interval $[0.7, 1.5)$, $[1.5, 3)$ or $[3, 7)$ where k is chosen to ensure that $\delta/10^k$ lies in $[0.7, 7)$. Now round off $\mu/10^k$ to the nearest integer m, and express the result of the ten batches of 10^5 simulations as $m * 10^k[d]$. For $(\mu, \delta) = (0.01257, 0.000282)$ or $(136.731 * 10^{-7}, 1.437 * 10^{-7})$ or $(221.386, 3.768)$ or

$(221.386, 37.68)$ or $(0.000000347, 0.000000303)$ the recipe gives:

$$0.0126[2] \qquad 137e-7[1] \qquad 221[5] \qquad 220[50] \qquad 3e-7[5]. \tag{1.3}$$

In the fourth example the fluctuations are relatively large. We should not be surprised at the outcome 400[200] for a different set of ten batches of 10^5 simulations. The first three estimates are *good* since $m > 10d$, the fifth is *poor* since $m \leq d$ and the fourth is *weak*. Fluctuations which are comparable to the average are an indication of heavy tails.

Let us mention two striking results of the paper. The first concerns LAD, a widely used estimator which is more robust than LS since it minimizes the sum of the absolute deviations rather than their squares. This makes the estimator less sensitive to outliers of the error. The LAD estimate of the regression line is a bisector of the sample. For $\xi > 1/2$ the outliers of the explanatory variable affect the stability of the LAD estimate, see [22], p.11. The table below lists some results from Section 6 for the empirical sd of the LAD-estimate:

$\xi \setminus \eta$	0	1/2	1	3/2	2	3	4
0	0.0969[2]	0.0951[2]	0.0917[5]	0.0869[2]	0.0810[5]	0.0681[5]	0.0560[5]
1/2	0.0641[2]	0.0690[2]	0.1[1]	3[5]	40[50]	1e7[1]	2e10[5]

Table (1) The empirical sd for $\hat{a}_{LAD}$.

The message from the table is clear. For errors with infinite second moment, $\eta \geq 1/2$, use LAD, but not when $\xi \geq 1/2$. Actually the expected loss for $\hat{a}_{\mathrm{LAD}}$ is infinite for $\eta \geq 1/2$ for all ξ. In this respect LAD is no better than LS. In Section 2 we prove:

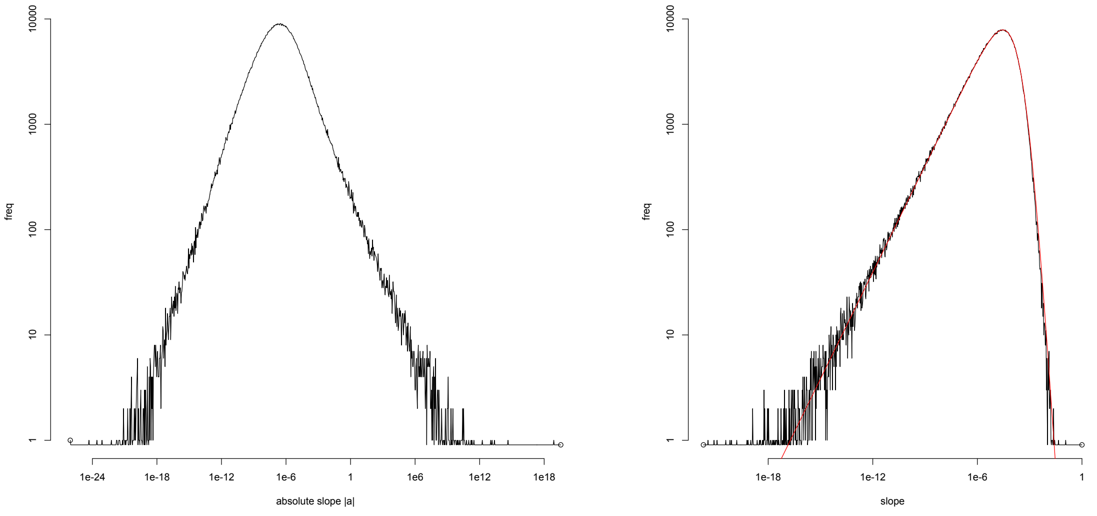
Theorem 2.1. *In the linear regression (1.1) let X have a non-degenerate distribution and let Y^* have an upper tail which varies regularly with non-positive exponent. Let the true regression line be the horizontal axis and let $\hat{a}_n$ denote the slope of the* LAD *estimate of the regression line for a sample of size n. For each $n > 1$ the quotient*

$$Q_n(t) = \mathbb{P}\{Y^* > t\}/\mathbb{P}\{\hat{a}_n > t\}$$

is a bounded function on $\mathbb{R}$.

The discrepancy between the empirical sd, based on simulations, and the theoretical value is disturbing. Should a risk manager feel free to use LAD in a situation where the explanatory variable is positive with a tail which decreases exponentially and the errors have a symmetric unimodal density? Or should she base her decision on the mathematical result in the theorem? The answer is clear: Hundred batches of a quintillion simulations of a sample of size $n = 100$ with X standard exponential and Y^* Cauchy may well give outcomes which are very different from 0.0917[5]. Such outcomes are of no practical interest. The empirical sd's computed in this paper and listed in the tables in Section 6 may be used for risks of the order of one in ten thousand, but for risks of say one in ten million – risks related to catastrophic events – other techniques have to be developed.

A million simulations allow one to make frequency plots which give an accurate impression of the density when a density exists. Such plots give more information then the empirical sd, they suggest a shape for the underlying distribution. We plot the log frequencies in order to have a better view of the tails. Log frequencies of Gaussian distributions yield concave parabolas. Figure 1 below shows loglog frequency plots for two estimators of the regression line for errors with a symmetric Student distribution for $(\xi, \eta) = (3, 4)$.



(a) LAD$(3, 4)$ (b) RM$(21)(3, 4)$

Figure 1: Frequencies of the absolute slope for $(\xi, \eta) = (3, 4)$. On the left the LAD estimator yields an empirical sd = 0e16[1]. On the right the Right Median estimator based on the 21 rightmost sample points, RM(21), yields an empirical sd = 0.00027[5]. The Hill estimate of the right tail index of the distribution of $|\hat{a}_{\text{LAD}}|$ is hR = 3.87; for $|\hat{a}_{\text{RM}(21)}|$ it is hR = 0.436.

First consider the loglog frequency plot of $|\hat{a}_{\mathrm{LAD}}|$ on the left. The range of $|\hat{a}_{\mathrm{LAD}}|$ in Figure 1a is impressive. The smallest value of $|a|$ is of the order of 10^{-24}, the largest of the order of 10^{20}. A difference of more than forty orders of magnitude. Accurate values occur when X_1 is large and $|Y^*|$ small. The LAD estimate of the regression line is a bisector of the sample. In extreme cases it will pass through the rightmost point and agree with the RMP estimate. The value will then be of the order of $1/X_1$. The minimum is determined by the minimal value of 10^8 simulations of a standard uniform variable. For $\xi = 3$ this gives the rough value 10^{-24} for the most accurate estimate. Large values of $|a|$ are due to large values of $|Y_1|$. Because of the tail index $\eta = 4$ the largest value of $|Y_1|$ will be of the order of 10^{24}. Then $|Y_1|/X_1$ is of the order of 10^{18}. For the tail index pair $(\xi, \eta) = (3, 4)$ the limits of computational accuracy for R are breached.

The asymptotes in the loglog frequency plots for $|\hat{a}|$ correspond to power tails for $|\hat{a}|$, at the origin and at infinity. The slope of the asymptote is the exponent of the tail. The plot on the right, Figure 1b, shows that it is possible to increase the slope of the right tail and thus reduce the inaccurate estimates. The value 0e16[1] for the sd of the LAD estimate of the slope is reduced to 0.00027[5] for RM(21). The Right Median estimate RM(21) with parameter 21 is a variation on the Rightmost Point estimate RMP. Colour the 21 rightmost sample points red and then choose the bisector of the sample which also is a bisector of the red points. This is a discrete version of the cake cutting problem: "Cut a cake into two equal halves, such that the icing is also divided fairly." The RM estimate passes through a red and a black point. Below the line are 49 sample points, ten of which are red; above the line too. The tail index of $\hat{a} = \hat{a}_{\mathrm{RM}(21)}$ is at most $2\eta/20 = 0.4$ by Theorem 3.10. The estimate $\hat{a}$ has finite sd even though the value 0.00027[5] shows that the fluctuations in the empirical sd for batches of a hundred thousand simulations are relatively large.

The smooth red curve in the right hand figure is the EGBP fit to the log frequency plot of $\log |\hat{a}|$.

EGBP variables have logconcave densities

$$f = ce^{-\psi} \qquad \psi(t) = r\psi_p(u_0 * (t - t_0)/v_0) \qquad p \in (0,1), r > 0, u_0 > 0, v_0 > 0.$$

The convex function ψ in the exponent is characterized by the location t_0 of the top, the curvature at the top and the absolute slopes of the two asymptotes. EGBP variables have the form $b + aX$ where $X = \log(U/V)$ is the difference of two independent loggamma variables $\log U$ and $\log V$. The appendix contains a short introduction to the class EGBP. The EGBP-fit is particularly good if Y^* has a symmetric distribution. This good fit is the second striking result of this paper. The result is empirical. Section 10 in the Appendix gives details.

The tables in Section 6 are meant to compare the performance of various estimators. It is not the value of the empirical sd's listed in these tables which are important, but rather the induced ordering of the estimators. For any pair (ξ, η) and any estimator E the empirical sd and the value of the parameter will vary when the df F^* of the error Y^* is varied, but the relative order of the estimators is quite stable as one sees by comparing the results for Student and Pareto errors with the same tail index η. Given a geometric estimator, one may apply the techniques of this paper. Determine the average losses in the estimate of the slope for ten batches of 10^5 simulations of a sample of a hundred points. Then compare the performance of this estimator to that of the estimators treated in this paper as listed in Section 6.

The remaining part of this section treats the asymptotic behaviour of estimators when the sample size tends to ∞ via a Poisson point process approach. This part may be skipped on first reading.

Recall that our interest in the problem of linear regression for heavy tails was triggered by a model in extreme value theory. One is interested in the behaviour of a vector $\mathbf{X}$ when a certain linear functional $Z = \zeta(\mathbf{X})$ is large. What happens to the distribution of the high risk scenario $\mathbf{X}^{H_t}$ for the half space $H_t = \{\zeta \geq t\}$ for $t \to \infty$? We consider the bivariate situation and choose coordinates such that ζ is the first coordinate. In the Heffernan-Tawn model one can choose the second coordinate such that the two coordinates of the high risk scenario are asymptotically independent. More precisely there exist normalizations of the high risk scenarios, affine transformations mapping the vertical half plane H_t onto H_1 such that the normalized high risk scenarios converge in distribution. The limit scenario lives on $H_1 = \{x \geq 1\}$. Its first component has a Pareto (or exponential) distribution and is independent of the second component. The normalizations which yield the limit scenario applied to the samples yield a limiting point process on $\{x > 0\}$. (By the Extension Theorem in [2].) If the limit scenario has density $r(x)f^*(y)$ on H_1 with $r(x) = \lambda/x^{\lambda+1}$ on $(1, \infty)$ for $\lambda = 1/\xi > 0$, the limit point process is a Poisson point process N_0 with intensity $r(x)f^*(y)$ where $r(x) = \lambda/x^{\lambda+1}$ on $(0, \infty)$. It is natural to use this point process N_0 to model the hundred points with the maximal ζ-values in a large sample from the vector $\mathbf{X}$.

For high risk scenarios estimators may be evaluated by looking at their performance on the hundred rightmost points of the Pareto point process. For geometric estimators the normalizations linking the sample to the limit point process do not affect the regression line since the normalizations belong to the group $\mathcal{G}$ used in the definition of geometric estimator above. The point process N_0 actually is a more realistic model than an iid sample.

For geometric estimators there is a simple relation between the slope A_n of the regression line for the n rightmost points $(\tilde{X}_i, Y_i^*)$, $i = 1, \ldots, 100$, of N_0 and the slope $\hat{a}_n$ of the regression line for a sample of n independent observations (X_i, Y_i^*):

$$A_n = Z_n^\xi \hat{a}_n \qquad \sqrt{\mathbb{E}A_n^2} = \zeta_n \sqrt{\mathbb{E}\hat{a}_n} \qquad \zeta_n = \sqrt{\mathbb{E}Z_n^{2\xi}}. \tag{1.4}$$

(The first n points of the standard Poisson point process divided by the next point, Z_n, are the order statistics from the uniform distribution on $(0, 1)$ and independent of the Gamma$(n+1)$ variable Z_n with density $x^n e^{-x}/n!$ on $(0, \infty)$.) A simple calculation shows that $\zeta_n = n^\xi + c(\xi) + o(1)$ for $n \to \infty$.

The point process N_a with points $(\tilde{X}_i, Y_i)$, $Y_i = Y_i^* + aX_i$, has intensity $r(x)f^*(y - ax)$. The step from n to $n + 1$ points in estimating the linear regression means that one also takes the $n + 1$st point of the point process N_a into account. This point lies close to the vertical axis if n is large and very close if $\xi > 0$ is large too. The new point will give more accurate information about the abscissa b of the regression line then the previous points since the influence of the slope decreases. For the same reason it will give little information on the value of the slope a. The point process approach allows us to step from sample size $n = 100$ to ∞ and ask the crucial question: Can one distinguish N_a and N_0 for $a \neq 0$?

Almost every realization of N_a determines the probability distribution of the error: If y_0 is a continuity point of F^* the fraction K_n/n of the n rightmost points of N_0 above the horizontal line $y = y_0$ tends to $1 - F^*(y_0)$ almost surely. This holds not only for horizontal lines but for any line $y = y_0 + ax$. Hence the point process N_a determines the error distribution. It need not determine the slope a of the regression line. Just as one cannot distinguish a sequence of standard normal variables W_n from the sequence of shifted variables $W_n + 1/n$ with absolute certainty, one cannot distinguish the Poisson point process N_a from N_0 for $\xi > 1/2$ if the errors have a scaled Student distribution. The distributions of the two point processes are equivalent. The one may be expressed in terms of the other by a strictly positive density. See Appendix 8 for details.

It may seem strange that realizations of the point process N_a determine the probability distribution of Y^* but not the slope a of the regression line. However for the latter one needs points which are not too close to the vertical axis. If ξ is large these are scarce.

The equivalence of the distributions of N_a and N_0 affects the asymptotic behaviour of the estimates A_n of the slope of the regression line for the Poisson point process. There exist no estimators for which A_n converges to the true slope. The limit, if it exists, is a random variable A_∞, and the loss $(A_\infty - a)^2$ is almost surely positive. Because of the simple scaling relation (1.4) between the estimate of the slope for iid samples and for N the limit relation $A_n \Rightarrow A_\infty$ implies $n^\xi \hat{a}_n \Rightarrow A_\infty$. Convergence $\mathrm{SD}_n = \sqrt{\mathbb{E}(A_n - a)^2} \to SD_\infty$ implies that the square root sd_n of the expected loss of $\hat{a}_n - a$ multiplied by n^ξ has this limit too: $n^\xi \, \mathrm{sd}_n \to \mathrm{SD}_\infty$.

Since the Poisson point process N_a determines the distribution of Y^* one may center the error and replace the regression equation $y = y^* + (b + ax)$ by the linear equation $y = y^* + ax$. Let $\hat{a}_n^0$ and A_n^0 denote the estimate of the slope a for the latter equation. These estimates will be more precise than $\hat{a}_n$ and A_n since now there is only the one parameter a. In particular SD_n^0 for large n should give a better idea of the size of SD_∞ then SD_n.

In order to make these observations more concrete we have plotted the values $\mathrm{SD}_n(\xi)$ and $\mathrm{SD}_n^0(\xi)$ for the Least Squares estimator for $\xi \in (0,3]$ and $n = 20, 50, 100, 200, 500, 1000$ in Figure 2. The dotted line is the plot for $\mathrm{SD}_\infty(\xi)$. The labels on the left and the right indicate two different scales. For $\xi > 1$ the plots of SD_n are almost indistinguishable apart from $n = 20$. On the interval $(0,1)$ convergence of SD_n to SD_∞ is slow. One may break up the uncertainty measured by SD_n into two parts, the uncertainty in the slope expressed in SD_n^0, and the uncertainty about the slope of the regression line due to the uncertainty about the line's position. The latter is quite large. From the figure one sees that for the precision of A_{20}^0 one needs a thousand points of the point process in the estimate A_{1000}.

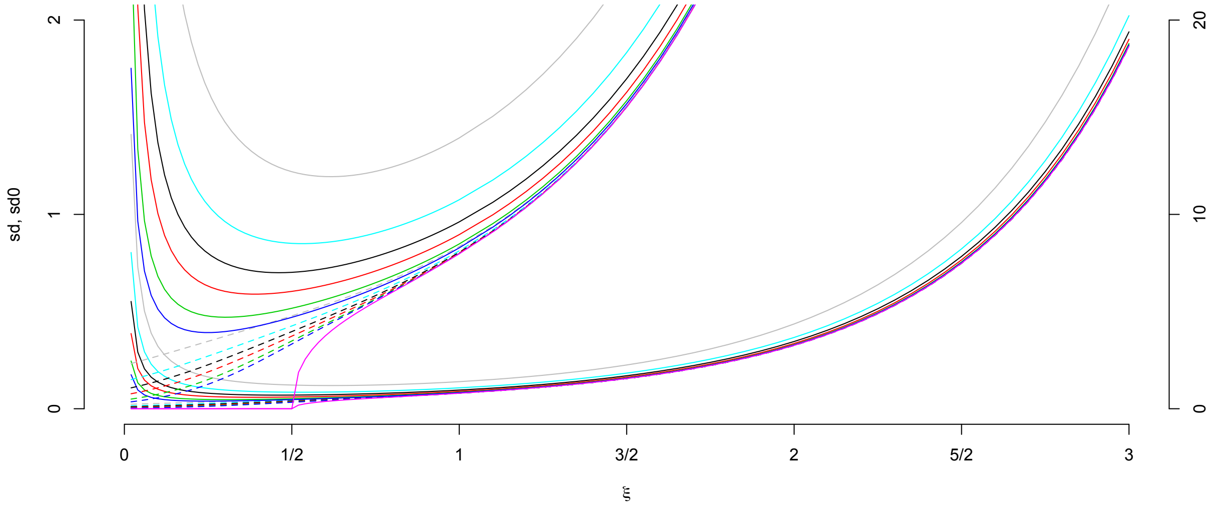


Figure 2: The sd (on two scales) of the slopes A_n and A_n^0 (dashed) of the LS estimates of the regression lines $y = y^* + ax + b$ and $y = y^* + ax$ for the $n = 20, 50, 100, 200, 500, 1000$ (grey, azure, black, red, green, blue) and ∞ (purple) rightmost points of the Poisson point process N_0 for errors Y_i^* with variance one.

The Least Squares estimator does not interest us particularly. It has the attractive property that there exist simple explicit formulas for the estimates and their sd's in terms of the configuration of the explanatory variables and the first two moments of the error. For a centered error and $\xi > 1/2$ one may write $A_\infty^0 - a = A_\infty - a = \sum \omega_n Y_n^*$ with coefficients ω_n depending on the sequence X_n, see [24]. The limit is the sum of the series $\sum \omega_n Y_n^*$. Its distribution depends on the distribution of the error Y^*. The limit distribution is not universal.

Figure 2 gives an impression of what to expect for sample size $n \neq 100$. It raises the question in how far similar results may be obtained for the other estimators considered in this paper.

The asymptotic behaviour of the estimators does not interest us particularly. This paper is about comparing the performance of a number of known and new estimators of the regression line for iid samples of size $n = 100$ when the explanatory variable has a Pareto distribution with tail index $\xi > 0$ and the error a Student distribution with tail index $\eta > 0$ or a Pareto distribution. The results in [24] indicate that for certain linear estimators the asymptotic results depend on the tails of the distribution of the error and of the explanatory variable, but are insensitive to the exact form. Our Ansatz is that this also is the case for the non-linear estimators treated below. The Poisson point process approach creates a framework within which the results of this paper fit and may be interpreted. The last few pages of this section may be regarded as an extended gloss to Figure 2. This figure should be kept in mind when evaluating the entries in the tables in Section 6. We shall return to this topic in Section 8.

2 Three simple estimators: LS, LAD and RMP

Least Squares (LS) and Least Absolute Deviation (LAD) are classic estimators which perform well if the tail index η of the error is small (LS), or the tail index ξ of the explanatory variable (LAD). For $\xi, \eta \geq 1/2$ one may use the bisector of the sample passing through the RightMost Point (RMP) as a simple but crude estimate of the regression line.

At the critical value $\eta = 1/2$ the second moment of the error becomes infinite and the least squares estimator breaks down. Samples change gradually when ξ and η cross the critical value of a half. We shall investigate the break down of the LS estimator by looking at its behaviour for $\xi = \eta = i/10$, $i = 2, \ldots, 7$, for errors with a Student distribution. It will be seen that the notation in (1.3) nicely expresses the decrease in the performance of the estimator on passing the critical exponent. We shall also show that even for bounded errors there may exist estimators which perform better than Least Squares. The estimator

LAD is more robust than Least Squares. Its performance declines for $\xi > 1/2$, but even for $\xi = 0$ (exponential distribution) or $\xi = -1$ (uniform distribution) its good performance is not lasting. The RightMost Point estimate is quite accurate most of the time but may be far off occasionally. That raises the question whether an estimator which is far off 1% of the time is acceptable.

Least squares (LS) is simple to implement and gives good results for $\eta < 1/2$. Given the two data vectors $\mathbf{x}$ and $\mathbf{y}$ we look for the point $a\mathbf{x} + b\mathbf{e}$ in the two-dimensional linear subspace spanned by $\mathbf{x}$ and $\mathbf{e} = (1, \ldots 1)$ which gives the best approximation to $\mathbf{y}$. Set $\mathbf{z} = \mathbf{x} - m\mathbf{e}$ where m is the mean value of the coordinates of $\mathbf{x}$. The vectors $\mathbf{e}$ and $\mathbf{z}$ are orthogonal. Choose a_0 and b_0 such that $\mathbf{y} - a_0\mathbf{z} \perp \mathbf{z}$ and $\mathbf{y} - b_0\mathbf{e} \perp \mathbf{e}$. Explicitly:

$$a_0 = \langle \mathbf{y}, \mathbf{z} \rangle / \langle \mathbf{z}, \mathbf{z} \rangle \qquad b_0 = \langle \mathbf{y}, \mathbf{e} \rangle / \langle \mathbf{e}, \mathbf{e} \rangle.$$

The point we are looking for is

$$a_0\mathbf{z} + b_0\mathbf{e} = a_0\mathbf{x} + (b_0 - ma_0)\mathbf{e} = a\mathbf{x} + b\mathbf{e}. \tag{2.1}$$

This point minimizes the sum of the squared residuals, where the *residuals* are the components of the vector $\mathbf{y} - (a\mathbf{x} + b\mathbf{e})$. Note that m is the mean of the components of $\mathbf{x}$ and $s^2 = \langle \mathbf{z}, \mathbf{z} \rangle$ the sample variance. Conditional on $\mathbf{X} = \mathbf{x}$ the estimate of the slope is

$$\hat{a}_{\mathrm{LS}} = \sum \zeta_i Y_i^* / s \qquad \zeta_i = (x_i - m)/s. \tag{2.2}$$

There is a simple relation between the standard deviation of Y^* and of the estimate $\hat{a}_{\mathrm{LS}}$ of the slope: (ζ_i) is a unit vector; hence conditional on the configuration $\mathbf{x}$ of $\mathbf{X}$

$$\mathrm{sd}(\hat{a}_{\mathrm{LS}}) = \mathrm{sd}(Y^*)/s.$$

If the sd of Y^* is infinite then so is the sd of the estimator $\hat{a}_{\mathrm{LS}}$. That by itself does not disqualify the LS estimator. What matters is that the expected loss is infinite.

Let us see what happens to the average loss L_r when the number r of simulations is large for distributions with tail index $\xi = \eta = \tau$ as τ crosses the critical value 0.5 where the second moment of Y^* becomes infinite. Figure 3 shows the log frequency plots of $\hat{a}_{\mathrm{LS}}$ for $\xi = \eta = \tau(i) = i/10$ for $i = 2, \ldots, 7$ based on ten batches of a hundred thousand simulations. The variable Y^* has a Student t distribution with $1/\eta$ degrees of freedom and is scaled to have interquartile distance one. The most striking feature is the change in shape. The parabolic form associated with the normal density goes over into a vertex at zero for $\tau = 0.5$ suggesting a Laplace density, and a cusp for $\tau > 0.5$. The cusp will turn into a singularity $f(x) \sim c/x^\tau$ with $\tau = 1 - 1/\eta > 0$ when Y^* has a Student distribution with tail index $\eta > 1$.

The change that really matters is not visible in the figure. It occurs in the tails.

The distribution of the average loss L_r depends on the tail behaviour of $L_1 = \hat{a}_{LS}^2$. The Student variable Y^* with $1/\eta$ degrees of freedom has tails $\mathbb{P}\{|Y^*| > t\} \sim c/t^{1/\eta}$. This also holds for $\hat{a}_{LS}$ which is a mixture of linear combinations of these variables by (2.2). For $\eta = i/10$, $i = 2, \ldots, 7$, the positive variable $L_1 = \hat{a}_{\mathrm{LS}}^2$ has upper tail $\mathbb{P}\{L_1 > t\} \sim c_i/t^{5/i}$. See Theorem 2.1 below.

First consider the behaviour of the average $Z_r(i)$ of $r = 10^6$ independent copies of the variable $Z(i) = 1/U^{i/5}$ where U is standard uniform. The Pareto variable $Z(i)$ has tail $1/z^{5/i}$ on $(1, \infty)$. Its expectation is $2/3, 3/2, 4, \infty, \infty, \infty$ for $i = 2, 3, 4, 5, 6, 7$. The average $Z_r(i)$ has the form $m_r(i) + s_r(i)W_r(i)$ where one may choose $m_r(i)$ to be the mean of $Z(i)$ if finite, and where $W_r(i)$ converges in distribution to a centered normal variable for $i = 2$, and for $i > 2$ to a skew stable variable with index $\alpha = i/5$ and $\beta = 1$. The asymptotic expression for Z_r is:

$$Z_r(i) = m_r(i) + s_r(i)W_r(i) \qquad i = 2, 3, 4, 5, 6, 7$$

- $i = 2$: $m_r(2) = 2/3 = \mathbb{E}Z(2)$, $s_r(2) = 1/\sqrt{r}$, $W_r(2) \Rightarrow c_2 W_2$;
- $i = 3$: $m_r(3) = 3/2 = \mathbb{E}Z(3)$, $s_r(3) = r^{-2/5}$, $W_r(3) \Rightarrow c_3 W_{5/3}$;
- $i = 4$: $m_r(4) = 4 = \mathbb{E}Z(4)$, $s_r(4) = r^{-1/5}$, $W_r(4) \Rightarrow c_4 W_{5/4}$;
- $i = 5$: $m_r(5) = \log r$, $s_r(5) \equiv 1$, $W_r(5) \Rightarrow c_5 W_1$;
- $i = 6$: $m_r(6) = s_r(6) = r^{1/5}$, $W_r(6) \Rightarrow c_6 W_{5/6}$;
- $i = 7$: $m_r(7) = s_r(7) = r^{2/5}$, $W_r(7) \Rightarrow c_7 W_{5/7}$.

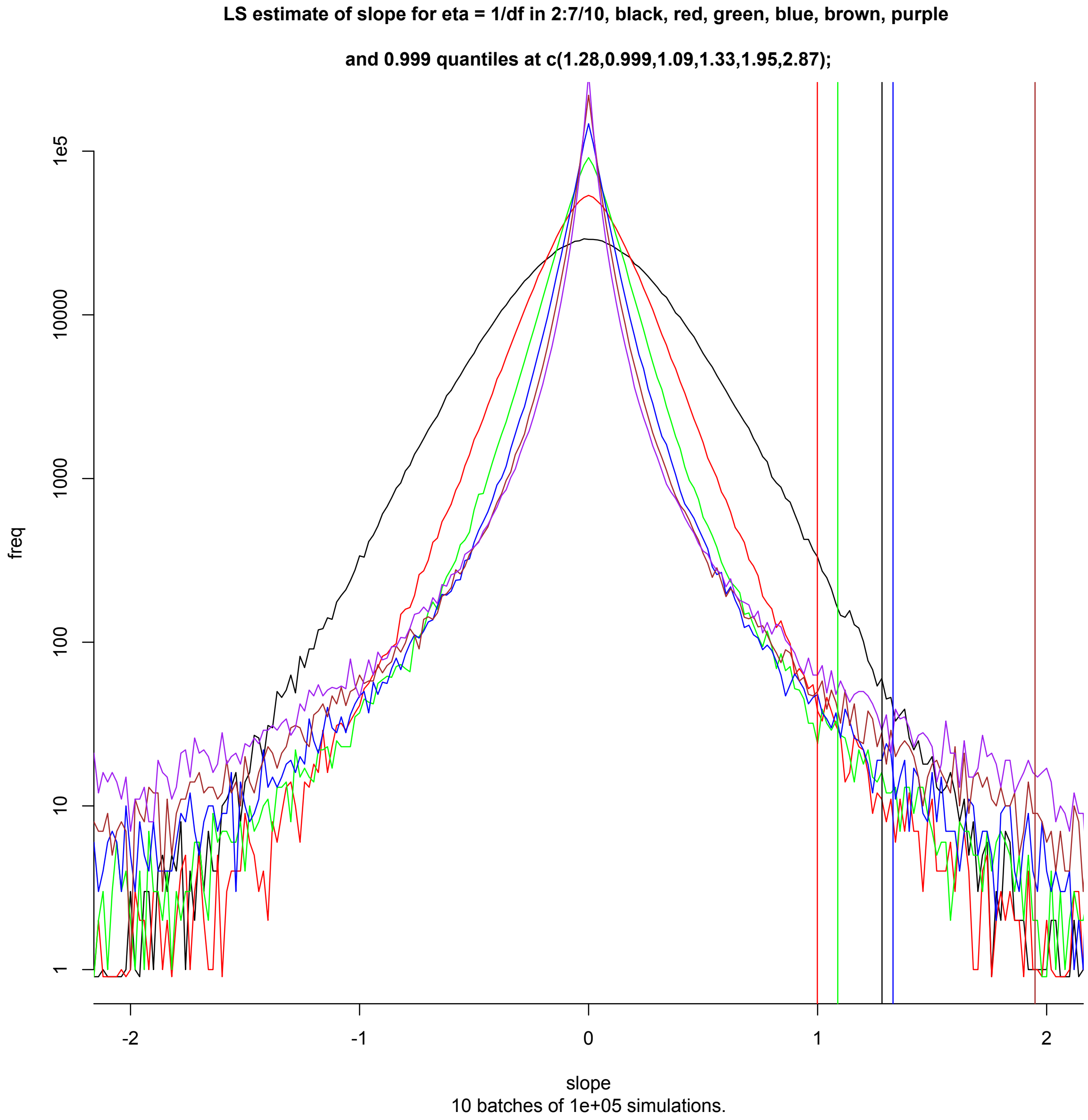


Figure 3: Log frequency of $\hat{a}_{\mathrm{LS}}$ for $\xi = \eta = i/10$ for $i = 2, \ldots, 7$, and 0.999 quantiles of $|\hat{a}_{\mathrm{LS}}|$. The Hill estimates of the tail index of $\hat{a}_{\mathrm{LS}}$ are based on the 500 largest absolute observations:

$\xi = \eta$	1/5	3/10	2/5	1/2	3/5	7/10
0.999 quantile	1.28	0.99	1.09	1.33	1.95	2.87
emp sd	0.312[1]	0.196[1]	0.16[1]	0.19[5]	0.3[1]	1[1]
Hill est	0.15[1]	0.23[1]	0.35[2]	0.46[2]	0.59[5]	0.69[5]

For an appropriate constant $C_i > 0$ the variable $C_iL_1 = C_i\hat{a}_{LS}^2$ has tails asymptotic to $1/t^{5/i}$, and hence the averages C_iL_r exhibit the asymptotic behaviour above. It is the relative size of the deterministic part $m_r(i)$ of L_r compared to the size of the fluctuations $s_r(i)W_r(i)$ of the random part which changes as $i/10$ passes the critical value 0.5. The quotients $s_r(i)/m_r(i)$ do not change much if one replaces L_r by $\sqrt{L_r}$, the batch sd. The theoretical results listed above are nicely reflected in the Hill estimate of the tail index, and the loss of precision in the empirical sd's in the table below Figure 3.

For individual samples it may be difficult to decide whether the parameters are $\xi = \eta = 4/10$ or 6/10. The pairs $(4,4)/10$ and $(6,6)/10$ belong to different domains in the classification in [24] but that classification is based on the behaviour for $n \to \infty$. The relation between the estimates $\hat{a}_{\mathrm{LS}}$ for $(4,4)/10$ and $(6,6)/10$ for samples of size $n = 100$ becomes apparent on looking at large ensembles of samples for parameter values $(i,i)/10$ when i varies from two to seven. The slide show was created in an attempt to understand how the change in the parameters affects the behaviour of the LS estimator. The estimate has a distribution which depends on the parameter. The dependence is clearly expressed in the tails of the distribution. The Hill estimates reflect nicely the tail index η of the error. A recent paper [18] gives similar results for samples with fixed size for LS in linear regression where the coefficient b in (1.1) is random with heavy tails.

The simplicity of the LS estimator makes a detailed analysis of the behaviour of the average loss L_r possible for $\hat{a}_{\mathrm{LS}}$. The critical value is $\eta = 1/2$. The relative size of the fluctuations rather than the absolute size of the sd signal the transition across the critical value. Note that the critical value $\eta = 1/2$ is not due to the "square" in Least Squares but to the exponent 2 in the loss function. There is a simple relation between the tails of the error distribution and of $\hat{a}_{\mathrm{LS}}$. Appendix 9 shows that $\mathbb{P}\{S < s\}/s^{[n/2]}$ is bounded. Lemma 3.4 in [18] then gives a very precise description of the tail behaviour of $\hat{a}_{\mathrm{LS}}$ in terms of the tails of Y^*. We formulate this lemma as a Theorem below.

Theorem 2.1. (Mikosch & de Vries) *Let $\hat{a}_{\mathrm{LS}}$ denote the slope of the* LS *estimate of the regression line in (1.1) for a sample of size $n \geq 4$ when the true regression line is the horizontal axis. Suppose Y^* has a continuous df and X a bounded density. Let*

$T(t) = \mathbb{P}\{|Y^*| > t\}/2$ vary regularly at infinity with exponent $-\lambda < 0$ and assume balance: there exists $\theta \in [-1,1]$ such that

$$\mathbb{P}\{\delta Y^* > t\}/T(t) \to 1 + \delta\theta \qquad t \to \infty, \qquad \delta = \pm 1.$$

Set $M = (X_1 + \cdots + X_n)/n$ and $Z_i = X_i - M$, $V = \sqrt{Z_1^2 + \cdots + Z_n^2}$. Define

$$C_i = \mathbb{E}|Z_i|^\lambda / V^{2\lambda} \qquad B_i = \mathbb{E}\,\mathrm{sign}(Z_i)|Z_i|^\lambda / V^{2\lambda} \qquad i = 1, \ldots, n. \tag{2.3}$$

If $\lambda < [n/2]$ then

$$\mathbb{P}\{\delta \hat{a}_{\mathrm{LS}} > t\}/T(t) \to \sum C_i + \delta\theta \sum B_i \qquad t \to \infty, \qquad \delta = \pm 1.$$

Proof Proposition 9.1 shows that there exists a constant $A > 1$ such that $\mathbb{P}\{V \leq s\} < As^{[n/2]}$ for $s > 0$. Set $\mu = ([n/2] + \lambda)/2$. Then $\mathbb{E}\|\mathbf{U}\|^\mu$ is finite for $\mathbf{U} = \mathbf{Z}/V^2$. Lemma 3.4 in [18] gives the desired result with $C_i = \mathbb{E}|U_i|^\lambda$ and $B_i = \mathrm{sign}(U_i)|U_i|^\lambda$. ¶

If one were to define the loss as the absolute value of the difference $\hat{a}_{\mathrm{LS}} - a$ rather than the square, the expected loss would be finite for $\eta < 1$. In particular the partial averages of $\hat{a}_{\mathrm{LS}}$ for an iid sequence of samples of fixed size n converge almost surely to the true slope. In this respect Least Squares is a good estimator for errors with tail index $\eta < 1$.

Least Absolute Deviations (LAD) also known as Least Absolute Value and Least Absolute Error is regarded as a good estimator of the regression line for errors with heavy tails. The LAD estimator has not achieved the popularity of the LS estimator in linear regression. Yet LAD has always been seen as a serious alternative to the simpler procedure LS. A century ago the astronomer Eddington in his book [6] discusses the problem of measuring the velocity of the planets and writes[2]: “This [LAD] is probably a preferable procedure, since squaring the velocities exaggerates the effect of a few exceptional velocities; just as in calculating the mean error of a series of observations it is preferable to use the simple mean residual irrespective of sign rather than the mean square residual”. In a footnote he adds: “This is contrary to the advice of most textbooks; but it can be

[2] I thank Michael Feast from the Department of Astronomy of the University of Cape Town for drawing my attention to these words.

shown to be true." Forty years earlier Edgeworth had propagated the use of LAD for astronomical observations in a series of papers in the Philosophical Magazine, see [14].

The LAD (Least Absolute Deviations) estimate of the regression line minimizes the sum of the absolute deviations rather than the sum of their squares. It was introduced (by Boscovitch) half a century before Gauss introduced Least Squares in 1806. Computationally it is less tractable, but nowadays there exist fast programs for computing the LAD regression coefficients even if there are a hundred or more explanatory variables. Dielman in [4] gives a detailed oversight of the literature on LAD.

The names "least squares" and "least absolute deviations" suggest that one needs finite variance of the variables Y^* for LS and a finite first moment for LAD. That is not the case. Bassett & Koenker in their paper [3] on the asymptotic normality of the LAD estimate for deterministic explanatory variables observe: "The result implies that for any error distribution for which the median is superior to the mean as an estimator of location, the LAE [LAD] estimator is preferable to the least squares estimator in the general linear model, in the sense of having strictly smaller asymptotic confidence ellipsoids." The median of a variable X is the value t which minimizes the expectation of $|X-t|$, but a finite first moment is not necessary for the existence of the median. The median of an odd number of points on a line is the middle point. It does not change if the positions of the points to the left and the right is altered continuously provided the points do not cross the median. Similarly the LAD-regression line for an even number of points is a bisector which passes through two sample points. The estimate does not change if the vertical coordinate of the points above and below are altered continuously provided the points do not cross the line. Proofs follow in the next section.

Under appropriate conditions the distribution of $\hat{a}_{LAD}$ is asymptotically normal. That is the case if the second moment of X is finite and the density of Y^* is positive and continuous at the median, see [9]. The LAD estimator of the regression line is not very sensitive to the tails of Y^* but it is sensitive to the behaviour of the distribution of Y^* at the median m_0. The sd of the normal approximation is inversely proportional to the density of Y^* at the median. LAD will do better if the density peaks at m_0 and worse if

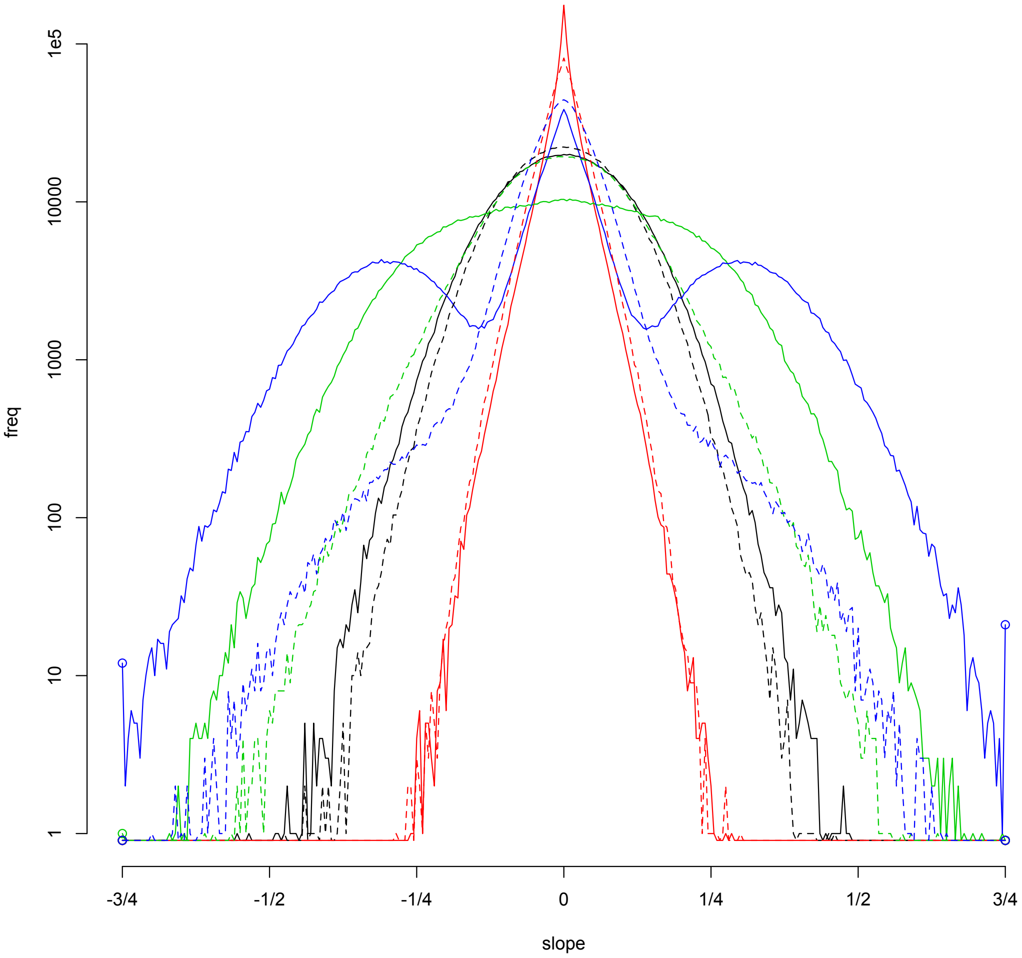


Figure 4: Log frequencies for the estimates $\hat{a}_{\mathrm{LAD}}$ (full line) and $\hat{a}_{LAD40}$ (dashed) of the slope of the regression line based on a million simulations of a sample of 100 points (X, Y^*), with X standard exponential and Y^* of the form $O * U^2$ (red), $O * U$ (black), $O * \sqrt{U}$ (green) and $O * (1 + U)/2$ (blue) where U is standard uniform on $(0, 1)$ and O is a fair sign independent of U. Note that $\hat{a}_{\mathrm{LAD}}$ depends on the density of Y^* at the median; $\hat{a}_{LAD40}$ on the density at the 0.4 and 0.6 quantiles.

$y^* =$	$O * U^2$	$O * U$	$O * \sqrt{U}$	$O * (1 + U)/2$
LS	0.0467[1]	0.0603[2]	0.0739[2]	0.0798[2]
LAD	0.0308[1]	0.0982[2]	0.1774[5]	0.2132[5]
LAD40	0.0376[1]	0.0879[2]	0.1069[2]	0.0677[5].

the density vanishes at m_0.

To illustrate this we consider the case where X has a standard exponential distribution ($\xi = 0$) and Y^* has a density f which is concentrated on $(-1, 1)$ and symmetric. We consider four situations $f(y) = 1/4\sqrt{|y|}$, $f \equiv 1/2$, $f(y) = |y|$ and $f \equiv 1$ on the complement of $[-1/2, 1/2]$. Figure 4 shows the log frequencies of the estimator $\hat{a}_{LAD}$ and $\hat{a}_{LAD40}$. Here LAD40 is a variation on LAD which depends on the behaviour of the distribution of Y^* at the 0.4 and 0.6 quantiles rather than the median. Ten batches of a hundred thousand simulations yield the log frequencies in Figure 4 and the given empirical sd's.

The Gauss-Markov Theorem states that the least squares estimate $\hat{a}_{LS}$ has the smallest sd among all estimates $\hat{a}$ of the slope which are linear combinations of the y_i. It clearly does not apply to LAD or LAD40. The incidental improvement of the performance by ten or thirty per cent is not sufficient to lure the reader away from LS. This paper is not about optimal estimators. A glance at the tables in Section 6 will show that for heavy tails there exist estimators whose performance is abominable. The aim of our paper is to show that there also exist estimators which perform well.

Rightmost point (RMP or RM(1)) (like LAD as we shall see below) is a weighted balance estimator. A balance estimate of the regression line is a bisector which passes through two of the hundred sample points. The regression line for RMP is the bisector which contains the rightmost sample point. The RMP estimate is accurate if X_1 is large, except in those cases where $|Y_1^*|$ is large too. In terms of the quadratic loss function employed in this paper it is a poor estimator for $\eta \geq 1/2$.

The table below lists the empirical sd of the estimate $\hat{a}$ of the slope for LS, LAD and RMP, based on ten batches of a hundred thousand simulations of a sample of size $n = 100$, for various values of the tail indices ξ and η. The explanatory variable X is Pareto with tail $1/x^{1/\xi}$ for $\xi > 0$ and standard exponential for $\xi = 0$; the dependent variable Y^* has a Student distribution with $1/\eta$ degrees of freedom for $\eta > 0$ and is normal for $\eta = 0$. The error is scaled to have Inter Quartile Distance IQD=1.

	$\xi = 0$	1/2	1	3/2	2	5/2	3
$\hat{a}_{\mathrm{LS}}$							
$\eta = 0$	0.0774[2]	0.0518[1]	0.00702[2]	0.00122[1]	0.000245[5]	0.000054[2]	0.0000131[5]
1/3	0.118[1]	0.0790[5]	0.0107[1]	0.00186[2]	0.00037[1]	0.000081[5]	0.000019[1]
1/2	0.23[1]	0.15[1]	0.020[1]	0.0034[2]	0.0007[1]	0.00015[5]	0.00004[1]
2/3	0[100]	30[50]	0[10]	1[1]	0.1[2]	0.01[2]	0.001[2]
1	0[100000]	20000[50000]	3000[5000]	0[1000]	0[100]	2[5]	0.1[1]
$\hat{a}_{\mathrm{LAD}}$							
$\eta = 0$	0.0971[2]	0.0641[2]	0.00859[2]	0.00149[1]	0.000295[5]	0.000066[2]	0.0000157[5]
1/3	0.0959[2]	0.0670[1]	0.00946[2]	0.00169[1]	0.000344[5]	0.0000764[5]	0.0000184[5]
1/2	0.0952[2]	0.0690[5]	0.0103[2]	0.00190[5]	0.00039[2]	0.000087[5]	0.000021[2]
2/3	0.0941[2]	0.072[1]	0.0120[2]	0.003[1]	0.00062[5]	0.00014[5]	0.00004[1]
1	0.0918[5]	0.11[5]	1[1]	0.0[1]	0.008[5]	0.00[1]	0.001[1]
$\hat{a}_{\mathrm{RMP}}$							
$\eta = 0$	0.1966[5]	0.0974[5]	0.0116[1]	0.00189[5]	0.00036[1]	0.000078[5]	0.000018[1]
1/3	0.290[5]	0.144[2]	0.0171[5]	0.0028[1]	0.00052[2]	0.000111[5]	0.000026[2]
1/2	0.6[1]	0.3[1]	0.04[1]	0.006[2]	0.0011[5]	0.0002[1]	0.00005[5]
2/3	4[5]	2[2]	0.3[5]	0.04[5]	0.01[1]	0.002[2]	0.0003[5]
1	300[500]	200[200]	20[50]	3[5]	1[1]	0.1[2]	0.02[5]

The empirical sd of the estimate of the slope, $\hat{a}_{\mathrm{LS}}$, $\hat{a}_{\mathrm{LAD}}$ and $\hat{a}_{\mathrm{RMP}}$, for $\eta \in [0, 1]$
based on ten batches of a hundred thousand simulations.
The break down of LS for $\eta > 1/2$ is dramatic. Even the simple estimator RMP performs better.
The sd's decrease as ξ increases, as is to be expected, and the relative size of the fluctuations increases too.
For $\eta \geq 1/2$ LAD gives the best performance.

Rousseeuw in [21] observes: "Unfortunately, [LAD] is only robust with respect to vertical outliers, but it does not protect against bad leverage points." This agrees with the deterioration of the LAD-estimate for $\eta \geq 1/2$ when ξ increases. The good performance of the LAD estimates for $\xi = 0$ and the relatively small fluctuations reflect the robustness which is supported by the extensive literature on this estimator. It does not agree with the theoretical result below:

Theorem 2.2. *In the linear regression (1.1) let X have a non-degenerate distribution and let Y^* have an upper tail which varies regularly with non-positive exponent. Let the true regression line be the horizontal axis and let $\hat{a}_n$ denote the slope of the* LAD *estimate of the regression line for a sample of size n. For each $n > 1$ the quotient*

$$Q_n(t) = \mathbb{P}\{Y^* > t\} / \mathbb{P}\{\hat{a}_n > t\}$$

is a bounded function on $\mathbb{R}$.

Proof Let $c_1 < c_2$ be points of increase of the df of X. Choose δ_1 and δ_2 positive such that the intervals $I_1 = (c_1 - \delta_1, c_1 + \delta_1)$ and $I_2 = (c_2 - \delta_2, c_2 + \delta_2)$ are disjoint, and $(c_1 - n\delta_1, c_1 + n\delta_1)$ and I_2 too. Let E denote the event that $X_1 \in I_2$ and the remaining $n-1$ values X_i lie in I_1. The LAD regression line L passes through (X_1, Y_1). (If it does not the line L' which passes through (X_1, Y_1) and intersects L in $x = c_1$ has a smaller sum of absolute deviations: Let δ denote the absolute difference in the slope of these two lines. The gain for X_1 is $(c_2 - \delta_2 - c_1)\delta$, and exceeds the possible loss $(n-1)\delta\delta_1$ for the the remaining $n-1$ points.) It is known that the LAD estimate of the regression line is a bisector. We may choose the vertical coordinate so that $y = 0$ is a continuity point of F^* and $0 < F^*(0) < 1$. A translation of Y^* does not affect the result. The event $E_1 \subset E$ that Y_1 is positive and more than half the points (X_i, Y_i), $i > 1$, lie below the horizontal axis has probability $p\mathbb{P}E$ where $p > 0$ depends on $F^*(0)$ and n. If E_1 occurs the regression line L will intersect the vertical line $x = c_1 - \delta_1$ below the horizontal axis. For $Y_1 = y > 0$ the slope A of L then exceeds y/c where $c = (c_2 + \delta_2) - (c_1 - \delta_1)$. Hence $\mathbb{P}\{A > t\} \geq p\mathbb{P}E\mathbb{P}\{Y^* > ct\}$ for $t > 0$. Regular variation of the upper tail of Y^* implies that $\mathbb{P}\{Y^* > ct\} \geq (c^\lambda/2)\mathbb{P}\{Y^* > t\}$ for $t \geq t_0$ where $\lambda \leq 0$ is the exponent of regular variation of $1 - F^*$. This yields the desired result for the quotient Q_n. ¶

How should one interpret this result? The expected loss (MSE) is infinite for $\eta \geq 1/2$. In that respect LAD is no better than LS. We shall introduce the notion of "light heavy" tails. Often heavy tails are obvious. If one mixes ten samples of ten observations each from a Cauchy distribution with ten samples of ten observations from a centered normal distribution, scaling each sample by the maximum of the ten absolute values to obtain point sets in the interval $[-1, 1]$, one will have no difficulty in selecting the ten samples which derive from the heavy-tailed Cauchy distribution, at least most of the time. In practice one expects heavy tails to be visible in samples of a hundred points. However heavy tails by definition describe the df far out. One can alter the density of a standard normal variable Z outside the interval $(-12, 12)$ to have the form c/z^2 for an appropriate constant c. If one takes samples from the variable Z' with the new density the effect of the heavy tails will be visible, but only in very large samples. For a sample of a trillion independent copies of Z' the probability that one of the points lies outside the interval

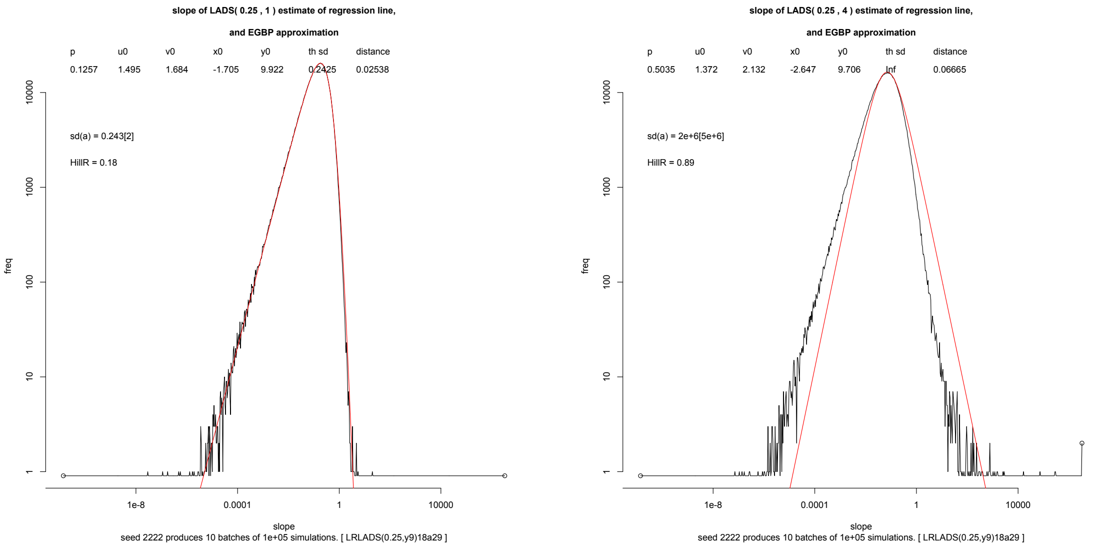
$(-12, 12)$ is less than 0.000 000 000 000 01. Here one may speak of “light heavy” tails. In the proof above it is argued that under certain circumstances LAD will yield the same estimate of the regression line as RMP. The slope of the bisector passing through the rightmost point is comparable to Y_1/X_1 and the upper tail of the df of this quotient is comparable to that of Y_1. In our set up a sufficient condition for LAD to agree with RMP is that $X_1 > 100X_2$. For a tail index $\xi = 3$ the probability of this event exceeds 0.2 as we saw in Section 1. If X has a standard exponential distribution the probability is less. The event $\{X_1 > 100X_2\} = \{U_1 < U_2/e^{100}\}$ for $X_i = -\log(U_i)$ has probability e^{-100}.

Here are two questions raised by the disparity between theory and simulations:

1) Suppose the error has very heavy tails, $\eta \geq 1/2$. Do there exist estimators E of the regression line for which the slope $\hat{a}_E$ has finite second moment? In Section 3 it will be shown that for the balance estimators RM(m) (Right Median) and HB0(d) (Hyperbolic Balance at the median) one may choose the parameters m and d, dependent on the tail indeces ξ and η, such that the estimate of the slope has finite second moment.

2) Is it safe to use LAD for $\xi < 1/2$? Not really. For $\xi < 1/2$ the estimate $\hat{a}_{\text{LAD}}$ is asymptotically normal as the sample size goes to infinity provided the error has a positive continuous density at the median. This does not say anything about the loss for samples of size $n = 100$. The empirical sds for $\xi = 0, 1/2$ and $\eta = 0, 1/2, 1, 3/2, 2, 3, 4$ were listed in Table (1). For $\xi = 0$ the performance of $\hat{a}_{\text{LAD}}$ is good; for $\xi = 1/2$ the performance for $\eta \geq 1$ is bad, for $\eta \geq 3$ atrocious. The empirical sd varies continuously with the tail indices. So what should one expect for $\xi = 1/4$? Ten batches of a hundred thousand simulations yield the second row in the table below: For $\eta \geq 3$ the performance is atrocious. The next sections describe estimators which perform better than LAD, sometimes even for $\xi = 0$. We shall construct an adapted version, LADGC, in which the effect of the large gap between X_1 and X_2 is mitigated by a gap correction.

$\xi \setminus \eta$	0	1/3	1/2	2/3	1	3/2	2	3	4
0	0.0969[2]	0.0959[2]	0.0951[2]	0.0941[2]	0.0917[2]	0.0869[2]	0.0810[5]	0.0681[5]	0.0560[5]
1/4	0.2328[5]	0.2363[5]	0.238[1]	0.2389[5]	0.243[2]	0.26[2]	0.4[2]	10000[10000]	2e+6[5]
1/2	0.0641[2]	0.0670[2]	0.0690[2]	0.072[1]	0.1[1]	3[5]	40[50]	1e+7[2]	2e+10[5]
1	0.00861[5]	0.00943[5]	0.0104[5]	0.012[1]	1[1]	20[20]	10000[10000]	2e+6[2]	0e+15[2].



(a) LAD$(1/4, 1)$

(b) LAD$(1/4, 4)$

Figure 5: Loglog frequencies of $|a_{\mathrm{LAD}}|$ for $\xi = 1/4$, $\eta = 1, 4$. The concave red curve is the EGBP fit. On the left the empirical sd is $0.243[2]$, the theoretical sd of the EGBP fit is 0.2425; on the right $2e+6[5e+6]$ and ∞. On the left the Hill estimate of the tail index based on the 1001 rightmost points is 0.18 yielding a finite fifth moment; on the right 0.89 yields an infinite second moment.

The empirical sd of $\hat{a}_{\mathrm{LAD}}$ for Student errors with tail index η. [LRLADS18a29]

In order to obtain a continuous transition for $\xi \to 0$ one should replace $X = 1/U^{\xi}$ with U uniformly distributed on $(0, 1)$ by $X = (U^{\xi} - 1)/\xi + \xi$ for $\xi \in (0, 1)$. In the table above the entries for $\xi = 1/4$ and $\xi = 1/2$ then have to be divided by 4 and 2 respectively. The rule of thumb (0.1) then is valid for all $\xi \in (0, 1)$. Since the situation $\xi \in (0, 1/2)$ plays no role in this paper we shall stick to the simple formula: $X = 1/U^{\xi}$ for $\xi > 0$.

The words above might evoke the image of a regime switch in the far tails when LAD is contaminated by the pernicious influence of the RMP estimator due to configurations of the sample where the distribution of the horizontal coordinates exhibits large gaps. This image is supported by the loglog frequency plots. For small values of η the plots suggest a smooth concave graph with asymptotic slope one on the left (due to a df of $|\hat{\alpha}_{\mathrm{LAD}}|$ asymptotic to cx for $x \to 0$), and a steeper slope on the right suggesting a tail index < 1 for the upper tail of $|\hat{a}_{\mathrm{LAD}}|$. The two plots for $\xi = 1/4$ and $\eta = 1, 4$ in Figure 5 have different shapes. The slope of the right leg of the right plot becomes less steep as one moves to the right. For two simulations $|\hat{a}|$ lies beyond the boundary value 10^6. The maximal absolute value is $5.3 * 10^9$. This single estimate makes a significant contribution to the average loss for $\eta = 4$. All this suggests that for $\eta = 4$ the tail of the df of $|\hat{a}_{\text{LAD}}|$ becomes heavier as one moves out further to the right.

3 Weighted balance estimators

Recall that a bisector is a line which divides the sample into two equal parts. It may be likened to the median of a one-dimensional sample. For odd sample size bisectors contain a sample point, for even sample size a bisector contains two sample points or none. The latter are called *free* bisectors. There are many bisectors, even if one restricts attention to bisectors through two points in a sample of size $n = 100$. The question is:

"How does one choose a bisector which is close to the regression line?"

For symmetrically distributed errors balance is a good criterium for selecting a bisector. It will be shown below how a decreasing sequence of non-negative weights allows one to define a bisector which is in balance. We give the intuitive background to the idea of a weighted balance estimator, some examples and the basic theory. The focus is on sample size $n = 100$. The extension to samples with an even number of observations is obvious. For a detailed exposition of the general theory and complete proofs the reader is referred to the companion paper [1].

The intuition behind the weighted balance estimators is simple. Assume the true regression line is the horizontal axis. Consider a sample of size $n = 2m$ and a free bisector L. If the slope of the bisector is negative the rightmost sample points will tend to lie above L; if the slope is positive the rightmost points tend to lie below L. Now introduce a decreasing sequence of weights, $w_1 \geq \cdots \geq w_n \geq 0$. The weight of the m points below the bisector will tend to increase as one increases the slope of the bisector. We shall prove that the increase in weight is indeed monotone. As the slope of bisector L increases the weight of the m points below the bisector increases. At a certain moment the weight of the m points below L will surpass half the total weight. That determines the line of balance. This line L is the weighted balance estimate WB0 for the weight sequence w_i.

For odd sample size, $n = 2m + 1$, the same argument works. We then consider bisectors which pass through one sample point to determine the WB0 estimate of the regression line for the given sample.

Strips might give more stable estimates. Consider a sample of size $n = 100$ and strips which contain twenty points such that of the remaining eighty points half lie above the strip and half below. Here too the weight $w(B)$ of the set B of forty points below the strip will increase if the slope of the strip increases and (by symmetry) the weight $w(A)$ of the forty points above the strip will decrease. By monotonicity, as one increases the slope from $-\infty$ to $+\infty$ there is a moment when $w(B)$ will surpass $w(A)$. The centerline of that strip is the WB40 estimate of the regression line for the given weight sequence.

Monotonicity allows one to determine the slope of the estimates WB0 and WB40 by a series of bisections. The Weighted Balance estimator is fast. It is versatile. Both the RightMost Point estimator and Least Absolute Deviation are weighted balance estimators. RMP for the weight sequence $1, 0, \ldots, 0$ and LAD for the random weight sequence $w_i = X_i$ as we shall see below.

We shall now first give some examples. Then we prove the monotonicity mentioned above. We then show that LAD is the weighted balance estimator for the weight sequence $w_i = X_i$. The second half of the section is devoted to an analysis of the tail behaviour of weighted balance estimators. Appendix 8 shows how weighted balance estimators fit in the Poisson point process model.

3.1 Three examples

In this paper we use three basic weight sequences. Two are deterministic.

1) Let $r = 2r_0 + 1$ be a positive odd number less than $n = 100$. Colour the rightmost r sample points red and select a bisector L passing though two sample points, one black and one red. The bisector L is in balance if there are r_0 red points below L and r_0 red points above L. This bisector is the Right Median RM(r) estimate of the regression line. If one rotates the line L anti clockwise over an infinitesimal angle around the point midway

between the red and the black sample point on L the slope increases and one obtains a free bisector. Since the red point lies to the right of the black point the set B of fifty points below this new line now contains r_0+1 red points. The weight sequence for RM(r) is $(1,\ldots,1,0,\ldots,0)$. The Right Median estimator is a variation on RMP which takes account of the position of the r rightmost points and thus avoids the occasional erroneous choice of RMP. If $r=1$ then RM(r) = RMP.

2a) The weight sequence $1, 1/2, \ldots, 1/100$ gives the rightmost point the largest weight. It is overruled by the next three points. $w_2+w_3+w_4>w_1$. If Y_1 is very large the bisector L through the rightmost point will have a large positive slope and the points $\mathbf{z}_2, \ldots, \mathbf{z}_9$ will tend to lie below L. The weight of the 49 points below L, augmented with the left point on L will then exceed half the total weight. For balance the slope has to be decreased. One can temper the influence of the rightmost point by choosing the weights to be the inverse of $2, \ldots, 101$ or $3, \ldots, 102$.

In general we define the *hyperbolic weight sequence* by

$$1/d, 1/(d+1), 1/(d+2), \ldots, 1/(d-1+n). \tag{3.1}$$

The parameter d is positive. If it is large the weights decrease slowly: $w_1/w_2=1+1/d$. Take $n=100$ and let d be a positive integer. Let $\Omega=\sum w_i$ denote the total weight. Consider a bisector L passing through two sample points. The left point $\mathbf{z}_L$ is lighter than the right point $\mathbf{z}_R$. There is a unique bisector such that the weight $w(B)$ of the set B of 49 points below L satisfies

$$w_L+w(B)<\Omega/2<w_R+w(B).$$

This is the line of balance. It is the Hyperbolic Weight HB0(d) estimate of the regression line.

2b) Instead of a line of balance one may look at a strip. Consider a strip S which contains twenty points with forty points above S and forty below. Assume that S is closed and of minimal width. The boundary lines of S each contain one of the twenty points. For certain slopes one of the boundary lines will contain two points and the strip will contain 21 points. Suppose the upper boundary contains two points, $\mathbf{z}_L$ to the left of

$\mathbf{z}_R$, and the set above S contains 39 points. Let $w(B)$ be the weight of the forty points below S and $w(A)$ the weight weight of the 39 points above S. Then S is a *strip of balance* if

$$w(A) + w_L < w(B) < w(A) + w_R.$$

The center line of the strip of balance is the Hyperbolic Weight HB40(d) estimate of the regression line.

Note that the series $\sum 1/(d-1+i)$ diverges. A bad estimate due to very large values of the vertical coordinates of the rightmost k points with total weight approximately $\log(k/d)$ may be overruled by the next k^2 points which have total weight $\approx \log k$. It helps if n is much larger than k^2 and if ξ is not too large.

3) We shall see below that LAD is the weighted balance estimator for the random weight $w_i = X_i$. Since the X_i form an ordered sample from a continuous Pareto distribution one cannot split the hundred sample points into two sets of fifty points with the same weight. It follows that there is a unique bisector L passing through two sample points such that the balance tilts according as one assigns the heavier point on L to the upper or lower set.

3.2 Monotonicity

In this section X and Y^* are assumed to have continuous dfs. Almost all samples of size $n > 1$ from (X, Y^*) then have the following properties:

- No vertical line contains two sample points;
- no two parallel lines contain four sample points.

In particular no line contains more than two sample points. Configurations which satisfy the two conditions above are called *unexceptional.* For unexceptional configurations there is a set of $\binom{n}{2}$ lines which each contain exactly two sample points. The slopes γ of these lines are finite and distinct. They form a set $\Gamma \subset \mathbb{R}$ of size $\binom{n}{2}$.

Definition 3. *A* weight w *is a sequence* $w_1 \geq \cdots \geq w_n \geq 0$ *with* $w_1 > w_n$.

Take an unexceptional sample of n points from (X, Y^*), a positive integer $m < n$ and a line with slope $\gamma \in \mathbb{R} \setminus \Gamma$. As one translates the line upwards the number of sample points below the line increases by steps of one since γ does not lie in Γ, and hence the lines contain at most one sample point. There is an open interval such that for all β in this interval the line $y = \beta + \gamma x$ contains no sample point and exactly m sample points lie below the line. The total weight of these m sample points will be denoted by $w_m(\gamma)$. It depends only on m and γ. As long as the line L moves around without hitting a sample point the set of m sample points below the line does not change and neither does its weight $w_m(\gamma)$. What happens when γ increases?

Lemma 3.1. *For any positive integer $m < n$ and any weight sequence w the function $\gamma \mapsto w_m(\gamma)$ is well defined for $\gamma \in \mathbb{R} \setminus \Gamma$, increasing, and constant on the components of $\mathbb{R} \setminus \Gamma$.*

Proof Consider lines which contain no sample points. Let $\gamma_0 \in \mathbb{R} \setminus \Gamma$ and let L_0 be a line with slope γ_0 which contains no sample points such that there are precisely m sample points below L_0. Let B denote the closed convex hull of these m sample points, and let A denote the closed convex hull of the $n - m$ sample points above L_0. One can move the line around continuously in a neighbourhood of L_0 without hitting a sample point. For any such line between the convex sets A and B the weight of the m points below the line equals $w_m(\gamma_0)$, the weight of B. If one tries to maximize the slope of this line then in the limit one obtains a line L with slope $\gamma \in \Gamma$. This line contains two sample points. The left point is a boundary point of B, the right one a boundary point of A. Consider lines which pass through the point $\mathbf{z} \in L$ midway between these two sample points. The line with slope $\gamma - d\gamma < \gamma$ lies between the sets A and B and $w_m(\gamma - d\gamma) = w_m(\gamma_0)$. For the line with slope $\gamma + d\gamma > \gamma$ there are m points below the line but the left point on L has been exchanged for the heavier right point. Hence $w_m(\gamma + d\gamma) \geq w_m(\gamma - d\gamma)$ with equality holding only if the two points on L have the same weight. ¶

This simple lemma is the crux of the theory of weighted balance estimators. An example will show how it is applied.

Example 1. Consider a sample of $n = 100$ points and take $m = 40$. Let $w_1 \geq \cdots \geq w_{100}$

be a weight sequence. We are looking for a closed strip S in balance: There are forty points below the strip and forty points above the strip, and the weights of these two sets of forty sample points should be in balance. The weight $w_{40}(\gamma)$ of the forty points below the strip depends on the slope γ. It increases as γ increases. The limit values for $\gamma \to \pm\infty$ are

$$\omega_0 = w_{40}(-\infty) = w_{61} + \cdots + w_{100} \qquad \omega_1 = w_{40}(\infty) = w_1 + \cdots + w_{40}.$$

Similarly the weight $w^{40}(\gamma)$ of the forty points above the strip decreases from ω_1 to ω_0. Both functions are constant on the components of $\mathbb{R} \setminus \Gamma$. Weight sequences are not constant. Hence $\omega_0 < \omega_1$. It follows from the monotonicity that here are points $\gamma_0 \leq \gamma_1$ in Γ such that

$$\{w_{40} < w^{40}\} = (-\infty, \gamma_0) \qquad \{w_{40} > w^{40}\} = (\gamma_1, \infty) \qquad \text{on } \mathbb{R} \setminus \Gamma. \tag{3.2}$$

If $\gamma_0 = \gamma_1$ there is a unique closed strip S of minimal height with slope $\gamma_0 = \gamma_1$. One of its boundary lines of S contains one sample point, the other two. This is the strip of balance. A slight change in the slope will cause one of the two points on the boundary line containing two points to fall outside the strip. Depending on which, the balance will tilt to one side or the other. Define the center line of S to be the estimate of the regression line for the estimator WB40 for this particular weight sequence. If $\gamma_0 < \gamma_1$ then for $i = 0, 1$ let L_i with slope γ_i be the center line of the corresponding strip S_i. Define the WB40 estimate as the line with slope $(\gamma_0 + \gamma_1)/2$ passing through the intersection of L_0 and L_1. ◊

The example shows how to define for any sample size n and any non-constant weight w and any positive integer $m < [n/2]$ for any unexceptional sample configuration a unique line, the WBm estimate of the regression line. Of particular interest is the case $m = [n/2]$. The strip then is a line and the line of balance will be denoted by WB0.

Definition 4. *For a sample of size n and $m < [n/2]$ the Hyperbolic Balance estimator* HB$m(d)$ *for $d > 0$ is the weighted balance estimator as in the example above (where $m = 40$) with the weight $w_i = 1/(d - 1 + i)$ in (3.1). If $m = [n/2]$ we write* HB0(d)*.* ◊

In many cases the line or strip of balance is unique. We shall speak of *exact balance* if there is a line or closed strip S with m sample points below S and m above, such that the two sets of m sample points below and above have the same weight. For sample size $n = 2m$ exact balance is not possible for the weight sequence $(1, \dots, 1, 0, \dots, 0)$ if the total weight is odd. Neither is it possible for the random weight sequence $w_i = X_i$ when X has a continuous df.

Proposition 3.2. *For sample size* $n = 100$ *exact balance is not possible for* HB0 *with the hyperbolic weight sequences* $w_i = 1/(d - 1 + i)$ *for parameter* $d \in \{1, \dots, 370261\}$ *neither for* HB40 *for* $d \leq 1000$.

Proof For HB0 the argument is simple. For $d \leq 370261$ the set $J_d = \{d, \dots, d + 99\}$ contains a prime power $q = p^r$, which may be chosen such that $2q > d + 99$. For exact balance there exist two disjoint subsets A and B of J_d containing fifty elements each such that the inverses have the same sum. That implies that $1/q = s$ where s is a signed sum $\sum \epsilon_i/i$ over the remaining 99 elements i with $\epsilon_i \in \{-1, 1\}$. Write s as an irreducible fraction $s = k/m$ and observe that m is not divisble by q since none of the 99 integers i in the sum is. Yet $1/q = k/m$ implies $k = 1$ and $m = q$.

For HB40 one needs more ingenuity to show that exact balance does not occur. One has to show that J_d does not contain two disjoint subsets A and B of forty elements each such that the sums over the inverse elements of A and of B are equal. We show that there exist at least 21 elements in J_d which cannot belong to $A \cup B$. Thus for $d = 1$ the elements $23i$, $i = 1, 2, 3, 4$, cannot belong to $A \cup B$ since for any non-empty subset of these four elements the sum of the signed inverses is an irreducible fraction whose denominator is divisible by 23. This has to be checked! It does not hold in general. The sum of $1/i$ over all $i \leq 100$ which are divisible by 25 is $1/12$. The sum $1/19 + 1/38 + 1/57 - 1/76$ equals $(12 + 6 + 4 - 3)/12/19 = 1/12$. Primes and prime powers in the denominators may disappear. One can ask R to check whether this happens. For $d = 1$ the set J_d contains 32 elements, 23, 29, 31, 37, 41, 43, 46, 47, 49, 53, 58, 59, 61, 62, 64, 67, 69, 71, 73, 74, 79, 81, 82, 83, 86, 87, 89, 92, 93, 94, 97, 98, which cannot lie in $A \cup B$. For $d = 2, \dots, 1000$ the number varies but is never less than 32. ¶

How does one characterize the line or strip of balance? In the case of exact balance there is a line which contains no points or a strip whose boundary lines contain no sample points such that the complement consists of two open half planes which both contain m sample points and which have the same weight. The basic image to keep in mind is that of the graphs of two piecewise constant functions, one increasing and the other decreasing, which cross in a point, a jump point of one of the functions, or which agree on an interval. In the case that the functions cross in one point we have strict balance and the line or strip of balance is unique. This is characterized by the following criteria:

The most intuitive situation is WB0 with even sample size $n = 2m$. The line of balance contains two sample points, $\mathbf{z}_L$ and $\mathbf{z}_R$. The weight w_L of the left point is less than the weight w_R of the right point. The set A of the $m-1$ sample points above the line L has weight $w(A)$ and the set B of the $m-1$ points above L has weight $w(B)$. Now assign one of the two points on L to either side. The weight tilts to the side which receives the heavier point $\mathbf{z}_R$.

If the sample size is odd, $n = 2m+1$, the line of balance L will contain two points. One of the adjacent open half planes will contain m points, the other $m-1$. Depending on which of the two points $\mathbf{z}_L$ or $\mathbf{z}_R$ on L we assign to the half plane with $m-1$ points the weight of the m points in that half plane will be less or more than the weight of the m points in the other half plane.

For the estimator WBm with $m < [n/2]$ the situation is similar to the second case above. The strip of balance S is a closed strip of minimal height. One of the adjacent half planes contains m points, the other $m-1$. The boundary of the latter contains two points. Depending on which of these two we assign to that half plane the weight of the m points in that half plane will be less or more than the weight of the other half plane.

For strict balance the inequalities are strict.

Monotonicity makes weighted balance estimators easy to handle.

Corollary 3.3. *Let n be the sample size and $m \leq n/2$ a positive integer, and let $\gamma_0 \in \mathbb{R}\backslash\Gamma$. Let S_0 be a closed strip of minimal height of slope γ_0 such that m sample points lie below S_0 and m above. If n is even and $m = n/2$ then S_0 is a line which contains no sample* *points, a free bisector; if $n = 2m + 1$ then S_0 is a line which contains one sample point; if $m < [n/2]$ both boundary lines of S_0 contain one sample point. Let w_0 denote the weight of the m sample points below S_0 and w^0 the weight of the m points above S_0. Let $\hat{\gamma}$ denote the slope of the* WB *estimate of the regression line.*

- *If $w_0 < w^0$ then $\gamma_0 < \hat{\gamma}$;*
- *if $w_0 > w^0$ then $\gamma_0 > \hat{\gamma}$;*
- *if $w_0 = w^0$ exact balance holds at γ_0.*

In the case of exact balance there is a maximal interval $J = (\gamma_1, \gamma_2)$ with $\gamma_1, \gamma_2 \in \Gamma$ such that $w_0 = w^0$ for all strips S_0 separating two sets of m sample points with slope $\gamma_0 \in J \setminus \Gamma$. Both γ_0 and $\hat{\gamma}$ lie in J. ♢

In order to find the weighted balance estimate for WB40 one needs a slope γ_1 for which the difference $D(\gamma_1) = w^{40} - w_{40}$ is negative and a slope $\gamma_2 > \gamma_1$ such that $D(\gamma_2)$ is positive. By repeated bisection of the interval (γ_1, γ_2) one quickly obtains good approximations to the slope of the WB40 estimate. Given a good approximation γ_0 one may then apply the discrete geometric approach of the example above.

Proposition 3.4. *The weight sequences (w_i) and $(cw_i + d)$ for $c > 0$ yield the same* WB *estimates.*

Proof The inequalities which define balance remain valid since there are m sample points on either side. ¶

The weight sequence $(6, 1, \dots, 1)$ yields the RMP estimator, but so does any weight which is close to this weight, for instance $(6, w_2, \dots, w_{100})$ with $1 < w_{100} \le w_2 \le 1.1$. A weight sequence for which w_1 is large and the remaining weights cluster together will perform poorly if the error has heavy tails.

Proposition 3.5. *Let* WB0 *be the Weighted Balance estimator of the regression line for the weight sequence (w_i), $i = 1, \dots, n = 2m$. If*

$$w_1 + w_{m+2} + \cdots + w_n > w_2 + \cdots + w_{m+1} \tag{3.3}$$

then WB0 = RMP.

Proof Suppose $\mathbf{z}_1$ does not lie on the line L of balance. Then the weight of $\mathbf{z}_1$ together with the $m-2$ points on the same side of L and the lighter point on L is not greater than half the total weight. ¶

3.3 LAD as a weighted balance estimator

We shall show that LAD = WB0 with weight $w_i = X_i$ for even sample size. The proof for odd sample size is similar. See [1].

Theorem 3.6. LAD = WB0 *with weight* $w_i = X_i$ *for sample size* $n = 2m$.

Proof First consider lines which contain no sample points. Suppose $k < m$ points lie below the line. If we translate the line upwards the $\mathbf{L}^1$ distance $d = \sum |Y_i - (\beta + \gamma X_i)|$, decreases since for the majority of the points the difference $Y_i - (\beta + \gamma X_i)$ is positive. So we may restrict attention to bisectors. Now assume that the weight of the m points below the line is less than half the total weight: $w_m < w^m$. Move the line about without hitting a sample point. If we alter β the distance d does not change since the change in the m positive terms in the sum is compensated by the change in the m negative terms. Now increase the slope γ to $\gamma + \delta$. The distance d decreases by δw^m due to the m positive terms of $Y_i - (\beta + \gamma X_i)$ and increases by δw_m due to the m negative terms. The assumption $w_m < w^m$ implies a decrease in d. So increase the slope as in the lemma above till we reach the line of balance. This line contains two sample points. For this line the distance increases when it is rotated around the point midway between the sample points both clockwise and anti clockwise. ¶

The LAD estimate of the regression line for a sample of size $n = 100$ is (almost surely) a bisector containing two sample points. The estimate is defined in terms of the sum of the absolute vertical distance of the sample points to this line, but it exhibits an almost complete lack of sensitivity to the vertical coordinate. Consider an unexceptional configuration of a sample of a hundred points and let L denote the LAD estimate of the regression line. Now move each of the sample points up or down the vertical line on which it lies, without crossing the bisector L. For the new configuration the line L still is the bisector in balance since the weights have not been altered. Hence L is the LAD estimate of the regression line for the new configuration too. This observation is due to Bassett and Koenker, see [3]. For weighted balance estimators it is trivial.

By treating LAD as a weighted balance estimator it becomes possible to replace the weight $X_i = 1/U_i^\xi$ by a weight $X_i = 1/U_i^\tau$ with $\tau \leq \xi$, for instance $\tau = \xi/12$, and thus avoid the situation where a configuration with x_1 much larger than x_2 forces the line of balance to pass through the rightmost point. The RightMost Point estimator gives very accurate estimates of the slope most of the time if x_1 is large, but in the case of heavy-tailed errors it will occasionally produce estimates which are way out. The power correction above diminishes this effect. Even for $\xi = 1/4$ the gaps x_1/x_2 or x_2/x_3 may still be large and give rise to erroneous estimates, yielding poor performance for errors with tail index $\eta \geq 3$ as we saw in Section 2. A gap correction ensures that the quotients w_i/w_{i+1} of successive weights do not exceed the quotients of the hyperbolic weights $v_i = 1/(d-1+i)$. By applying these corrections one obtains weights which reflect the configuration of the horizontal sample points but the excessive behaviour associated with extreme points of heavy-tailed distributions has been tempered. The tables in Section 6 show that the estimators LADPC (LAD with power correction) and LADHC (LAD with gap correction and a hyperbolic deterministic correction) perform well.

The Weighted Balance estimators are geometric with respect to the group $\mathcal{G}$ for deterministic weight sequences w_i, and so is LAD. The adapted LAD-estimators, LADPC and LADGC, are not. They are geometric with respect to the subgroup $\mathcal{G}_0$ of all transformations

$$(x, y) \mapsto (cx + q, dy + ax + b) \qquad c, d > 0, q = 0. \tag{3.4}$$

The transformations in $\mathcal{G}_0$ map the right half plane $x \geq 0$ onto itself. A scale transformation of the horizontal axis yields a scale transformation of the weights X_i of the LAD-estimator but also of the weights $W_i = X_i^\tau$ of the power corrected LAD estimator. If $X_i' = cX_i$ and $W_i = X_i^\tau$ then the power correction $W_i' = c_i^\tau W_i$ of X_i' produces the same estimate as W_i. Scale the weight sequences so that $W_1 = 1 = W_1'$ and they are identical.

Gap correction acts on the quotients and is not affected by scaling. Gap correction of W_i by the deterministic sequence v_i produces the weight sequence V_i defined by

$$V_i/V_{i+1} = (W_i/W_{i+1}) \wedge (v_i/v_{i+1}) \qquad V_1 = 1. \tag{3.5}$$

Here $v_i = 1/(d-1+i)$ is the hyperbolic weight sequence and $W_i = X_i^\tau$ where $\tau = 1 \wedge 1/\xi$. So $W_i = 1/U_i$ for $\xi \geq 1$.

3.4 Tails of $\hat{a}_{\mathrm{LAD}}$

The performance of LAD is poor for $\xi \geq 1/2$ when the error has heavy tails. We shall now prove that the tails of $\hat{a}_{\mathrm{LAD}}$ for $\xi \geq 0$ are as heavy as those of the error. If the upper tail of the df of Y^* varies regularly with negative exponent the quotient

$$Q_n(t) = \mathbb{P}\{Y^* > t\}/\mathbb{P}\{\hat{a} > t\} \qquad t > 0 \tag{3.6}$$

is bounded.

Proposition 3.7. *Let X have a Pareto distribution with tail index $\xi \geq 0$ and let the error Y^* have a continuous df F^* such that $F^*(0) = 1/2$. Let $1 - F^*(t)$ vary regularly at infinity with negative index. Let the sample size be $n \geq 4$. Then the function $t \mapsto \mathbb{P}\{Y^* > t\}/\mathbb{P}\{\hat{a}_{\mathrm{LAD}} > t\}$ is bounded on $\mathbb{R}$.*

Proof Define E to be the event that X_1 lies in the interval $[n, 2n]$ and the other $n-1$ variables X_i for $i > 1$ satisfy $X_i < 2$. Let E_0 be the intersection of E with the event that at least m of the variables Y_i^*, $i > 1$, are negative for $m = [n/2] + 1$. Both E and E_0 have positive probability. Let $t > 0$. The event $Y_1^* > t$ is independent of E_0 and has probability $1 - F^*(t)$. For LAD the regression line L is the unique line of balance. For configurations in E_0 it passes through the rightmost point by (3.3). Condition on E_0. The event $Y_1^* > t$ implies $\hat{a} > t/2n$ since there are at most $[n/2]$ points below L. Hence one of the m points below the horizontal axis lies on or above L, which implies that the abscissa is negative. Hence $\mathbb{P}(E_0 \cap \{Y_1^* > t\}) \leq \mathbb{P}\{\hat{a} > t/2n\}$. Independence implies $\mathbb{P}\{\hat{a} > s\} \geq \mathbb{P}(E_0)\mathbb{P}\{Y^* > 2ns\}$. Regular variation of the upper tail of the df of Y^* then ensures that the quotient is bounded on $(0, \infty)$. ¶

It might be supposed that estimators like LADPC and LADGC which take the structure of the sequence (X_i) into account will outperform HB0 which is insensitive to the structure of the sample apart from the dependence of the parameter d on the tail indices. Recall however that the median is blind to the values of the sample points and only sees their order, but is an excellent estimate of the center of a symmetric distribution for heavy tails. Figure 1a shows that LADGC is a good estimator for errors with a Student distribution, in particular when η is large.

Definition 5. *A weight sequence is called* responsive *if the components $w_i = w_i(\mathbf{x})$ depend continuously on the vector $\mathbf{x}$ of the explanatory variables and*

$$x_j < x_i \Rightarrow w_j < w_i \qquad x_j = x_i \Rightarrow w_j = w_i \qquad 1 \leq i, j \leq n. \tag{3.7}$$

The weight sequences of LAD, LADPC and LADGC are responsive.

The power correction and gap correction of LAD were constructed to ensure good performance even when ξ is large. The tables in Section 6 show that this goal is achieved. Yet the proposition above also holds for LADPC and LADGC. It holds for all responsive weight sequences.

Theorem 3.8. *Let the regression (1.1) hold with sample size $n \geq 4$. Assume the dfs F of X and F^* of Y^* are continuous. Let $\hat{a}$ be the slope of the* WB0 *estimate of the regression line for the weight w. If the weight is responsive, see (3.7), and if the upper tail of the error, $y \mapsto 1 - F^*(y)$, varies regularly with negative exponent the quotient Q_n in (3.6) is bounded.*

Proof See [1]. $\Diamond$

3.5 Tails of $\hat{a}_{\mathrm{RM}}$

For $\hat{a}_{\mathrm{LAD}}$ the expected loss is infinite for $\eta \geq 1/2$. So too for $\hat{a}_{\mathrm{LADPC}}$ and $\hat{a}_{\mathrm{LADGC}}$. One might be tempted to conclude that Weighted Balance estimators are of little use. Actually the situation is not as dark as it seems. Table 1 in Section 6 shows that LADPC performs well. It is optimal for $\eta > 1/2$ when $\xi = 1/2, 1, 3/2, 2, 3$. If we consider estimators which depend on a parameter then Table 2 shows that for errors with a Student distribution LADHC is optimal or indistinguishable from optimal if η is large and Table 3 shows that for errors with a Pareto distribution LADHC performs quite well, even though it is not up to the Weighted Theil-Sen estimator. The estimator LADHC is a variation of LAD which we shall introduce below. For ten batches of a hundred thousand simulations its performance is indistinguishable from that of LADGC. On the theoretical side there is some light too: There exist weighted balance estimators for which the loss has finite second moment for errors with tail index $\eta \geq 1/2$. The next proposition treats a concrete case to show the basic argument.

Proposition 3.9. *Let X have a bounded density and the error Y^* a continuous df. Assume that there exist positive constants $\lambda < 1$ and C_0 such that $\mathbb{P}\{|Y^*| > t\} \leq C_0/t^\lambda$ for $t > 0$. Assume sample size $n = 100$. Let $\hat{a}$ be the slope of the* RM(r) *regression line for $r = 2r_0 + 1 = 33$. There exists a constant $C > 0$ such that $\mathbb{P}\{|\hat{a}| > t\} \leq C/t^{17\lambda}$ for $t > 1$.*

Proof Colour the 33 rightmost points red. Let L denote the unique line of balance for an unexceptional configuration of the hundred points. There are 16 red points above L and one on L. There are 49 points above L in total, hence at most 33 points with index $i > 50$. Hence there are at least 17 points with index $i > 50$ on or below L. If L is steep either the 17 red points above or on L have large vertical coordinates or the vertical coordinates of the 17 points with index $i > 50$ are large in absolute value. To make this precise let $(x_0, y_0) \in L$ be the point such that x_0 lies midway between x_{33} and x_{51}, and set $d = (x_{33} - x_{51})/2$. Suppose L has slope $a > t > 0$. If y_0 is non-negative there is a set $J \subset \{1, \ldots, 33\}$ of 17 indices such that $Y_j^* > dt$ holds for all $j \in J$. If y_0 is negative there is a set $J \subset \{51, \ldots, 100\}$ such that $Y_j^* < -dt$ holds for all $j \in J$. The number of such subsets J is $M = \binom{33}{17} + \binom{50}{17}$. Hence $\mathbb{P}\{|\hat{a}| > t\} \leq (M\mathbb{P}\{|Y^*| > dt\})^{17}$. Actually $d = D$ is random: $D = (X_{33} - X_{51})/2$. The condition that X has a bounded density ensures that there is a constant $C_1 > 0$ such that $\mathbb{P}\{D \leq s\} \leq C_1 s^{18}$. Since D is independent of the sequence of errors Y_i^* one can bound $\mathbb{P}\{Y_j^* > Dt, j \in J\}$ for any set $J \subset \{1, \ldots, 100\}$ of 17 indices by $C_2/t^{17\lambda}$ on $(0, \infty)$. This is the desired result since by symmetry a similar argument holds for negative slopes. ¶

For the bound on $\mathbb{P}\{D \leq s\}$ and on $\mathbb{P}\{Y_j^* > Dt \mid j \in J\}$ the reader is referred to [1], where she will also find the proof of the general result:

Theorem 3.10. *Let $\hat{a}$ denote the slope of the* RM0(r) *estimate of the regression line for $r = 2r_0 + 1 < k = [(n+1)/2]$ red points, where n denotes the sample size. Let the true regression line be the horizontal axis. Let Y^* have a continuous df and X a bounded density. Suppose there exist positive constants B, β such that $\mathbb{P}\{|Y^*| > y\} \leq (B/y)^\beta$ for $y > 0$. Then there exists a constant $C > 0$ such that*

$$\mathbb{P}\{|\hat{a}| > t\} \leq \begin{cases} C/t^{k-r} & k - r < (r_0+1)\beta \\ C(\log t)/t^{k-r} & k - r = (r_0+1)\beta \\ C/t^{(r_0+1)\beta} & (r_0+1)\beta < k - r. \end{cases} \tag{3.8}$$

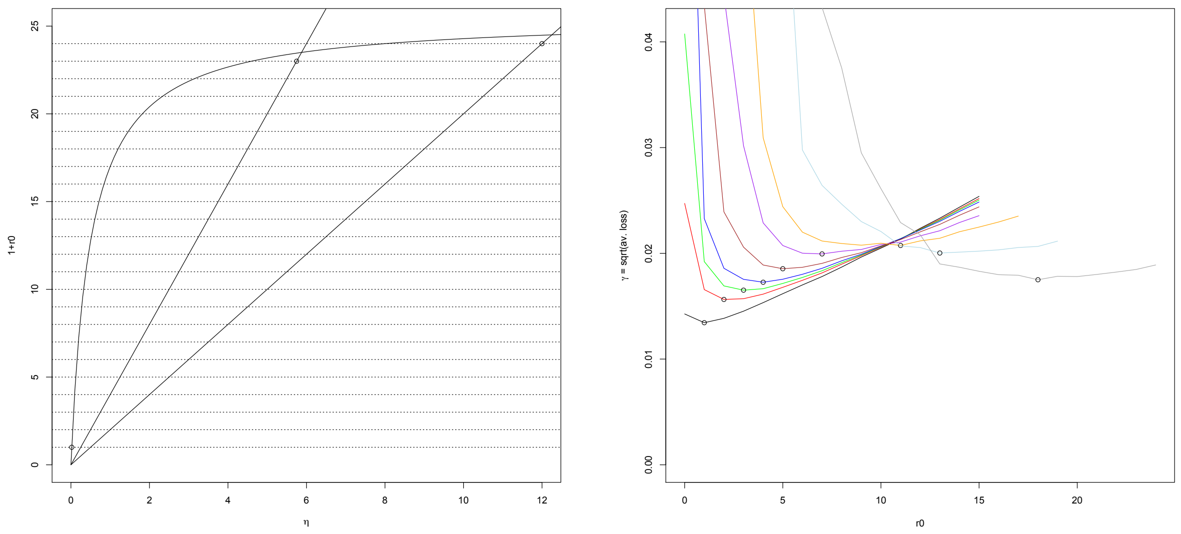
Given the tail index $\eta > 0$ of the error can one choose r_0 such that RM(r) for $r = 2r_0+1$ has finite second moment? For sample size $n = 100$ the condition $(r_0+1)\beta < k - r$ translates into $(2 + 1/\eta)(r_0 + 1) < 51$. Together with the condition $(r_0 + 1)/\eta > 2$ this yields

$$2\eta < r_0 + 1 < 51/(2 + 1/\eta). \tag{3.9}$$

The concave increasing function $s_2(\eta) = 51/(2+1/\eta)$ exceeds $s_1(\eta) = 2\eta$ on $0 < \eta < 49/4$. The condition that r_0 is an integer complicates (3.9). Figure 6a plots s_1 and s_2 on $(0, 49/4)$. Let $\eta_2(s) < \eta_1(s)$ denote the inverse functions on $(0, 49/2)$. It is clear that for $\eta \in (\eta_2(1), \eta_1(24))$ there exists an integer $r_0 \in \{0, \dots, 23\}$ such that (3.9) holds (since $\eta_2(i+1) < \eta_1(i)$ for $i = 1, \dots, 23$).

Proposition 3.11. *Let X have a bounded density and Y^* a continuous df with tail exponent $\eta > 0$. For $\eta \in (1/49, 12)$ and sample size $n = 100$ one may choose an odd integer $r = 2r_0 + 1$ in $\{1, \dots, 47\}$ such that the slope $\hat{a}$ of the* RM(r) *estimate of the regression line has finite second moment. One may choose $r_0 = [2\eta]$.*

Similarly $\mathbb{E}\hat{a}^4$ is finite for $\eta \in (1/49, 23/4)$ for the RM(r) estimator with $r = 2r_0 + 1$ and $r_0 = [4\eta]$.



(a) Finite moments

(b) Optimal parameters

Figure 6: The figure on the left shows that for sample size $n = 100$ and errors with tail index $\eta < 12$ ($\eta < 23/4$) one may choose the parameter $r = 2r_0 + 1$ in the Right Median estimate of the regression line such that (3.9) holds and the slope has finite second moment (finite fourth moment). On the right the square root of the average loss of $\hat{a}_{\mathrm{RM}}$ for a batch of 10^5 simulations for X Pareto with tail index $\xi = 1$ and Y^* Pareto scaled by its IQD, with tail index $\eta = 0, 1/3, 1/2, 2/3, 1, 3/2, 2, 3, 4$ (black, red, green, blue, brown, purple, orange, blue, grey) for various values of r_0. The optimal value of $r = 2r_0 + 1$ is $3, 5, 7, 9, 11, 15, 23, 27, 37$. The graphs all have values ≈ 0.02 in $r0 = 11$.

3.6 Tails of $\hat{a}_{\mathrm{WB}}$

For general weights the argument has to be adapted. Since we measure performance by the loss function $L(u) = u^2$ we are particularly interested in weight sequences for which the slope $\hat{a}$ has finite second moment.

Assume sample size $n = 100$. Define points with index $i < 50$ to be *heavy* and points with indices $i > 51$ to be *light*. There is a positive integer b_H depending only on the weight w such that for any line of balance the set of 49 points above the line augmented with the heavier point on the line contains at least b_H heavy points. The argument is simple. Let L be a bisector which passes through two points. Suppose there are only k heavy points on or above L. Then the total weight of the 49 points above L together with the rightmost point on L is at most $w_1 + \cdots + w_k + w_{50} + \cdots + w_{99-k}$. If k is small the sum may be less than $\Omega/2$. In that case L cannot be a line of balance and $b_H > k$. Similarly balance implies that there are at least b_L light points on or above the line of balance, and by symmetry also at least b_L light points on or below the line. A four line program in R will yield b_H as the smallest integer k for which the sum above equals or exceeds $\Omega/2$. So too for b_L. The minimum $b = b_H \wedge b_L$ is called the *balance minimum* for the weight w. For the hyperbolic weight $w_i = 1/(d - 1 + i)$, $i = 1, \ldots, 100$, we obtain the following results:

d	1	2	5	10	20	50	100	200	500
b_H	4	6	8	10	12	14	14	15	15
b_L	3	5	7	9	11	13	14	14	15

The heavy and light balance minima for HB(d) for sample size $n = 100$ and various values of d.

One can now use the argument which was used for the Right Median in the proposition above. Set $D = (X_{49} - X_{52})/2$. Then $\mathbb{P}\{D \leq s\} \leq C_1 s^3$, and for $t > 0$ the event $\{|\hat{a}| > t\}$ is included in the union of a finite number of events $\{|Y_j^*| > tD \mid j \in J\}$ where J is a subset of $\{1, \ldots, 49\}$ or of $\{52, \ldots, 100\}$ containing b elements. This has been done in [1]. Since the tail index of the error plays an important role in the present paper we use that to formulate a simple corollary to the theorem.

Proposition 3.12. *Let the variables X and Y^* in the linear regression (1.1) have continuous dfs. Suppose the error has tail index $\eta > 0$ and the weight has balance minimum $b \geq 1$. The slope $\hat{a}$ of the corresponding* WB *estimator of the regression line has finite second moment if $\eta/b < 1/2$.*

Proposition 3.13. *Suppose the sample size $n = 2m$ is even. Let w be a weight sequence and define the dual weight w^* by $w_i^* = w_1 - w_{n+1-i}$. Then the light balance minimum of w is the heavy balance minimum of w^*.*

Proof For even sample size the heavy points for w^* are the light points for w and vice versa. ¶

Example 2. For sample size $n = 1000$ and parameter $d = 250$ we find $b_H = b_L \geq 121$ for the hyperbolic weight $1/d, 1/(d+1), \ldots$. Hence the slope $\hat{a}$ of the HB0(d) estimate of the regression line for $d = 250$ will have finite second moment if X has a Pareto distribution with tail index $\xi > 0$ and if the error has a continuous df with tail index $\eta \leq 60$. ◊

Example 3. Assume $n = 100$. There exist weights with balance minimum $b = 0$. No weight has balance minimum $b > 25$. For the weight $1, 0, \ldots, 0$ associated with RMP the set A of fifty points with indices $i = 2, \ldots, 51$ contains no light points, but the weight of A is less than half the total weight. Hence $b_L = 0$. Suppose $b > 25$. Then b_L and b_H both exceed 25. Hence any set A of fifty points which contains 25 heavy points has weight $< \Omega/2$ where $\Omega = \sum w_i$ is the total weight. Now let A be the set of points with indices $i \in \{1, \ldots, 25\} \cup \{51, \ldots, 75\}$. Since w is decreasing and not constant it follows that $w(A) > w(B)$ where B is the complementary set of fifty points. ◊

3.7 Tails of $\hat{a}_{\text{LADHC}}$

On the one hand there are tails which are as heavy as the tails of the error, on the other hand tails which decrease so fast that there is a finite second moment even when the error has tail index $\eta = 4$. How does one harmonize these different tail behaviours for the slope $\hat{a}$? There exists a simple technical solution. Before we formulate that let us remark that one of the attractions of weighted balance estimators is the transparent tail behaviour. One may be unhappy about the heavy tails of LAD and its adapted forms LADPC and LADGC, but the tail behaviour is clear, it has a simple description and the reason for the heavy tails is clear too: For certain configurations the weight will be close to an affine transformation of the weight $(1, 0, \dots, 0)$ of RMP and for these configurations the estimator will exhibit the bad tail behaviour of RMP. Similarly the tail behaviour of RM and HB has a simple explanation: The balance condition implies that there exists a positive integer b, the balance minimum, which depends only on the weight such that for any configuration the line of balance will be steep only if at least b sample points have vertical coordinates which are large in absolute value. For sample size $n = 100$ the balance minimum of ten ensures that the slope $\hat{a}$ has finite second moment if the tail exponent of the error is less than five.

There is a simple formula for combining the responsiveness of LADGC with the good tail behaviour of HB.

Definition 6. *The Deterministic Correction of the random weight V_i by the deterministic weight w_i is the weight W_i which agrees up to a scale factor with the weight V_i in the right half of the points and with the weight w_i in the left half. Set $W_m = 1$ for $m = [n/2]$ and*

$$W_i = V_i/V_m \quad i \leq m; \qquad W_i = w_i/w_m \quad i \geq m. \tag{3.10}$$

The Hyperbolic Correction of LAD*,* LADHC(d) *is the deterministic correction of the random sequence V_i of* LADGC *by the hyperbolic weight $w_i = 1/(d - 1 + i)$.*

In simulations the gap correction of LAD and the hyperbolic correction (with the same parameter d) are indistinguishable. This only aggravates the problem of the disparity in the tail behaviour. For tail indices $(\xi, \eta) = (2, 4)$ the tail of $\hat{a}_{\mathrm{LADGC}}$ is bounded below by $c/t^{1/4}$; the tail of $\hat{a}_{\mathrm{LADHC}}$ is bounded above by $C/t^{5/2}$. How does one choose between an estimator with tails of the order of $c/t^{1/4}$ and one with tails of the order of $C/t^{5/2}$? In order to compare the tail behaviour of the gap correction and the hyperbolic correction one would need to have information on the constants c and C. Sharp bounds on such constants are hard to obtain and will depend not only on the parameters but also on the shape of the underlying dfs. In the absence of such constants we have to accept the lower bound on the tail of $\hat{a}_{\mathrm{LADGC}}$ as a weakness of the estimator. A weakness shared by all weighted balance estimators that are responsive to the configuration of the horizontal coordinates of the sample points. One would like to know how many simulations are needed to reveal the flaw. For LAD a million suffice if the error has very heavy tails, $\eta \geq 3$, even when the tail index of X is small, $\xi = 1/4$, as was shown in Section 2. For LADGC the flaw is not visible in the simulations analyzed in this paper. It is known for which configurations the estimate will be poor: The rightmost x-coordinate is isolated and the remaining x-coordinates all cluster together. Such configurations pose problems for many estimators. Pivot points have received considerable attention in the statistical literature. The hyperbolic correction solves the estimation problems associated with these configurations. It combines sensitivity to the configuration of the horizontal coordinates in the right half of the sample points with the good tail behaviour of the HB0 estimator, while achieving the same performance as the gap corrected version of LAD.

The weight for LADHC(d) is random. Hence so is the balance minimum. However for the deterministic correction of a random weight (V_i) by a deterministic weight (w_i) there is a lower bound b^0 for the balance minimum which only depends on the sequence (w_i). We give the arguments for the heavy balance minimum below for $n = 2m$. Recall that $k < b_H$ holds if a set of m sample points which contains at most k heavy points (with index $i < m$) has weight less than $\Omega/2$ where Ω is the total weight of all sample points. Maximizing the weight of this set of m sample points gives the inequality

$$\sum_{1}^{k} + \sum_{m}^{n-k-1} < \sum_{k+1}^{m-1} + \sum_{n-k}^{n}$$

which may be written as

$$D = \sum_{k+1}^{m-1} - \sum_{1}^{k} > \delta = \sum_{m}^{n-k-1} - \sum_{n-k}^{n}. \tag{3.11}$$

The left hand side is random, the right hand side deterministic.

Proposition 3.14. *Let W be the deterministic correction of a random weight V by the deterministic weight w for sample size $n = 2m$. Let $\Omega = \sum W_i$. There exists an integer b_H^0 which depends only on w such that any set A of m sample points which contains less* *than b_H^0 points with index $i < m$ has weight $W(A) < \Omega/2$. The value b_H^0 is optimal: There exists a weight V_0 and a set A_0 of m sample points, of which b_H^0 have index $i < m$, such that $W_0(A_0) \geq \Omega/2$ holds with positive probability.*

Proof Assume that w has been scaled to satisfy $w_m = 1$. Introduce weights $z = z(t)$ for $1 \leq t \leq w_1$ by

$$z_i(t) = \begin{cases} tw_i/w_1 \vee 1 & i = 1, \dots, m \\ w_i & i = m, \dots, n. \end{cases}$$

Observe that $z(1) = w \wedge 1$ and $z(w_1) = w$. In general for $t \in (w_1/w_j, w_1/w_{j+1})$ the weight $z(t)$ satisfies $z_i(t) = 1$ for $i > j$ and $z_i(t) > 1$ for $i = 1, \dots, j$. Let $d = d(t)$ denote the difference D on the right hand side of (3.11) for the weight $z(t)$. One can show that $t \mapsto d(t)$ is continuous on $[1, w_1]$ with a derivative $\dot{d}(t)$ which is constant on intervals $(w_1/w_j, w_1/w_{j+1})$, negative on $(1, w_1/w_{2k})$, and increasing on $(w_1/w_k, w_1)$. Hence d is a piecewise linear convex function on $[w_1/w_k, w_1]$. It is minimal in w_1/w_j for an index $j \geq 2k$ where $j = m$ or j is the first index i for which $w_1 + \cdots + w_k < w_{k+1} + \cdots + w_i$. Let d_0 denote this minimum. Now observe that conditional on $W_k = c \geq 1$ the difference D in (3.11) is bounded below by $d(t)$ if we choose t such that $z_k(t) = c$. Hence $D \geq d(T) \geq d_0$ for $T = w_1 W_k / w_k$. Define b_H^0 to be the minimal integer k for which $d_0 = d_0(k) \leq \delta$. If $j_0 \geq 2k_0$ is the index j associated with $k_0 = d_H^0$ then the weight $V_0 = z(t_0)$ with $t_0 = w_1/w_{j_0}$ satisfies $V_0(A) \geq \Omega/2$ where A is the set of sample points with index $i \in \{1, \dots, k_0\} \cup \{m, \dots, n-1-k_0\}$. ¶

If L is a bisector for the weight W and balance holds then the set of $m-1$ points above L together with the heavier point on L contain at least b_H^0 points with index $i > m$. One may define the light balance minimum b_L^0 similarly. These two integers depend only on the deterministic weight w. For the hyperbolic weights $w_i = 1/(d-1+i)$ and sample size $n = 100$ the optimal lower bounds b_H^0 and b_L^0 for the heavy and light balance minima are listed below for various values of the parameter d.

d	1	2	3	4	5	6	8	10	12	15	20	25	30	40	50	60	80	100
b_H	4	6	7	7	8	9	9	10	11	11	12	12	13	13	14	14	14	14
b_H^0	3	5	5	6	7	7	8	8	9	9	10	10	11	11	11	12	12	12
b_L	3	5	6	7	7	8	9	9	10	11	11	12	12	13	13	13	14	14
b_L^0	2	4	5	6	6	7	7	8	8	9	10	10	11	11	11	12	12	12

The heavy and light balance minima b_H and b_L for HB(d) for sample size $n = 100$ and various values of d and the optimal lower bounds b_H^0 and b_L^0 for the corresponding Hyperbolic Correction of a random weight.

4 Theil's estimator and Kendall's τ

Least squares chooses a regression line such that the residuals $y_i - (ax_i + b)$ have zero mean and are uncorrelated with the n values x_i of the explanatory variable. Theil in [28] showed that the median of the $\binom{n}{2}$ lines passing through two sample points is an estimator of the regression line for which Kendall's τ vanishes. The Theil-Sen estimator is widely used in papers on climate change.

Kendall's τ is a robust measure of the correlation or strength of association in a bivariate sample. It is non-parametric. It counts the number of inversions. An inversion holds for two indices $i \neq j$ if $(y_j - y_i)/(x_j - x_i)$ is negative. The pair (i, j) is then called discordant. In a sample of n points there are $\binom{n}{2}$ pairs. If the number of inversions is $\binom{n}{2}$ and the x_i are in decreasing order then the y_i are in increasing order. If there are no inversions the two sequences have the same order. By definition:

$$\tau = 2\frac{n_c - n_d}{n(n-1)} \tag{4.1}$$

where n_c is the number of concordant pairs and n_d the number of discordant pairs. For independent variables τ is centered and asymptotically normal (for $n \to \infty$) with variance $2(2n + 5)/(9n(n - 1))$. See [17].

Let x and y be sequences of the same length with distinct values and define $\tau(y, x)$ as in (4.1).

Proposition 4.1. *The function $a \mapsto \tau(a) = \tau(y + ax, x)$ is increasing.*

Proof It suffices to prove $\tau(a) > \tau(0)$ for $a > 0$ (replace y by $y + a_1 x$). Assume $j < i$. Set $c = y_j - y_i$ and $d = x_j - x_i > 0$. Now observe that $(c + ad)/d$ lies between c/d and a and may be positive if c/d is negative but can not be negative if c is positive. ¶

It follows that one may determine $a = \hat{a}$ such that $\tau(a) = 0$ by finding a value a_1 where τ is negative and a value $a_2 > a_1$ where τ is positive and then using successive bisections.

Theil in 1950 proposed to use the median of the $\binom{n}{2}$ quotients $(y_j - y_i)/(x_j - x_i)$ of the plane sample (x, y) as an estimator of the slope of the regression line through the sample. He proved:

Theorem 4.2. *Let $\hat{a}_T$ denote the Theil estimator of the slope. Then $\tau(\hat{a}_T) = 0$.*

The estimator has a simple structure. It is called complete in [28] since it makes use of the complete set of quotients. Sen in [25] extended the estimator to the case where points may have the same x-coordinate. Siegel in [26] introduced a variation where one first computes for each point the median of the quotients involving that point, and then takes the median of these medians. This is related to the RM$(2k + 1)$ estimator. The estimate $\hat{a}_{\mathrm{RM}(m)}$ is the median of the slope of the lines through the $m = 2k + 1$ rightmost points dividing the sample into two equal parts.

Jaeckel in [13] proposed a weighted version with weights $w_{ij} = x_j - x_i$, $j < i$. For deterministic explanatory variables $x_{i,n}$ under appropriate conditions this weighted Theil-Sen estimate is asymptotically normal, see [27]. These conditions do not apply in our situation. For large values of ξ the weights with $j = 1$ will tend to dominate the sum. For the Weighted Balance estimator replacing the weights X_i by the hyperbolic weights yielded good results. We shall therefore use the weights

$$w_{ij}(d) = \frac{1}{i + d} - \frac{1}{j + d} \qquad i < j \tag{4.2}$$

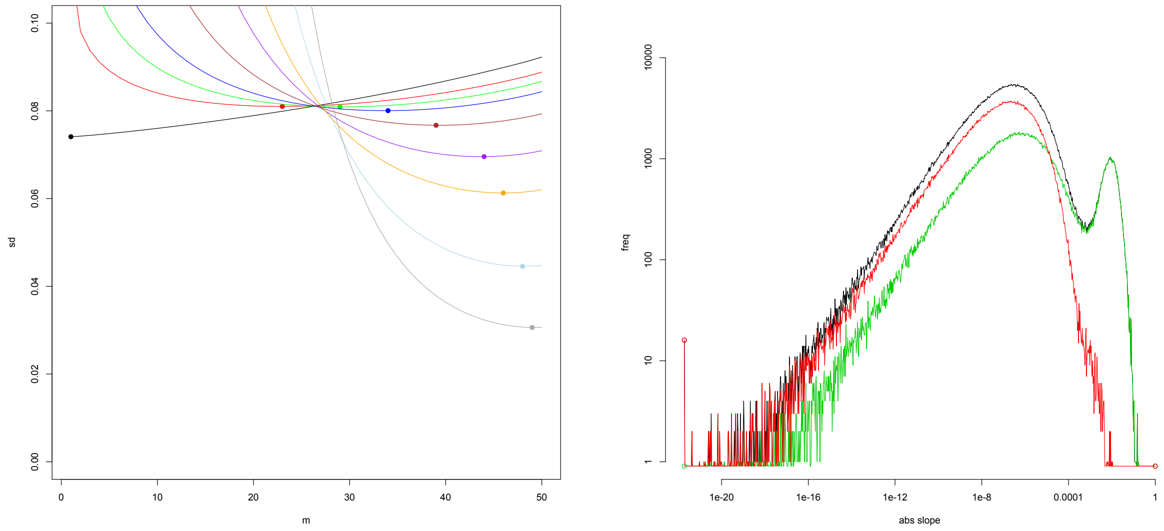
which promote pairs for which the smaller index is close to one and for which i and j are far apart. The parameter $d > 0$ determines how strong this bias for the rightmost points is. The unweighted Theil-Sen estimator outperforms the weighted Theil-Sen estimator with $d = 1$, but by increasing d and thus decreasing the bias of the weights the empirical sd of the estimator may be decreased by a factor 2 or more.

5 Trimming

Trimming is an excellent procedure for getting rid of the noisy outer observations in a sample from a heavy-tailed distribution. If one arranges the observations from a Student distribution with tail index η in decreasing order, $Y_1 > Y_2 > \cdots > Y_n$, the maximal term Y_1 has tail index η, the second largest term Y_2 has tail index $\eta/2$, the third largest tail index $\eta/3$, etc. For the very heavy tails with index $\eta = 4$ deleting the eight largest and the eight most negative observations leaves us with variables $Y_9, \ldots, Y_{n-8}$ which have finite variance. Trimming reduces the sample size (and destroys independence), but this is compensated by the good behaviour of the remaining sample points. The number of sample points that have to be trimmed to get good performance depends on the tail index of the distribution.

Take samples of a hundred Student variables with tail index $\eta \in \{0, 1/3, 1/2, 2/3, 1, 3/2, 2, 3, 4\}$ scaled by their IQD. Perform ten batches of a hundred thousand simulations, compute the average $\hat{a}$ of the trimmed sample, the square root α of the average loss for each of the ten batches, and the average, $\bar{\alpha}$, and sd of these ten values of α. Do this for various values of m, where m is the number of observations which are deleted to the right and to the left. Thus $\hat{a}$ is the estimate of the center of the distribution based on the $100 - 2m$ centermost observations. Figure 7a plots the values of $\bar{\alpha}$ for $m = 0, \ldots, 49$ for the nine values of the tail index η. Note that - as a result of the scaling by the IQD - the nine plots fit in the same frame, the minimal values are comparable and for some unknown reason the plots all pass through approximately the same point, $(27, 0.08)$. Note too that even for $\eta = 1/3$ when Y has finite second moment almost half the sample points are deleted to obtain the estimate $\hat{a}_m(\eta)$ with the minimal loss. For $\eta = 4$ the optimal trimmed average is the median.

Rousseeuw in [21] suggested a trimming procedure for linear regression with heavy tails. Fix a positive integer $m < n/2 - 1$. For any slope $\gamma \in \mathbb{R}$ consider a closed strip S with slope γ such that there are m points above the strip and m points below. Compute the LS estimate $\hat{b}(\gamma) + \hat{a}(\gamma)x$ of the regression line based on the $n - 2m$ points in the strip and the sum $Q(\gamma)$ of the squared residues. Define the Least Trimmed Squares estimate



(a) m-trimmed Student averages

(b) RM(21)$(3,4)$

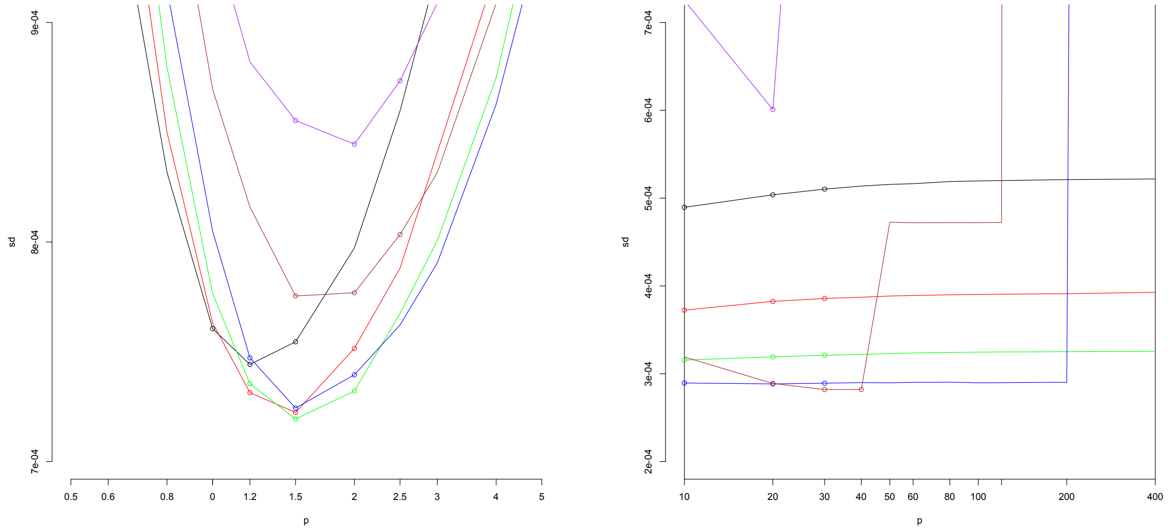
Figure 7: On the left: $\bar{\alpha}$, for ten batches of 10^5 simulations of a sample of 100 Student variables scaled by their IQD and trimmed by m on both sides, is minimal for $m = 1, 23, 29, 34, 39, 44, 46, 48, 49$ for tail index $\eta = 0, 1/3, 1/2, 2/3, 1, 3/2, 2, 3, 4$. On the right a bimodal loglog frequency plot for $\hat{a}_{\mathrm{TLS}(m,p)}$ for $(\xi, \eta) = (3, 3)$, $(m, p) = (22, 1)$ and Pareto errors. The red (green) curve describes the 370659 negative (629341 positive) outcomes. Here $\bar{\alpha} = 0.0029[1]$ for the state function $\log Q - p \sum\{1/i \mid \mathbf{z}_i \in S\}$ with $p = 1$ and $0.0137[1]$ without reward ($p = 0$). The minimal value is $0.00016[5]$ for $p = 15$. An extra parameter yields $\bar{\alpha} = 0.000020[1]$ for $(m, p, r) = (25, 30, 0.6)$.

for this value of the trimming parameter m as the regression line $L(\gamma)$ for which $Q(\gamma)$ is minimal.

Recall from the section on the Weighted Balance estimators that for unexceptional configurations of a sample of size n the slopes of the lines passing through two sample points are distinct. The set Γ of these slopes divides $\mathbb{R} \setminus \Gamma$ into $\binom{n}{2} + 1$ open intervals, the components of $\mathbb{R} \setminus \Gamma$. The set of sample points above the strip S with slope γ does not change when γ varies over one of these components, neither does the set of sample points below S. It suffices to choose a point γ_i in each of the components, compute the LS regression line $\hat{L}(\gamma_i)$ associated with this value γ_i, and the sum $Q(\gamma_i)$ of squared residues over the sample points in the strip with slope γ_i. Then $\hat{a}_{\text{LTS}(m)}$ is the slope of the line $\hat{L}(\gamma_i)$ for which $Q(\gamma_i)$ is minimal. The functions $\hat{a}(\gamma)$, $\hat{b}(\gamma)$ and $Q(\gamma)$ are constant on the components of $\mathbb{R} \setminus \Gamma$. There are differences with the theory of WB-estimators. The function Q is roughly V-shaped and the minimum is almost surely unique if X and Y have continuous dfs but Q is not decreasing to the left of its minimum or increasing to the right. The sum Q of the squared residuals in the strip S depends only on the way in which the strip S partitions the sample into three subsets of m, $n - 2m$ and m points, but in order to determine the minimum one has to compute Q for all these different partitions. This makes TLS a computer intensive procedure. Note too that LTS uses the points inside the strip to estimate the regression line and WBm the points outside the strip.

As in the example above on the trimmed average of a hundred Student variables we vary m to determine the optimal value, the value $m = m_0$ for which the average loss of $\hat{a}_{\text{LTS}}(m)$ over a million simulations is minimal.

Instead of selecting the strip $S = S(\gamma)$ with slope γ which minimizes $Q(\gamma)$, the sum of the squared residuals of the sample points in S, one could minimize the sum of the absolute values of the residuals, or the width of S, the difference between the maximal and the minimal residual. One can also choose the closed strip $S(\gamma)$ minimal. That would ensure that the boundary lines of S would each contain a sample points (for $\gamma \in \mathbb{R} \setminus \Gamma$). Now choose $\hat{a} = \gamma$ to be the slope of the strip $S(\gamma)$ with minimal vertical height. The minimal vertical height is not constant on the components of $\mathbb{R} \setminus \Gamma$ but increasing or decreasing depending on the positions of the two sample points on the boundary. This yields the Least Central Strip estimator mentioned in the Introduction. For all these estimators one then can ask how the optimal value of the trimming parameter m depends on the tail indices (ξ, η).



(a) Student errors

(b) Pareto errors

Figure 8: The average square root $\bar{\alpha}$ of the average loss for $\hat{a}_{\mathrm{LTS}}$ for Student (left, $m = 30$) and Pareto (right, $m = 29$) errors for $(\xi, \eta) = (3, 3)$ and ten batches of 10^5 simulations. The curves describe the behaviour of $\bar{\alpha}$ as a function of p for different values of $r = 0.15, 0.2, 0.25, 0.3, 0.4, 0.5$ (Student) and $r = 3 : 8/10$ (Pareto) with colours black, red, green, blue, brown, purpe. The erratic behaviour on the right is symptomatic for Pareto errors.

In this paper we restrict attention to LTS. By insisting on a minimal value of the squared residuals of the remaining points the estimator does not pay special attention to the rightmost sample points. To compensate for this neglect we introduce a reward. Divide Q by the product P of $1 + 1/i$ over the indices i of the $n - 2m$ points in the strip. Since one wants to minimize Q the reward for including the rightmost point is generous: Q is halved. One can introduce a positive parameter p to temper the reward, replacing $1 + 1/i$ by $(1 + 1/i)^p$. With the extra parameter p one can reduce the loss, but for certain values of the tail indices the loglog frequency plot of $|\hat{a}(m, p)|$ for Pareto errors and for optimal m and p turns out to be bimodal, suggesting a dichotomy: choose γ to minimize

Q or to include $\mathbf{z}_1$ in the strip. See Figure 7b.

The reward used in this paper is more complex. It depends on two positive parameters p and r. We choose the strip $S(\gamma)$ which minimizes the state function

$$T_{p,r}(\gamma) = \log Q(\gamma) - p \sum_{\mathbf{z}_i \in S(\gamma)} \frac{1}{1 + r(i-1)}. \tag{5.1}$$

The parameter r regulates how fast the reward decreases with the index i. The optimal values lie in $(0, 2)$. The parameter p determines the total effect on the state function. It may exceed 100. Now there is a good fit of the distribution of $\log|\hat{a}|$ in the class of EGBP-distributions.

Given the trimming parameter m one can compute for each γ a vector of T-values $T_{p,r}$, $(p, r) \in \Delta$, and estimates $\hat{a}_{p,r}$ for a finite set $\Delta \subset (0, \infty)^2$. By appropriate updating one obtains a vector $\hat{a}_{\mathrm{TLS}(m,p,r)}$, $(p, r) \in \Delta$. Now compute the empirical sd for ten batches of hundred thousand simulations. The square root $s = s(p, r)$ of the average loss over the million simulations is a function of (p, r). In Figure 8 $\Delta = \{p_1, \dots, p_{11}\} \times \{r_1, \dots, r_7\}$ and we plot $s(p, r)$ as a function of p for various values of r.

The performance of the resulting estimator WLTS is impressive.

The erratic dependence of p on the tail indices (ξ, η) suggests that it might be difficult to determine a good parameter triple (m, p, r) for a given sample of size $n = 100$ even if one has good estimates of the tail indices. The difference between the left and right plots in Figure 8 suggests that the parameters p and r may be sensitive to changes in the distribution of the error. For $(\xi, \eta) = (1, 1)$ the optimal trimming parameter m is the same for Student and Pareto errors, but the optimal values of the parameter p in the reward differ by a factor a thousand. In Figure 9 we plot in one figure the square root of the average loss in the estimate $\hat{a}_{\mathrm{WLTS}}$ for Student and for Pareto errors for various values of r as a function of p. The minimal values for the two error distributions are not far apart but the difference in the optimal value of p and the difference in the structure of the graphs make it difficult to see how one should choose good parameter values if the distribution of the error is not known. These features place the WLTS(m, p, r) estimator hors concours.

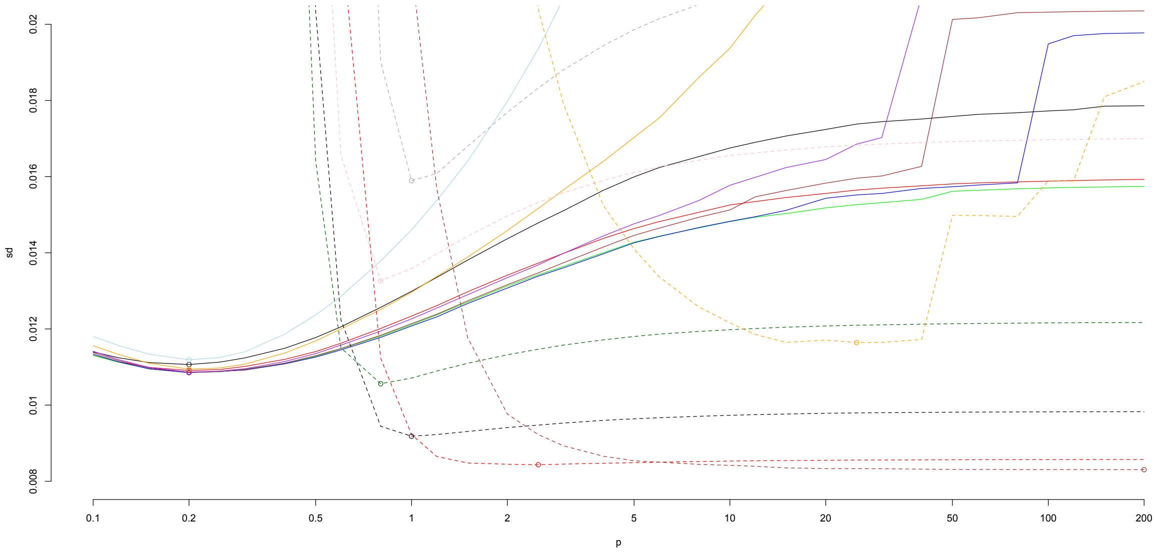


Figure 9: The square root of the average loss for WLTS$(16, p, r)$ at $(\xi, \eta) = (1, 1)$ for Student errors (full lines) and Pareto errors (dashed) scaled by the IQD for p ranging from 0.1 to 200 and $r = 0.01$ (grey), 0.02 (pink), 0.05 (dark green), 0.1 (black), 0.2 (red), 0.3 (green), 0.4 (blue), 0.5 (brown), 0.6 (purple), 1 (orange), 2 (azure). The optimal values of (p, r) are $(0.2, 0.4)$ (Student) and $(200, 0.5)$ (Pareto).

We now turn to trimmed LAD. Let us first say a few words on terminology. Trimmed Least Squares as opposed to Least Trimmed Squares was investigated in [23]. In that paper two procedures are compared. The trimming is based on a preliminary estimate of the slope of the regression line or on the Koenker-Bassett regression quantiles. The term Least Trimmed Squares introduced by Rousseeuw makes clear that one minimizes over all possible trimmings.

The LAD estimate is a bisector passing through two sample points. Hence we consider trimming around a bisector. Given a positive integer $m < n/2 - 1$, for each bisector L of the sample we consider the minimal closed strip $S \supset L$ such that m sample points lie in the open half plane above S and m in the open half plane below. The boundary lines of S contain one sample point each. Define the state variable $T = T_m$ as the sum of the absolute residuals of the $n - 2m$ points in the strip with respect to the bisector L. Note that L also is a bisector of the sample restricted to the strip S. Define $\hat{a}_m$ to be the slope of the bisector L for which the state variable T_m is minimal. The optimal value of the trimming parameter m is random. It minimizes the average loss over a million simulations. If the error has a symmetric Student distribution the optimal value of m is approximately 25. This holds for all values of the tail indices ξ and η. It also holds if we define the state function to be the sum of the squared residuals, or the difference between the maximal and minimal residual values. We therefore define the TB1, TB2 and TB∞ estimator of the regression line as the bisector L for which the corresponding state function $T = T_m$ is minimal for $m = 25$. These three Trimmed Bisector estimators are based on trimming around a bisector. The difference between the three estimators is small.

It is not clear why the optimal value of m for trimming around bisectors should be $m = 25$. This value need not be optimal if the errors have a Pareto distribution.

6 Tables

There are three sets of six tables listing the empirical sd and bias for various estimators and error distributions for the six values $\xi = 0, 1/2, 1, 3/2, 2, 3$ of the tail exponent of the explanatory variable.

The value of ξ can be estimated from the data. This determines the table which applies. In the first two sets the error has a symmetric Student distribution; in the third a Pareto distribution. The entries typically have the form $m/10^k[d]$ where $x_E = m/10^k$ is the mean of the ten quantities $\gamma_i = \sqrt{A_r}$ with $A_r = (\hat{a}_1^2 + \cdots + \hat{a}_r^2)/r$ the average loss for $\hat{a} = \hat{a}_E$ over a batch of $r = 10^5$ simulations of a sample of a hundred observations (X_i, Y_i^*). Here $\hat{a} = \hat{a}_E$ is the slope of the regression line estimated by E. The true regression line is the horizontal axis. The digit $d \in \{1, 2, 5\}$ gives an indication of the size of the fluctuations in the ten observed values γ_i. The sd of these ten values is rounded to $d/10^k$. See (1.3) for the precise prescription.

The entries in the tables depend on the seed used in the simulations. We have used the seeds $2223, \ldots, 1002222$ for the ten batches of a hundred thousand simulations. A different sequence of seeds will give different outcomes. The difference will in general be of the order of d in the last digit of m.

Colours are used to make the information in the tables more accessible. The value $x_E = m/10^k$ may be a bad indicator of the performance. The entry $0[1e + 17]$ for Least Squares at $(\xi, \eta) = (0, 3)$ describes a poorer performance than $0.0567[5]$ for the Trimmed LAD, TB1, in the same row. In each row let y_* denote the minimum of the sums $(m+3d)/10^k$ over the six entries in the row corresponding to the different estimators, and x_* the corresponding value of $x = m/10^k$. (In Tables 2 and 3 Weighted Least Trimmed Squares, WLTS, is excluded in determining the minimum.) The colour scheme is:

- red: $0 < x_E \leq x_*$ (minimum);
- green: $0 < x_E \leq y_*$ (indistinguishable from the minimum);
- blue: $0 < x_E \leq 5y_*/4$ (excellent);

- purple: $0 < x_E \leq 2y_*$ (good).

A colourful table indicates that there are quite a few estimators which perform well.

The estimators are geometric. The estimated regression line does not change if one changes the coordinates, though the coordinates of the line do, see (1.2). The estimates of LADPC and LADHC are aberrant, they are not invariant under translations of the horizontal axis, see (3.4).

In the second and third set of tables the estimator depends on a parameter. The parameter may vary with the values of the tail indices ξ and η. It is chosen to minimize the average loss. The value of this optimal parameter is given in the fourth set of tables.

It may seem foolish to choose a quadratic loss function to measure the performance of estimators in an environment where the underlying variables have heavy tails. The results in the tables seem to justify the choice.

6.1 Table 1, estimators without parameters, Student errors

LS, Least Squares, and LAD, Least Absolute Deviation, are treated in Section 2. LAD is a Weighted Balance estimator with weight X_i; LADPC, power corrected LAD uses the weight $X_i^{1/12}$. It is geometric with respect to the subgroup $\mathcal{G}_0$ of transformations which map the right half plane onto itself. It is only defined when the explanatory variables are positive, see Section 3. TS, Theil-Sen, is a robust estimator which chooses the regression line for which Kendall's tau vanishes, rather than the covariance as in LS. The slope is the median of the slopes of the 4950 line segments connecting two sample points, see Section4. The three estimators $\mathrm{TB1}, \mathrm{TB2}, \mathrm{TB}\infty$ are based on samples for which 25 points above and 25 points below the bisector have been trimmed. The estimator chooses the bisector for which a certain state function T of the residuals is minimal. For TB1 T is the sum of the absolute values of the fifty remaining points, for TB2 the sum of the squares and for TB∞ T is the width, the difference between the largest and smallest residual, see Section 5.

LS is optimal if and only if the error has a Gaussian distribution. LAD is good when η

is small and the power correction LADPC is good when η is large. TS and TB1 are good, but not as good as LADPC. For the three estimators trimmed around the bisector TB∞ is best for small values of η, TB1 for large values of η and TB2 for values in between.

$\eta \setminus 0$	LS	LAD	TS	TB1	TB2	TB∞
0	0.0774[2]	0.0969[2]	0.0912[2]	0.2578[5]	0.2371[5]	0.2227[5]
1/3	0.1179[5]	0.0959[2]	0.1001[2]	0.2094[5]	0.1916[5]	0.1896[5]
1/2	0.26[5]	0.0951[2]	0.1041[5]	0.1888[5]	0.1726[5]	0.1770[5]
2/3	1.6[5]	0.0941[2]	0.1080[2]	0.1697[5]	0.1561[5]	0.1669[5]
1	200[100]	0.0917[2]	0.1147[2]	0.1390[5]	0.1302[5]	0.1516[5]
3/2	1e+6[2e+6]	0.0869[2]	0.1218[5]	0.1062[5]	0.1056[5]	0.1407[5]
2	3e+9[5e+9]	0.0810[5]	0.1257[5]	0.0838[5]	0.0936[5]	0.139[1]
3	0[$1e+17$]	0.0681[5]	0.1242[5]	0.0567[5]	0.095[1]	0.153[1]
4	2e+24[5e+24]	0.0560[5]	0.1159[5]	0.0417[5]	0.124[5]	0.192[5]

$\eta \setminus 1/2$	LS	LAD	LADPC	TS	TB1	TB2	TB∞
0	0.0518[1]	0.0641[2]	0.0711[2]	0.0822[2]	0.258[1]	0.230[1]	0.2159[5]
1/3	0.0788[5]	0.0670[2]	0.0720[2]	0.0913[2]	0.2044[5]	0.1829[5]	0.1824[5]
1/2	0.17[2]	0.0690[2]	0.0723[2]	0.0957[5]	0.1827[5]	0.1640[5]	0.170[1]
2/3	1.0[5]	0.072[1]	0.0726[2]	0.1001[2]	0.163[1]	0.148[1]	0.1604[5]
1	120[50]	0.1[1]	0.0727[2]	0.1083[5]	0.1331[5]	0.1236[5]	0.1452[5]
3/2	0[$1e+6$]	3[5]	0.0716[2]	0.1182[5]	0.1023[5]	0.1026[5]	0.135[1]
2	1e+9[1e+9]	40[50]	0.0695[5]	0.1253[5]	0.0835[5]	0.0960[5]	0.134[1]
3	1e+16[1e+16]	1e+7[2e+7]	0.0624[5]	0.1298[5]	0.0638[5]	0.113[2]	0.153[2]
4	2e+23[5e+23]	2e+10[5e+10]	0.055[1]	0.1257[5]	0.057[2]	0.160[5]	0.21[1]

$\eta \setminus 1$	LS	LAD	LADPC	TS	TB1	TB2	TB∞
0	0.00703[5]	0.00861[5]	0.01189[5]	0.01660[5]	0.0647[2]	0.0558[2]	0.0514[2]
1/3	0.0106[2]	0.00943[5]	0.01224[5]	0.0187[1]	0.0491[2]	0.0424[2]	0.0424[2]
1/2	0.022[5]	0.0104[5]	0.01239[5]	0.0197[1]	0.0432[2]	0.0376[2]	0.0393[1]
2/3	0.1[1]	0.012[1]	0.01259[5]	0.02084[5]	0.0382[2]	0.0336[2]	0.0369[2]
1	10[10]	1[1]	0.0130[1]	0.0230[1]	0.0305[5]	0.0279[5]	0.0333[1]
3/2	30000[20000]	20[20]	0.01328[5]	0.0260[1]	0.0234[2]	0.0238[5]	0.0312[5]
2	1e+8[2e+8]	10000[10000]	0.01339[5]	0.0286[1]	0.0197[2]	0.0237[2]	0.0316[5]
3	0[$1e+16$]	2e+6[2e+6]	0.0130[1]	0.0314[2]	0.0171[2]	0.032[1]	0.038[1]
4	2e+23[5e+23]	0[$2e+15$]	0.0123[5]	0.0320[1]	0.0182[5]	0.049[1]	0.055[1]

$\eta \setminus 3/2$	LS	LAD	LADPC	TS	TB1	TB2	TB∞
0	0.00122[1]	0.00149[1]	0.00260[2]	0.00432[2]	0.0220[2]	0.0183[1]	0.0164[1]
1/3	0.00184[5]	0.00167[2]	0.00271[2]	0.00491[2]	0.0160[1]	0.0133[1]	0.0131[1]
1/2	0.004[1]	0.00188[5]	0.00276[2]	0.00521[2]	0.0137[1]	0.0116[1]	0.0120[1]
2/3	0.02[2]	0.0028[5]	0.00282[2]	0.00556[2]	0.0120[1]	0.0103[1]	0.0112[1]
1	3[5]	0.1[1]	0.00296[2]	0.00626[5]	0.0093[2]	0.0085[1]	0.0101[1]
3/2	5000[5000]	10[10]	0.00313[2]	0.00730[5]	0.0072[1]	0.0074[1]	0.0095[1]
2	2e+7[5e+7]	1e+6[2e+6]	0.00328[2]	0.00830[5]	0.0063[1]	0.0078[1]	0.0099[2]
3	1e+15[2e+15]	1e+8[2e+8]	0.00344[5]	0.00968[5]	0.0062[2]	0.0117[5]	0.0129[5]
4	0[$1e+23$]	3e+12[5e+12]	0.00347[5]	0.0104[1]	0.0074[2]	0.0187[5]	0.020[1]

$\eta \setminus 2$	LS	LAD	LADPC	TS	TB1	TB2	TB∞
0	0.000243[5]	0.000296[5]	0.000645[5]	0.00126[1]	0.0086[1]	0.00685[5]	0.00597[5]
1/3	0.00036[1]	0.00034[1]	0.00068[1]	0.00145[1]	0.00596[5]	0.00482[5]	0.00463[5]
1/2	0.0008[5]	0.00041[5]	0.000693[5]	0.00154[1]	0.00502[5]	0.00415[5]	0.00422[5]
2/3	0.004[5]	0.001[1]	0.00071[1]	0.00165[1]	0.00431[5]	0.00363[5]	0.00388[5]
1	1[1]	0.01[2]	0.00075[1]	0.00189[2]	0.0033[1]	0.00295[5]	0.00348[5]
3/2	1000[1000]	1[1]	0.00082[1]	0.00227[1]	0.00254[5]	0.00266[5]	0.00334[5]
2	0[$1e+7$]	0[100000]	0.00089[1]	0.00266[5]	0.00229[5]	0.00292[5]	0.0036[1]
3	1e+14[5e+14]	0[$1e+7$]	0.00101[2]	0.00328[2]	0.0025[1]	0.0047[2]	0.0050[2]
4	1e+22[2e+22]	1e+15[2e+15]	0.00108[5]	0.00372[5]	0.0032[2]	0.0079[2]	0.0081[5]

$\eta \setminus 3$	LS	LAD	LADPC	TS	TB1	TB2	TB∞
0	0.0000129[5]	0.000016[1]	0.000049[1]	0.000133[2]	0.00161[5]	0.00122[2]	0.00099[2]
1/3	0.000019[2]	0.000018[1]	0.000053[1]	0.000155[2]	0.00103[2]	0.00079[2]	0.00071[2]
1/2	0.00004[2]	0.000022[1]	0.000054[1]	0.000166[2]	0.00084[2]	0.00065[1]	0.00063[2]
2/3	0.0002[2]	0.00005[5]	0.000055[2]	0.000179[5]	0.00070[1]	0.00057[1]	0.00058[2]
1	0.02[5]	0.001[1]	0.000060[2]	0.000209[5]	0.00051[2]	0.00045[1]	0.00051[2]
3/2	30[50]	0.1[1]	0.000068[2]	0.000261[5]	0.00040[1]	0.00042[1]	0.00051[1]
2	100000[200000]	0[1000]	0.000077[2]	0.00032[1]	0.00038[2]	0.00051[5]	0.00057[5]
3	0[$1e + 13$]	300000[500000]	0.00010[1]	0.00044[1]	0.00049[5]	0.0009[1]	0.0009[1]
4	1e+20[5e+20]	1e+13[2e+13]	0.00012[1]	0.00055[2]	0.0007[1]	0.0016[1]	0.0016[5]

Empirical sd of the slope $\hat{a}$ for LS, LAD, LADPC, TS, TB1, TB2 and TB∞.
Sample size $n = 100$, $\xi = 0, 1/2, 1, 3/2, 2, 3$. Student errors with tail index η, scaled by their IQD.

6.2 Table 2, estimators with a parameter, Student errors

The Hyperbolic balance estimators HB40(d) and HB0(d) are comparable. They use the weight sequence $1/d, 1/(d+1), \dots, 1/(d-1+n)$ see (3.1). The parameter d depends on the value of the tail indices. The estimator HB0 performs slightly better than HB40 when η is large. This may be due to the fact that for large η the Student density at the 0.4 and 0.6 quantiles is much smaller than at the median. The Right Median estimator RM(r) chooses the bisector which divides the $r = 2r_0 + 1$ rightmost (red) points equally into two sets of r_0 with one red point, the median, on the bisector, see Section 3. It is intuitive but its performance is less good than that of the Hyperbolic Balance estimators. LADHC, the Hyperbolic Correction of LAD is indistinguishable from the Gap Correction, LADGC, see (3.10). It performs well when η is large. Apart from these four weighted balance estimators we consider the weighted Theil-Sen estimator WTS(p) introduced in Section 4 which performs well for small values of η. The Least Trimmed Squares, LTS(m, p, r), yields the smallest empirical sd's but is hors concours since it is not clear how its parameters should be chosen, see Section 5.

$\eta \setminus 0$	RM[r]	HB40[d]	HB0[d]	LADHC[d]	WTS[p]	WLTS[m, p, r]
0	0.1215[5]	0.0916[2]	0.0993[5]	0.0972[2]	0.0810[2]	0.0706[2]
1/3	0.1200[5]	0.0918[2]	0.0979[2]	0.0964[2]	0.0894[2]	0.0722[2]
1/2	0.1188[2]	0.0917[2]	0.0969[2]	0.0955[2]	0.0935[2]	0.0726[2]
2/3	0.1173[5]	0.0913[2]	0.0956[2]	0.0944[2]	0.0973[2]	0.0727[2]
1	0.1140[5]	0.0904[2]	0.0927[2]	0.0918[2]	0.1044[2]	0.0717[1]
3/2	0.1072[5]	0.0878[2]	0.0868[2]	0.0862[2]	0.1128[5]	0.0680[2]
2	0.0991[5]	0.0840[2]	0.0800[2]	0.0797[5]	0.1190[5]	0.0626[2]
3	0.0806[5]	0.0738[5]	0.0645[5]	0.0643[5]	0.125[1]	0.0475[2]
4	0.0633[5]	0.0626[2]	0.0498[2]	0.0497[2]	0.129[1]	0.0331[2]

$\eta \setminus 1/2$	RM[r]	HB40[d]	HB0[d]	LADHC[d]	WTS[p]	WLTS[m, p, r]
0	0.0889[5]	0.0632[1]	0.0687[2]	0.0658[2]	0.0553[1]	0.0398[1]
1/3	0.0930[5]	0.0663[2]	0.0710[2]	0.0680[2]	0.0632[2]	0.0421[1]
1/2	0.0944[5]	0.0677[2]	0.0717[2]	0.0691[2]	0.0674[2]	0.0428[1]
2/3	0.0955[5]	0.0689[2]	0.0725[5]	0.0704[2]	0.0717[2]	0.0430[1]
1	0.0973[2]	0.0712[2]	0.0735[2]	0.0719[2]	0.0801[2]	0.0424[1]
3/2	0.0972[5]	0.0729[2]	0.0730[2]	0.0725[2]	0.0917[5]	0.0396[1]
2	0.0947[5]	0.0731[5]	0.0709[2]	0.0711[2]	0.1018[2]	0.0356[1]
3	0.0833[5]	0.0692[5]	0.0621[5]	0.0625[5]	0.117[1]	0.0264[1]
4	0.0706[5]	0.0624[5]	0.0512[5]	0.0512[2]	0.129[1]	0.01825[5]

$\eta \setminus 1$	RM[r]	HB40[d]	HB0[d]	LADHC[d]	WTS[p]	WLTS[m, p, r]
0	0.0116[1]	0.00870[5]	0.00952[5]	0.00920[5]	0.00754[2]	0.00706[5]
1/3	0.0138[1]	0.00957[5]	0.01030[5]	0.00980[5]	0.00894[5]	0.00860[5]
1/2	0.0147[1]	0.01009[5]	0.01071[5]	0.01020[5]	0.00978[5]	0.00936[5]
2/3	0.0156[1]	0.01066[5]	0.01119[5]	0.01069[5]	0.01071[5]	0.0100[1]
1	0.01703[5]	0.01160[5]	0.01203[5]	0.01171[5]	0.0127[1]	0.0109[1]
3/2	0.0187[2]	0.0128[1]	0.0129[1]	0.01270[5]	0.0156[2]	0.0110[1]
2	0.0197[5]	0.01375[5]	0.01346[5]	0.01328[5]	0.0185[1]	0.01040[5]
3	0.0195[5]	0.0144[2]	0.0133[2]	0.0132[1]	0.0240[5]	0.00839[5]
4	0.0180[2]	0.0141[2]	0.0120[2]	0.0119[1]	0.0288[5]	0.00617[5]

$\eta \setminus 3/2$	RM[r]	HB40[d]	HB0[d]	LADHC[d]	WTS[p]	WLTS[m,p,r]
0	0.00189[5]	0.00153[1]	0.00168[2]	0.00176[1]	0.00133[1]	0.00123[1]
1/3	0.00250[5]	0.00174[2]	0.00187[2]	0.00188[2]	0.00160[2]	0.00159[2]
1/2	0.00277[5]	0.00188[2]	0.00200[2]	0.00197[2]	0.00178[1]	0.00183[2]
2/3	0.00307[5]	0.00205[2]	0.00216[2]	0.00208[2]	0.00201[2]	0.00204[2]
1	0.00362[5]	0.00236[5]	0.00245[2]	0.00236[2]	0.00250[5]	0.00236[2]
3/2	0.0044[1]	0.00278[5]	0.00285[5]	0.00278[5]	0.00330[5]	0.00261[5]
2	0.0050[2]	0.00316[5]	0.00314[2]	0.00311[2]	0.00418[5]	0.00269[5]
3	0.0056[2]	0.0037[1]	0.0035[1]	0.00345[5]	0.0062[2]	0.00243[5]
4	0.0056[2]	0.00394[5]	0.00343[5]	0.00343[5]	0.0082[5]	0.00200[5]

$\eta \setminus 2$	RM[r]	HB40[d]	HB0[d]	LADHC[d]	WTS[p]	WLTS[m,p,r]
0	0.00036[1]	0.000306[5]	0.00034[1]	0.000384[5]	0.000271[5]	0.000245[5]
1/3	0.00052[1]	0.00036[1]	0.00039[1]	0.00043[1]	0.000328[5]	0.00033[1]
1/2	0.00058[1]	0.00040[1]	0.000423[5]	0.000448[5]	0.000370[5]	0.00038[1]
2/3	0.00069[5]	0.00044[1]	0.00047[1]	0.000476[5]	0.000426[5]	0.00044[1]
1	0.00084[2]	0.00053[2]	0.00056[2]	0.00054[1]	0.00055[2]	0.00054[2]
3/2	0.0012[1]	0.00066[2]	0.00069[2]	0.00067[2]	0.00077[2]	0.00065[2]
2	0.00137[5]	0.00081[1]	0.00080[1]	0.00082[1]	0.00104[2]	0.00070[2]
3	0.0018[1]	0.00107[5]	0.00102[5]	0.00100[2]	0.0018[1]	0.00072[2]
4	0.00191[5]	0.00122[5]	0.00107[2]	0.00108[2]	0.0025[1]	0.00066[1]

$\eta \setminus 3$	RM[r]	HB40[d]	HB0[d]	LADHC[d]	WTS[p]	WLTS[m,p,r]
0	0.000018[1]	0.000017[1]	0.000018[1]	0.000027[1]	0.0000150[5]	0.0000130[5]
1/3	0.000025[2]	0.000020[1]	0.000021[1]	0.000030[1]	0.000018[1]	0.000018[1]
1/2	0.000034[1]	0.000022[1]	0.000024[1]	0.000032[2]	0.0000210[5]	0.000021[2]
2/3	0.000040[5]	0.000026[1]	0.000029[2]	0.000033[2]	0.000025[1]	0.000025[1]
1	0.000054[2]	0.000032[2]	0.000035[2]	0.000038[2]	0.000033[1]	0.000033[2]
3/2	0.000086[5]	0.000046[5]	0.000050[5]	0.000049[2]	0.000054[5]	0.000046[5]
2	0.000122[5]	0.00006[1]	0.00007[1]	0.000065[2]	0.000077[5]	0.000053[5]
3	0.00020[2]	0.00011[2]	0.00010[1]	0.00010[1]	0.00017[5]	0.000069[5]
4	0.00024[1]	0.00013[1]	0.00012[1]	0.00012[1]	0.00025[5]	0.000073[5]

Empirical sd of the slope for RM(r), HB40(d), HB0(d), LADHC(d), WTS(p) and WLTS(m, p, r). Sample size $n = 100$, $\xi = 0, 1/2, 1, 3/2, 2, 3$. Student errors with tail index η, scaled by their IQD.

6.3 Table 3, estimators with a parameter, Pareto errors

The estimators are the same as for Table 2 but now applied to Pareto errors. Weighted Theil-Sen is the best choice here. It yields the minimal square root of the average loss for 51 of the 54 rows. The bias is positive, roughly one tenth of the sd. Only WLTS has a negative bias.

$\eta \setminus 0$	RM[r]	HB40[d]	HB0[d]	LADHC[d]	WTS[p]	WLTS[m, p, r]
0	0.1194[5]	0.0877[2]	0.0969[2]	0.0949[2]	0.0712[1]	0.0778[2]
	0.0096[5]	0.0086[5]	0.0079[5]	0.0088[2]	0.0089[2]	-0.0280[5]
1/3	0.1165[5]	0.0844[2]	0.0942[2]	0.0927[2]	0.0679[1]	0.0823[2]
	0.0122[5]	0.0096[5]	0.0100[5]	0.0112[2]	0.0099[2]	-0.0462[5]
1/2	0.1149[5]	0.0826[2]	0.0926[2]	0.0910[1]	0.0640[1]	0.0823[2]
	0.0134[5]	0.0097[5]	0.0109[5]	0.0118[2]	0.0093[2]	-0.0561[2]
2/3	0.1128[2]	0.0806[2]	0.0906[2]	0.0893[5]	0.0604[2]	0.0812[2]
	0.0129[5]	0.0104[5]	0.0110[2]	0.0120[1]	0.0088[2]	-0.0650[2]
1	0.1085[5]	0.0761[5]	0.0865[2]	0.0854[5]	0.0517[2]	0.0702[5]
	0.0145[5]	0.0104[2]	0.0118[2]	0.0124[2]	0.0068[2]	-0.0890[2]
3/2	0.1006[5]	0.0685[2]	0.0793[2]	0.0786[2]	0.0422[2]	0.072[5]
	0.0137[2]	0.0096[2]	0.0125[2]	0.0133[1]	0.0055[1]	-0.0762[2]
2	0.0917[5]	0.0601[2]	0.0715[5]	0.0708[5]	0.0333[2]	0.055[1]
	0.0128[2]	0.0087[2]	0.0123[2]	0.0126[2]	0.0039[1]	-0.0657[1]
3	0.0733[5]	0.0434[2]	0.0556[5]	0.0553[2]	0.0196[1]	0.0376[5]
	0.0108[2]	0.0062[1]	0.0110[2]	0.0103[2]	0.00149[5]	-0.0417[1]
4	0.0571[5]	0.0296[2]	0.0412[5]	0.0408[2]	0.0115[1]	0.0236[2]
	0.0071[2]	0.0040[1]	0.0082[1]	0.0078[2]	0.00054[2]	-0.02674[5]

$\eta \setminus 1/2$	RM[r]	HB40[d]	HB0[d]	LADHC[d]	WTS[p]	WLTS[m,p,r]
0	0.0918[2]	0.0626[2]	0.0690[2]	0.0658[2]	0.0498[1]	0.0517[2]
	0.0134[5]	0.0101[2]	0.0105[2]	0.0098[2]	0.0077[2]	-0.0131[2]
1/3	0.0947[2]	0.0636[2]	0.0704[2]	0.0671[2]	0.0518[2]	0.0534[2]
	0.0165[5]	0.0117[2]	0.0129[2]	0.0128[2]	0.0099[2]	-0.0218[2]
1/2	0.0956[5]	0.0639[2]	0.0710[5]	0.0675[2]	0.0529[2]	0.0533[2]
	0.0172[5]	0.0124[2]	0.0143[2]	0.0140[2]	0.0108[2]	-0.0263[1]
2/3	0.0964[5]	0.0639[5]	0.0710[5]	0.0682[5]	0.0524[5]	0.0535[5]
	0.0179[5]	0.0134[2]	0.0149[2]	0.0154[1]	0.0110[2]	-0.0284[1]
1	0.0969[5]	0.0632[5]	0.0709[5]	0.0689[5]	0.0470[5]	0.0517[2]
	0.0196[5]	0.0138[2]	0.0163[2]	0.0171[2]	0.0095[2]	-0.0328[2]
3/2	0.0951[5]	0.0609[5]	0.0693[5]	0.0669[5]	0.0374[2]	0.0463[2]
	0.0199[5]	0.0137[2]	0.0175[2]	0.0175[2]	0.0064[1]	-0.0351[1]
2	0.091[1]	0.0569[5]	0.0662[5]	0.0639[5]	0.0309[2]	0.0423[5]
	0.0203[5]	0.0128[2]	0.0178[2]	0.0165[2]	0.0048[1]	-0.0290[1]
3	0.079[1]	0.0453[5]	0.0560[5]	0.0549[5]	0.0197[2]	0.0299[2]
	0.0174[2]	0.0098[2]	0.0156[2]	0.0146[2]	0.00217[5]	-0.0183[1]
4	0.064[1]	0.0325[5]	0.0444[5]	0.0434[2]	0.0122[1]	0.0195[2]
	0.0127[2]	0.0066[1]	0.0122[2]	0.0115[2]	0.00102[5]	-0.01029[5]

$\eta \setminus 1$	RM[r]	HB40[d]	HB0[d]	LADHC[d]	WTS[p]	WLTS[m,p,r]
0	0.0134[1]	0.00907[5]	0.00995[5]	0.00944[5]	0.00692[5]	0.00733[5]
	0.00236[5]	0.00158[2]	0.00165[2]	0.00148[2]	0.00096[2]	-0.00080[2]
1/3	0.0155[1]	0.00987[5]	0.0108[1]	0.0102[1]	0.00748[5]	0.00770[5]
	0.00306[5]	0.00205[2]	0.00216[5]	0.00203[5]	0.00130[2]	-0.00161[2]
1/2	0.0162[1]	0.0102[1]	0.0113[1]	0.01053[5]	0.0078[1]	0.00786[5]
	0.00329[5]	0.00223[2]	0.00244[5]	0.00229[5]	0.00145[2]	-0.00178[2]
2/3	0.0170[2]	0.0105[1]	0.0116[1]	0.0111[1]	0.0081[1]	0.00788[5]
	0.00353[5]	0.00241[5]	0.00263[5]	0.00258[5]	0.00159[2]	-0.00209[2]
1	0.0183[2]	0.0111[1]	0.0123[1]	0.0121[1]	0.0089[2]	0.0078[1]
	0.00408[5]	0.00272[5]	0.00303[5]	0.00309[5]	0.00183[2]	-0.00257[5]
3/2	0.0196[2]	0.0115[1]	0.0131[2]	0.0134[2]	0.0100[5]	0.00729[5]
	0.00462[5]	0.00298[5]	0.00348[5]	0.00366[5]	0.00200[5]	-0.00329[2]
2	0.0203[2]	0.0115[1]	0.0135[2]	0.0130[2]	0.0086[5]	0.00670[5]
	0.0048[1]	0.00309[5]	0.00373[5]	0.00367[5]	0.00164[2]	-0.00300[2]
3	0.0194[5]	0.0104[2]	0.0128[2]	0.0123[2]	0.0048[1]	0.0052[1]
	0.0048[1]	0.00279[5]	0.00366[5]	0.00358[5]	0.00078[2]	-0.00233[2]
4	0.0172[5]	0.0083[2]	0.0111[2]	0.0107[1]	0.0030[1]	00361[5]
	0.00389[5]	0.00210[5]	0.00315[5]	0.00309[5]	0.00038[1]	-0.00184[1]

$\eta \setminus 3/2$	RM[r]	HB40[d]	HB0[d]	LADHC[d]	WTS[p]	WLTS[m,p,r]
0	0.00236[5]	0.00165[2]	0.00182[2]	0.00182[2]	0.00125[1]	0.00133[1]
	0.000368[5]	0.000247[5]	0.000269[5]	0.000237[5]	0.000146[5]	-0.000073[5]
1/3	0.00301[5]	0.00189[2]	0.00209[2]	0.00198[2]	0.00139[2]	0.00142[1]
	0.00052[1]	0.000342[5]	0.000373[5]	0.000329[5]	0.000203[5]	-0.000132[5]
1/2	0.00333[5]	0.00201[2]	0.00221[5]	0.00206[1]	0.00148[2]	0.00147[2]
	0.00060[1]	0.000385[5]	0.000417[5]	0.00037[1]	0.000232[5]	-0.000167[5]
2/3	0.0037[1]	0.00215[5]	0.00239[5]	0.00219[2]	0.00158[5]	0.00152[2]
	0.00069[1]	0.000434[5]	0.000475[5]	0.00043[1]	0.000259[5]	-0.000159[5]
1	0.0042[1]	0.00238[5]	0.00267[5]	0.00249[2]	0.00180[5]	0.00150[5]
	0.00085[2]	0.000518[5]	0.000569[5]	0.000530[5]	0.000311[5]	-0.000239[5]
3/2	0.0049[1]	0.00271[5]	0.00302[5]	0.0030[1]	0.0023[2]	0.00154[5]
	0.00106[1]	0.00063[1]	0.00069[1]	0.00068[1]	0.000381[5]	-0.000302[5]
2	0.0056[2]	0.00291[5]	0.0033[1]	0.0033[1]	0.0028[2]	0.00142[2]
	0.00122[2]	0.00071[1]	0.00080[1]	0.00078[1]	0.00041[1]	-0.000429[5]
3	0.0059[2]	0.0030[1]	0.0035[1]	0.00342[5]	0.003[1]	0.00113[1]
	0.00139[2]	0.00075[1]	0.00089[1]	0.00087[1]	0.000312[5]	-0.000380[5]
4	0.0058[2]	0.0026[1]	0.0034[1]	0.00329[5]	0.00100[5]	0.00082[2]
	0.00132[2]	0.00067[1]	0.00086[1]	0.00086[1]	0.000138[5]	-0.000297[2]

$\eta \setminus 2$	RM[r]	HB40[d]	HB0[d]	LADHC[d]	WTS[p]	WLTS[m, p, r]
0	0.00046[2]	0.000340[5]	0.000374[5]	0.000416[5]	0.000260[5]	0.000272[5]
	0.000054[1]	0.000041[1]	0.000044[1]	0.000042[2]	0.0000244[5]	-44e-7[5e-7]
1/3	0.00064[2]	0.00041[1]	0.00044[1]	0.000452[5]	0.00030[1]	0.000296[5]
	0.000089[2]	0.000059[1]	0.000063[1]	0.000059[1]	0.000035[1]	-0.000017[1]
1/2	0.00076[2]	0.00044[2]	0.00049[1]	0.000471[5]	0.00032[1]	0.000306[5]
	0.000113[2]	0.000069[1]	0.000074[1]	0.000068[2]	0.000040[1]	-0.000016[1]
2/3	0.00084[2]	0.00049[2]	0.00053[2]	0.00051[1]	0.00034[1]	0.000314[5]
	0.000130[2]	0.000079[1]	0.000084[1]	0.000079[2]	0.000046[1]	-0.000025[1]
1	0.00107[5]	0.00057[2]	0.00061[2]	0.00059[2]	0.00041[2]	0.000324[5]
	0.000186[5]	0.000099[1]	0.000107[2]	0.000099[1]	0.000057[1]	-0.000040[1]
3/2	0.00133[5]	0.00069[2]	0.00076[2]	0.00073[2]	0.00055[5]	0.00033[2]
	0.000249[5]	0.000129[2]	0.000140[2]	0.000132[2]	0.000073[1]	-0.000036[1]
2	0.0016[1]	0.00079[5]	0.00088[5]	0.00089[5]	0.0007[1]	0.00033[2]
	0.000304[5]	0.000157[2]	0.000174[2]	0.000164[5]	0.000084[2]	-0.000049[1]
3	0.0020[2]	0.00090[5]	0.00106[5]	0.00104[5]	0.0009[5]	0.000282[5]
	0.00039[1]	0.000191[5]	0.000215[5]	0.000207[5]	0.000077[2]	-0.000067[1]
4	0.0021[2]	0.00089[5]	0.00111[5]	0.00107[5]	0.00035[5]	0.000217[5]
	0.00043[1]	0.000193[5]	0.000235[5]	0.000227[2]	0.000043[1]	-0.000052[1]

$\eta \setminus 3$	RM$[r]$	HB40$[d]$	HB0$[d]$	LADHC$[d]$	WTS$[p]$	WLTS$[m,p,r]$
0	0.000024[2]	0.000019[1]	0.000021[1]	0.000030[1]	0.000015[1]	0.000015[1]
	16e-7[1e-7]	135e-8[5e-8]	144e-8[5e-8]	18e-7[1e-7]	86e-8[5e-8]	6e-8[5e-8]
1/3	0.00004[1]	0.000024[2]	0.000026[2]	0.000032[1]	0.000018[2]	0.000017[1]
	28e-7[2e-7]	207e-8[5e-8]	219e-8[5e-8]	25e-7[1e-7]	126e-8[5e-8]	-14e-8[5e-8]
1/2	0.000044[5]	0.000027[2]	0.000029[2]	0.000033[1]	0.000019[2]	0.000017[1]
	39e-7[1e-7]	25e-7[1e-7]	26e-7[1e-7]	28e-7[1e-7]	148e-8[5e-8]	-26e-8[5e-8]
2/3	0.000053[5]	0.000030[5]	0.000034[5]	0.000036[5]	0.000021[5]	0.000018[1]
	47e-7[1e-7]	30e-7[1e-7]	31e-7[1e-7]	33e-7[1e-7]	172e-8[5e-8]	-32e-8[5e-8]
1	0.00008[1]	0.00004[1]	0.00004[1]	0.000044[5]	0.000027[5]	0.000019[1]
	74e-7[2e-7]	39e-7[1e-7]	42e-7[1e-7]	43e-7[1e-7]	223e-8[5e-8]	-43e-8[5e-8]
3/2	0.00012[1]	0.000053[5]	0.00006[1]	0.000052[5]	0.00004[1]	0.000020[1]
	0.0000159[2]	57e-7[2e-7]	61e-7[2e-7]	59e-7[2e-7]	31e-7[1e-7]	-74e-8[5e-8]
2	0.00015[2]	0.00007[1]	0.00008[1]	0.000069[5]	0.00006[2]	0.000021[2]
	0.0000191[5]	75e-7[2e-7]	82e-7[2e-7]	78e-7[2e-7]	40e-7[2e-7]	-112e-8[5e-8]
3	0.00025[5]	0.00010[2]	0.00011[2]	0.00012[1]	0.00013[5]	0.000020[1]
	0.000027[1]	0.0000109[5]	0.0000126[5]	0.0000120[5]	59e-7[5e-7]	-159e-8[5e-8]
4	0.0006[5]	0.00012[2]	0.00014[2]	0.00014[1]	0.0003[2]	0.000018[1]
	0.000034[2]	0.0000131[5]	0.0000158[5]	0.0000149[2]	7e-6[1e-6]	-22e-7[1e-7]

Empirical sd and bias of the slope $\hat{a}$ for RM(r), HB40(d), HB0(d), LADHC(d), WTS(p) and WLTS(m,p,r). Sample size $n = 100$, Pareto errors with tail index η, scaled by their IQD.

6.4 Table 4, parameter values

The parameter values, like the empirical sd and the bias in the tables above, depend on the seeds used for the simulations. They are chosen to be optimal on the basis of one batch of a hundred thousand simulations. Since the dependence of the average loss on the parameter is often locally quadratic the values of the parameter are imprecise. The dependence on the tail indices is monotonic. The values for $\xi = 0$ are aberrant since here the $1/\sqrt{n}$ asymptotic normality will apply. We have not been able to establish a simple functional relationship between the parameter and the tail indices even if the values at $\xi = 0, 1/2$ are omitted. The parameter values for Pareto and Student errors differ. For the parameters p and r in the WLTS estimator this difference may be large.

$\eta \setminus \xi$	0	1/2	1	3/2	2	3
0	21	5	1	1	1	1
1/3	21	7	3	3	1	1
1/2	23	9	5	3	3	3
2/3	25	11	7	5	3	3
1	27	13	9	7	5	5
3/2	29	17	13	11	7	7
2	31	21	15	13	11	11
3	35	29	23	19	17	15
4	41	39	33	27	23	21

$\eta \setminus \xi$	0	1/2	1	3/2	2	3
0	11	4	1	0	0	0
1/3	11	5	2	1	1	0
1/2	11	6	3	2	2	1
2/3	13	7	4	2	2	1
1	13	8	5	4	4	2
3/2	16	11	7	5	5	5
2	18	12	11	6	6	5
3	20	16	13	11	9	6
4	23	20	18	14	13	6

Parameter $r = 2r0 + 1$ for Right Median, RM(r), with Student (left) and Pareto (right) errors.

$\eta \setminus \xi$	0	1/2	1	3/2	2	3
0	12	3	1.5	1	0.8	0.8
1/3	15	4	2	1.2	1	1
1/2	15	4	2	1.5	1.2	1
2/3	15	5	2.5	1.5	1.2	1
1	20	6	3	2	1.5	1.2
3/2	25	8	4	2.5	2	1.5
2	30	10	5	4	3	2
3	40	20	10	6	6	4
4	60	30	15	12	12	6

$\eta \setminus \xi$	0	1/2	1	3/2	2	3
0	15	3	1.5	1	1	0.8
1/3	15	4	2	1.2	1.2	1
1/2	15	4	2	1.5	1.2	1
2/3	20	5	2.5	1.5	1.5	1
1	25	6	3	2	2	1.2
3/2	30	8	4	3	2.5	1.5
2	40	10	5	4	4	2
3	50	20	10	8	6	4
4	100	40	20	15	12	6

Parameter $d > 0$ for Hyperbolic Balance 40, HB40(d), with Student (left) and Pareto (right) errors.

$\eta \setminus \xi$	0	1/2	1	3/2	2	3
0	15	3	1.5	1.2	1	1
1/3	15	4	2	1.5	1.2	1
1/2	20	4	2	1.5	1.2	1.2
2/3	20	5	2	1.5	1.5	1.5
1	20	6	3	2	2	1.5
3/2	25	8	4	3	2.5	2
2	30	10	5	4	3	2.5
3	40	20	10	8	6	4
4	60	40	20	15	12	8

$\eta \setminus \xi$	0	1/2	1	3/2	2	3
0	10	3	1.5	1.2	1	1
1/3	15	5	2	1.5	1.2	1.2
1/2	20	6	2.5	2	1.5	1.2
2/3	20	6	3	2	1.5	1.5
1	30	10	4	3	2	1.5
3/2	60	20	8	4	3	2
2	100	40	12	6	5	2.5
3	300	150	40	15	10	4
4	1000	500	120	40	20	6

Parameter $d > 0$ for Hyperbolic Balance 50, HB0(d), with Student (left) and Pareto (right) errors.

$\eta \setminus \xi$	0	1/2	1	3/2	2	3
0	12	3	1.5	1	0.8	0.8
1/3	15	4	2	1.2	1	1
1/2	15	4	2	1.5	1.2	1
2/3	15	5	2.5	1.5	1.2	1
1	20	6	3	2	1.5	1.2
3/2	25	8	4	2.5	2	1.5
2	30	10	5	4	3	2
3	40	20	10	6	6	4
4	60	30	15	12	12	6

$\eta \setminus \xi$	0	1/2	1	3/2	2	3
0	15	3	1.5	1	1	0.8
1/3	15	4	2	1.2	1.2	1
1/2	15	4	2	1.5	1.2	1
2/3	20	5	2.5	1.5	1.5	1
1	25	6	3	2	2	1.2
3/2	30	8	4	3	2.5	1.5
2	40	10	5	4	4	2
3	50	20	10	8	6	4
4	100	40	20	15	12	6

Parameter $d > 0$ for LAD with Hyperbolic Correction, LADHC(d), with Student (left) and Pareto (right) errors.

$\eta \setminus \xi$	0	1/2	1	3/2	2	3
0	15	2.5	1	1	1	1
1/3	15	3	1.2	1	1	1
1/2	15	4	1.5	1	1	1
2/3	20	5	2	1.2	1	1
1	25	6	3	2	1.5	1
3/2	25	8	5	4	3	1
2	30	12	8	6	5	4
3	50	30	15	8	6	5
4	80	50	25	12	10	8

$\eta \setminus \xi$	0	1/2	1	3/2	2	3
0	2.5	1	1	1	1	1
1/3	3	1	1	1	1	1
1/2	4	1	1	1	1	1
2/3	5	1.2	1	1	1	1
1	10	2.5	1	1	1	1
3/2	15	8	1.2	1	1	1
2	25	15	2.5	1.2	1.2	1
3	100	60	15	3	2.5	1.2
4	400	150	50	20	12	1.5

Parameter $p > 0$ for Weighted Theil-Sen, WTS(p), with Student (left) and Pareto (right) errors

$\eta \setminus \xi$	0	1/2	1	3/2	2	3
0	(5,0,0)	(6,0,0)	(1,0.04,0.25)	(1,0.4,2.5)	(1,0.25,1.5)	(1,0.3,1.5)
1/3	(12,0.008,0.01)	(12,0.005,0.1)	(5,0.1,2)	(3,0.12,2.5)	(4,0.2,0.8)	(4,0.2,1.2)
1/2	(14,0.008,0.06)	(16,0.02,0.3)	(9,0.12,1.5)	(8,0.15,0.8)	(6,0.25,0.6)	(5,0.3,1.25)
2/3	(16,0.012,0.05)	(19,0.04,0.8)	(12,0.15,1)	(10,0.15,0.8)	(11,0.3,0.6)	(10,0.3,1)
1	(22,0.04,0.08)	(26,0.08,0.3)	(16,0.2,0.5)	(15,0.4,0.4)	(14,0.3,0.6)	(16,0.8,1.5)
3/2	(25,0.1,0.06)	(33,0.25,0.4)	(24,0.5,0.4)	(24,0.8,0.4)	(18,0.6,0.3)	(20,3,2.5)
2	(28,0.15,0.05)	(38,0.5,0.4)	(30,0.8,0.3)	(28,1,0.25)	(26,1,0.3)	(24,3,2.5)
3	(34,0.5,0.015)	(42,1,0.3)	(35,1.2,0.2)	(32,1.5,0.25)	(30,1.5,0.25)	(26,8,5)
4	(39,0.8,0.03)	(45,1.5,0.25)	(38,1.5,0.15)	(34,1.5,0.15)	(32,1.5,0.2)	(28,10,5)

Parameters (m, p, r) for Weighted Least Trimmed Squares, WLTS(m, p, r), with Student errors

$\eta \setminus \xi$	0	1/2	1	3/2	2	3
0	(4,25,0.04)	(5,100,0.06)	(5,80,0.12)	(6,100,0.2)	(4,100,0.15)	(3,80,0.25)
1/3	(7,1.2,0.08)	(9,300,0.15)	(9,8,0.2)	(8,12,0.25)	(8,12,0.25)	(5,50,0.2)
1/2	(9,1,0.1)	(11,6,0.15)	(10,20,0.25)	(9,5,0.25)	(10,40,0.4)	(9,60,0.4)
2/3	(11,1.2,0.12)	(13,80,0.25)	(12,25,0.3)	(10,30,0.4)	(12,30,0.4)	(9,20,0.4)
1	(25,6,0.3)	(17,8,0.3)	(16,100,0.4)	(15,50,0.5)	(15,200,0.4)	(15,100,0.6)
3/2	(34,20,0.4)	(25,12,0.4)	(22,20,0.4)	(23,10,0.6)	(18,150,0.6)	(16,80,0.6)
2	(38,10,0.3)	(33,15,0.5)	(25,10,0.5)	(27,25,0.5)	(25,30,0.6)	(21,15,0.6)
3	(42,8,0.25)	(39,6,0.4)	(32,25,0.6)	(31,12,0.5)	(29,100,0.5)	(25,30,0.6)
4	(43,6,0.2)	(42,6,0.4)	(36,5,0.4)	(33,25,0.5)	(31,12,0.5)	(28,20,0.5)

Parameters (m, p, r) for Weighted Least Trimmed Squares, WLTS(m, p, r), with Pareto errors

6.5 An example

An example may help to explain how the parameter in the estimator functions for samples of size $n \neq 100$. Consider a sample of size 231. The error distribution is not symmetric: the upper tail decreases like c/y, the lower tail like c'/y^2. The explanatory variables have a Pareto distribution with tail index $\xi = 1$. We want to apply the HB100(d) estimator. The estimate is the central line of a strip with a hundred points above the strip and a hundred points below. These two subsets of a hundred points are in balance with respect to the hyperbolic weight $w_i = 1/(d - 1 + i)$. This estimator, like HB40 for sample size $n = 100$, is not very sensitive to the behaviour of the error density at the median.

If the error distribution is known one determines the optimal value of the parameter d by a series of simulations for batches of a hundred thousand simulations. Throughout this paper we choose d to have the form $d_0 * 10^k$ with $d_0 \in \{1, 1.2, 1.5, 2, 2.5, 3, 4, 5, 6, 8\}$. For the optimal value we have computed the empirical sd and bias of $\hat{a}_{\mathrm{HB100}(d)}$ over ten batches of a hundred thousand simulations:

$$d = 3 \qquad \text{emp sd} = 0.00403[2]; \qquad \text{bias} = 0.00021[1]. \tag{6.1}$$

Here is the error density f^*: Start with a unimodal symmetric Pareto distribution with tail $1/(2+2y)$, replace observations $Y_i^* < -1$ by $-\sqrt{|Y_i^*|}$, and scale by the IQD $= 2$. This yields the error density $f^*(y) = 1/(1+2|y|)^2$ on $(-1, \infty)$ and $4|y|/(1+4y^2)^2$ on $(-\infty, -1)$.

If one knows the 231 sample points (X_i, Y_i), but the tail indices and error distribution are unknown, then one has to estimate the tail indices and determine the lack of symmetry in the error distribution. Use the Hill estimator $\hat{\xi}$ for the tail index of the horizontal coordinate. Apply LADPC or one of the other estimators in Table 1 which do not contain a parameter to obtain a preliminary estimate $\hat{a}_0$ of the slope of the regression line. Use the residues $z_i = y_i - \hat{a}_0 x_i$ to determine an estimate $\hat{\eta}$ of the tail index of the error and to determine the lack symmetry of the error distribution. Our program deletes the twenty rightmost points (since for these points the effect of an error in the estimate $\hat{a}_0$ on the value of z_i may be large). The remaining 211 values z_i are shifted to make the median the origin and arranged in increasing order. Now delete the middle points, retaining the 58 largest and the 58 smallest values. Compute the standardized Wilcoxon rank statistic T for these 116 values. For $|T| \leq 2$ we assume that the distribution is symmetric and apply the Hill estimator to the 116 absolute values to determine $\hat{\eta}$; for $T \geq 6$ we assume extreme asymmetry and apply the Hill estimator to the 58 positive values; for $2 < T < 6$ we augment these 58 values with the largest of the remaining absolute values, the number depending linearly on T, to estimate η. For $T < -2$ we do the same with signs changed. Now determine the optimal value d_S of the parameter d for a sample of 231 observations from a Student distribution with tail index $\hat{\eta}$ and d_P for a sample of 231 points from a Pareto distribution with this tail index $\hat{\xi}$. Finally apply the HB100(d) estimator to the sample where d is d_S or d_P or a geometric average of d_S and d_P depending on the value of $|T|$.

What is the performance of this two step estimator? For the initial LADPC estimate ten batches of 10^5 simulations give an empirical sd of 0.00470[2] and bias 0.00017[1]. Construct tables with the optimal value of the parameter d for various values of the tail indices for samples of size $n = 231$ for Student errors and Pareto errors as in Table 3 above:

$\eta \setminus \xi$	0	1/2	1	3/2	2	3
0	30	3	1.5	1	0.8	0.6
1/3	30	4	2	1.5	1	0.6
1/2	30	5	2	1.5	1.2	0.8
2/3	30	5	2	1.5	1.5	1.2
1	30	6	3	2	2	1.5
3/2	40	8	4	3	2	2
2	50	10	5	4	3	2.5
3	60	20	8	6	5	3
4	80	30	15	10	10	5

$\eta \setminus \xi$	0	1/2	1	3/2	2	3
0	30	4	1.5	1.2	1	0.8
1/3	40	6	2.5	1.5	1.2	1
1/2	40	8	3	2	1.5	1.2
2/3	40	8	3	2	1.5	1.2
1	50	8	4	2.5	2	1.5
3/2	60	15	5	3	2.5	2
2	60	20	8	5	3	2
3	120	40	15	10	5	4
4	200	60	25	12	10	8

The optimal parameter d for Hyperbolic Balance, HB100(d).
Sample size $n = 231$ with Student (left) and Pareto (right) errors.

These two tables may be used to determine the optimal value of the parameter d for any pair (ξ, η) in the rectangle $[0, 3] \times [0, 4]$ by interpolation. We calculated for S and P for the six values $\xi = 0, 1/2, 1, 3/2, 2, 3$ a linear approximation to $\log d$ as a function of η and used these linear approximations for the interpolation. Of the million simulations 269 024 were given the predicate "symmetric", and 16 114 "extremely asymmetric". Our estimates of (ξ, η) vary around a mean value $(1.00, 1.3)$ with sd's $(0.066, 0.17)$ and correlation $-2/10^3$. For the log of the parameter d we find a mean value of 0.95 and sd 0.03. The empirical sd and bias for ten batches of 10^5 simulations are

$$\text{emp sd} = 0.00405[2]; \qquad \text{bias} = 0.00023[1].$$

These are indistinguishable from the values in (6.1). Knowledge of the tail indices ξ and η and of the distribution of the error hardly improves the estimate! That is remarkable but perhaps not unreasonable. In the two step estimate the parameter adapts to the configuration of the sample.

7 Conclusions

There are a number of estimators which perform well for linear regression with heavy tails.

The conclusions may be found in the tables in Section 6. For sample size $n = 100$ and errors Y^* with a scaled Student or Pareto distribution and explanatory variables X with a Pareto distribution the tables list the performance of a number of estimators for certain values of the tail index η of the error and of the tail index ξ of the explanatory variable. The similarity in the results for errors with a symmetric Student distribution and for errors with a one sided Pareto distribution suggests that these exemplary results extend to a wide class of error distributions for tail indices $\eta \in [0, 4]$ and $\xi \in [0, 3]$. The example in Section 6.5 shows how these results may be used to obtain estimates of the regression line and the error distribution for samples of any size when the value of the tail indices is not known.

Least Squares performs poorly for errors with infinite second moment, $\eta \geq 1/2$. Least Absolute Deviation too unless $\xi < 1/2$. The other eleven estimators perform well. One may decide to use the Theil-Sen estimator. The performance is good for symmetric errors. It may be improved by adding a weight. Weighted Theil-Sen is overall the best estimator for errors with a Pareto distribution and a good estimator for errors with a Student distribution. If one prefers to work with Least Absolute Deviations the tables will warn you that LAD performs badly if both ξ and η exceed a half. However there exist variations which do well. The Power Corrected LAD overall is the best estimator for Student errors for estimators which do not contain a parameter. The Gap Correction and the Hyperbolic Correction of LAD do better but here one needs to choose a parameter depending on the tail indices. Actually the situation is quite complex. The gaudy Figure 1a shows that for Student errors nine estimators from the dozen which are investigated in the paper are optimal at at least one of the 54 points (ξ, η) at which the loss is calculated.

Statisticians are not particularly concerned with the asymptotic behaviour of estimators for sample size $n \to \infty$. One is confronted with a sample of fixed size. One is interested in the distribution of the estimator for samples of this size. If there is a universal limit like the normal distribution the limit distribution may give information on the distribution of the estimator for the sample at hand. This yields an efficient procedure to evaluate estimators. But if there is no universal limit? It then may be convenient to take a fixed sample size, say $n = 100$, as benchmark. Samorodnitsky et al in [24] show that for heavy tails for $\xi > 1/2$ certain linear estimators of the slope of the regression ray have a limit distribution for $n \to \infty$. The limit is a functional of a Poisson point process on $(0, \infty) \times \mathbb{R}$. It depends on the distribution of the error. It is not universal. This also is the case for non-linear estimators. Unorthodox methods are needed to handle this situation. We have chosen to compare the performance of a small number of estimators for a fixed sample size at six values of the tail index ξ of the explanatory variable and nine values of the tail index η of the error. Throughout the paper we restrict attention to samples of size $n = 100$. The paper uses tables and programs rather than theorems and proofs.

The credit for this paper should not go to the two authors mentioned on the title page, but to R, more precisely to the men and women who developed R. What the telescope is to the human eye R is to statistical intuition.

From our background in extreme value theory it is natural that we should focus the power of R on linear regression with heavy tails. The relation to the theory of conditional extremes and high risk scenarios is explained in Section 1. Our task was to determine the framework: Ten batches of a hundred thousand simulations of a sample of size $n = 100$ and distributions with tail index $\eta = 0, 1/3, 1/2, 2/3, 1, 3/2, 2, 3, 4$ for the error and $\xi = 0, 1/2, 1, 3/2, 2, 3$ for the explanatory variable. The rectangle $[0, 3] \times [0, 4]$ is a small subset of the positive quadrant of the ξ, η-plane but it is ample for most practical applications. The bounds of the rectangle are dictated by the software, R version 3.2.1, the sample size and the number of simulations by the hardware, the operating system OS X 10.6.8 the 3.06 GHz Intel Core 2 Duo processor and 4 GB 1067 MHz DD3 memory of the imac used for obtaining the results of this paper. For most estimators computing the estimate of the regression line for a batch of a hundred thousand samples of size $n = 100$ takes a few minutes. The distributions, Pareto for the explanatory variable and Student for the error, are standard. We added Pareto errors to see how asymmetry would affect the results.

In extreme value theory the step from Pareto samples to a Poisson point process on $(0, \infty)$ with a power tail is a small one, see [8]. The importance of Poisson point processes for linear regression for heavy tails has been pointed out in [24]. The alternative model is a Poisson point process which may be regarded as a sample of size $n = \infty$. Samples of finite size n then correspond to a restriction, one only considers the rightmost n points of the Poisson point process. R allows us to give a detailed picture of the performance of Least Squares for Gaussian errors and of the Right Median estimator for Cauchy errors as a function of ξ for various values of n, see Figure 2 in Section 1 and Figure 10 in Appendix 8.

The first task for R was to show how standard estimators, Least Squares, Least Absolute Deviation and RMP, the bisector through the rightmost sample point, break down: LS for $\eta \geq 1/2$, LAD for $\xi \geq 1/4$ and RMP for $\eta > 1/2$. This analysis is documented in Section 2.

We consider a dozen estimators. The weighted balance estimators are new but seem well suited to handle heavy tails. For weighted balance estimators the estimate of the regression line is a bisector of the sample. One chooses the bisector which ensures balance for a given sequence of weights $w_1, \ldots, w_n$. Weighted balance estimators have several attractive features. They are versatile. They are fast. There exist simple bounds on the tails of the estimators. In some instances not only $\hat{a}_n$ but also $\hat{a}_\infty$ is well defined. Both LAD, Least Absolute Deviation, and RMP, the bisector through the rightmost point, are weighted balance estimators. Simple variations on LAD, like the power correction LADPC and the hyperbolic correction LADHC ensure good performance for $\xi > 1/2$ where LAD breaks down. Similarly RM, the Right Median, yields a more robust estimator of the slope than RMP. The Hyperbolic Balance estimators, HB, use the weight sequence $1/d, 1/(d+1), 1/(d+2), \ldots, 1/(d-1+n)$ which may be regarded as a smoothed version of the weight sequence $1, \ldots, 1, 0, \ldots, 0$ for RM, or as a deterministic version of the weight sequence $w_i = X_i$ used in LAD.

We also look at the performance of the Theil-Sen and the weighted Theil-Sen estimator, and at a number of estimators where the sample is trimmed to avoid the influence of outliers of the error. The reader may restrict attention to the section dealing with the estimator of her choice.

There is no clear winner. In such a situation the reader may expect clear and objective descriptions of the various estimators and of the relations between the estimators. She will be interested in the tail behaviour of the various estimators. The authors may be expected to supply a framework within which the results of the tables may be evaluated.

It is surprising that a criterium which might be supposed to need a finite second moment, a quadratic loss function for the estimates, performs well for errors and explanatory variables with very heavy tails. The RMSE (Root Mean Squared Error) with the term "mean" interpreted as average is the measure used in this paper to compare the performance of the estimators. For symmetric error distributions RMSE is the empirical sd of the batch of a hundred thousand simulations. For Pareto errors we also list the empirical bias. The expected loss may be infinite and yet both the average losses over the ten batches and the fluctuations in these averages may be small. The conclusion in this paper is that the square root of the average quadratic loss is a good measure of performance even for errors and explanatory variables with very heavy tails, and that one should trust the outcomes of R rather than the theoretical expressions for the tails of $\hat{a}$. For risks where the stakes are high one might consider increasing the number of simulations, and perform a hundred batches of ten million simulations rather than ten of a hundred thousand.

What would have happened if we had chosen the absolute value loss function $L(u) = |u|$ to measure performance rather than the quadratic loss function $L(u) = u^2$? By the Mikosch-de Vries Theorem 2.1 the LS estimate $\hat{a}_{\mathrm{LS}}$ has finite expectation for $\eta < 1$. The critical value for the tail index of the error would then be $\eta = 1$ rather than $\eta = 1/2$. Since in applications errors with infinite first absolute moment "do not occur" one might then conclude that there is no need for this paper. However in last instance it is the distribution of the estimator $\hat{a}$ which determines its performance. Since the loglog frequency plots all have a roughly concave shape the loss functions u^2 and $|u|$ both are determined by the right asymptote and will give similar results. For quadratic loss LS is not optimal even for $\eta = 1/3$.

The tables in Section 6 are a poor reflection on the power of R. For each of the 54 points in the ξ, η-plane where the estimator is applied to ten batches of a hundred thousand simulations R not only determines the average loss per batch, it also determines a loglog frequency plot of the absolute estimate $|\hat{a}|$ for the million simulations (and in the case of asymmetric error distributions also of the positive and negative parts of $\hat{a}$). R determines the extreme outcomes of $\hat{a}$ and records the seeds which produced these extreme estimates so that the samples which produce poor estimates can be reconstructed. In addition R gives an estimate of the slope of the right asymptote of the loglog plot based on the thousand largest values of $|\hat{a}|$. This slope is the exponent in the right tail of the df of $|\hat{a}|$. An absolute slope larger than two is an indication of a finite second moment.

Considerations of space forced us to record the empirical sds rather than the loglog frequency plots for the various estimators at the 54 selected points in the ξ, η-plane. For a statistician it is the frequency plot which is of prime interest. By plotting the log of the frequency of $\log |\hat{a}|$ one obtains figures which give a clear representation of the relevant data. The slope of the right asymptote is the exponent of the right tail of the distribution of $|\hat{a}|$, the slope of the left asymptote is the exponent of the df of $|\hat{a}|$ at the origin. The right asymptote tells us how far off the estimates may be, the asymptote on the left tells us how accurate the estimates may be. The curvature at the top reflects the concentration of the distribution of $\log |\hat{a}|$.

Comparison of the shapes of these loglog frequency plots yield a disappointing result. Recall that our estimators are geometric. Transformations $\Gamma \in \mathcal{G}$ have no effect on the estimated regression line. Here $\mathcal{G}$ is the group of all linear transformations of the plane which map right half planes into right half planes and preserve orientation, $\Gamma : (x, y) \mapsto (px + q, b + ax + cy)$ with $p, c > 0$, see (1.2). The loglog frequency plots of the good estimators, up to random fluctuations and a transformation Γ in $\mathcal{G}$, all have the shape of the graph of $\log g_0$ for the probability density $g_0(s) = (1/\pi)/\cosh(s)$. Appendix 10 contains more information on this mysterious result.

The focus on the square root of the average loss does not do justice to the versatility of R in investigating the performance of estimators of the regression line. Let us give two examples.

LAD, and more generally the weighted balance estimators, may be regarded as the counterpart to the median in linear estimation. The estimate is affected by the behaviour of the error density at the median. For errors with a Pareto distribution the bias is a considerable fraction of the empirical sd. A natural question is: Is the positive bias term due to the asymmetry of the tails of the Pareto distribution, or is it due to the local asymmetry of the density at the median? For R the answer is simple. Construct Pareto variables normalized such that $F^*(-1/2) = 1/4$ and $F^*(1/2) = 3/4$ to ensure that IQD $= 1$, draw a hundred samples Y_i^* from this distribution, but flip the sign of the error if Y_i^* lies in the interval $(-1/2, 1/2)$. This reverses the effect of the local asymmetry but does not affect the tails. Simulate ten batches of a hundred thousand samples and note the effect on the bias of estimate $\hat{a}$. We find that roughly a third of the bias is due to the local asymmetry.

For the Cauchy distribution one can easily write down the equations for the MLE estimator of the slope of the regression line. How does this estimator perform? The program includes a call to the optimization function optim() which needs a starting point. If one chooses the origin as initial point the performance is slightly better than HB0 or HB40. However the maximum determined by optim() is local. An initial point far out may result in a different (larger) local maximum. By considering a variety of initial points one increases the chance of finding the absolute maximum. One obtains a better estimate of the MLE. Initial points far out may result in a larger empirical sd. The performance of MLE now is slightly worse than HB.

The criticism that the paper is no more than a collection of examples and a tabulation of rote results obtained by varying the parameters in a small set of simple programs and that it hardly contains any material of theoretical significance is justified. The paper may be seen as a first exploration of a rough terrain. We hope that the tables are helpful to the practicing statistician and that the examples will stimulate theoretical statisticians to develop the mathematics which may result in a better understanding of the results presented above.

8 Appendix 1. Linear Regression for Poisson point processes

Recall the alternative model for the regression equation introduced at the end of Section 1. Instead of a sample of n points (X_i, Y_i) in the plane with $Y_i = Y_i^* + b + aX_i$ we look at the n rightmost points of a Poisson point process N_a with points (X_i, Y_i), $Y_i = Y_i^* + aX_i$. Here $X_i = 1/U_i^\xi$ where $U_1 < U_2 < \dots$ are the points of the standard Poisson point process on $(0, \infty)$. The tail index ξ is positive. The points $X_1 > X_2 > \dots$ then form a Poisson point process on $(0, \infty)$ with mean measure $\rho(x, \infty) = 1/x^\lambda$ for $\lambda = 1/\xi$. The iid sequence (Y_i^*) from the error distribution F^* is independent of the Poisson point process (X_i). If F^* has density f^* then N_a has intensity $f^*(y - ax)\lambda dx/x^{\lambda+1}$ on $(0, \infty) \times \mathbb{R}$. Almost every realization of N_a determines the error distribution as we shall see below. Hence the value of the abscissa b in the regression equation is of little interest. The question is: Do the realizations of N_a determine a?

If the tail index ξ of ρ exceeds a half and the error has a Student or Gaussian distribution then N_a does not determine a. The distributions π_a of the Poisson point processes N_a are *equivalent*. One may write $d\pi_a = f_a d\pi_0$ and $d\pi_0 = g_a d\pi_a$. We shall explain this more precisely below. We then give conditions on the error distribution which ensure equivalence of the distributions π_a, $a \in \mathbb{R}$, for $\xi > 1/2$. Roughly speaking the density f^* of the error Y^* should be positive and smooth. We shall also briefly investigate the influence of irregularities in the error distribution. The second half of this section contains an analysis of the behaviour of the estimate $\hat{a}_n$ of the slope of the regression line based on the n rightmost points of the Poisson point process N_0 for $n \to \infty$ for two estimators: Least Squares and Right Median. For both one may define $\hat{a}_\infty$. For LS

Figure 2 at the end of Section 1 plotted the sd of $\hat{a}_n(\xi)$, $0 \leq \xi \leq 3$, for various sample sizes, $n = 20, 50, 100, 200, 500, 1000$ and for $n = \infty$. Below we shall present similar plots for the empirical sd's of the RM(r) estimates for the optimal value of the parameter r. The figures for LS and for RM(r) both suggest convergence for $n \to \infty$. It will be shown that convergence holds.

8.1 Distributions and densities of Poisson point processes

One can distinguish a biased coin from a fair coin by repeated trials. Similarly one can distinguish a normal variable with variance one and positive mean from a standard normal variable. One can distinguish a Poisson point process on $(0, \infty)$ with intensity $c > 1$ from the standard Poisson point process. But what happens if the bias varies over time, if the mean tends to zero and if the intensity is not constant but a function which tends to one? Suppose the Poisson point process on $(0, \infty)$ has intensity $j(t)$ which tends to one for $t \to \infty$. If $\int_0^t (j(s) - 1)ds = o(\sqrt{t})$ the difference in the number of points on $(0, t]$ between the Poisson point process with intensity j and the standard Poisson point process is masked by the random fluctuations in the standard point process which are of the order of $\sqrt{t}$. Does this imply that one cannot distinguish samples from the two point processes?

To answer this question one has to look at the distributions. If the probability measures which describe the distributions of the point processes are singular one can distinguish samples with certainty; if the distributions are equivalent one can not.

Densities of random variables or vectors are generally taken with respect to Lebesgue measure. One can also consider the density of a variable X with respect to a standard variable U. If X is $N(c, 1)$ and U is $N(0, 1)$ the density of X with respect to U is $f(U) = Z/C$ where $Z = e^{cU}$ and $C = \mathbb{E}Z = e^{c^2/2}$. If (X_n) is a sequence of independent $N(c_n, 1)$ variables with mean $c_n = 1/n$ and (U_n) are independent standard normal variables, the density of the sequence $\mathbf{X}$ with respect to the sequence $\mathbf{U}$ is $f(\mathbf{U}) = Z/C$ where $Z = \exp(U_1 + U_2/2 + \cdots)$ and $C = \exp(1 + 1/4 + \cdots) = e^{\pi^2/6}$. The Monotone Convergence Theorem applied to $Z_n = \exp(U_1/1 + \cdots + U_n/n)$ shows that $\mathbb{E}(Z/C) = 1$. Samples from

(X_n) and from (U_n) cannot be distinguished with certainty. Our aim is to show that this also is the case for samples from the Poisson point processes N_a on $(0,\infty)\times\mathbb{R}$ for $\xi > 1/2$ if the error Y^* has a Student distribution.

The situation is not quite symmetric. If the density of X with respect to U is $f(U)$ then the density of U with respect to X is $g(X)$ where $g = 1/f$, but only if f is U-a.s. positive. Thus if the intensity j of the Poisson point process N_j on $(0,\infty)$ vanishes on the interval $(0,1)$ a sample which has a point in this interval evidently derives from the standard Poisson point process on $(0,\infty)$ and not from N_j. We shall now consider Poisson point processes N and M on a separable metric space with mean measures ν and μ. Assume $d\nu = gd\mu$ with $g = e^\gamma$.

Theorem 8.1. *Let M be a Ppp on a separable metric space E with mean measure μ and N the Ppp with mean measure $gd\mu$ for $g = e^\gamma$. The distribution of N has a density $h(M)$ with respect to the distribution of M in the following situations:*

- *$\mu E < \infty$ and $g \equiv 0$: $h(M) = 1_{\{M=0\}}/e^{\mu E}$;*
- *$\mu E < \infty$, $g = e^\gamma > 0$, $\int gd\mu < \infty$: $h(M) = e^{\int \gamma dM}/e^{\int g-1d\mu}$;*
- *$g = e^\gamma > 0$, $\int \gamma^2 d\mu < \infty$, $\int |g-1-\gamma|d\mu < \infty$: $h(M) = e^{\int \gamma d(M-\mu)}/e^{\int g-1-\gamma d\mu}$.*

Proof If K and K_0 are Poisson variables with expectation c, c_0 the density of K with respect to K_0 is Z/C where $Z = (c/c_0)^{K_0}$ and $C = \mathbb{E}Z = e^{c-c_0}$. Note the similarity with the normal variables. Poisson and normal both are exponential families. Let $g = c_1 1_{E_1} + \cdots + c_m 1_{E_m}$ for disjoint subsets $E_1, \ldots, E_m$ of E and $d\nu = gd\mu$. Set $\gamma_i = \log(c_i)$ and $K_i = M(E_i)$. Then the density of N with respect to M is Z/C where

$$Z = c_1^{K_1} 1_{E_1} + \cdots + c_m^{K_m} 1_{E_m} = e^{\int \gamma dM}$$

and $C = \mathbb{E}Z = e^{\int g-1d\mu}$. The extension to positive μ-integrable Borel functions g is standard, and so is the $\mathbf{L}^2$ extension if $\mu E = \infty$. ¶

The Poisson point process N_a on $(0,\infty)\times\mathbb{R}$ has intensity $\lambda f^*(y-ax)/x^{\lambda+1}$ for $\lambda = 1/\xi$ where f^* is the error density. We are interested in the density Z/C of N_a with respect to

N_0. The restrictions of the intensities to the half plane $\{x \geq 1\}$ are probability densities. If f^* is strictly positive the restrictions of N_a have equivalent distributions by the second condition above. This is not surprising. One can hardly expect to determine with certainty the slope of the regression line from the few points if any of N_a in this half plane. But what if one knows the position of all points?

First observe that for any $\xi > 0$ almost every realization of N_0 determines the error distribution. Recall that for the standard Poisson point process N on $(0, \infty)$ the number $N_t = N(0, t)$ of points of N in the interval $(0, t)$ is almost surely asymptotic to t for $t \to \infty$. Indeed the Law of the Iterated Logarithm applies:

$$\limsup_{t\to\infty} \frac{(N_t - t)}{\sqrt{2\log\log t}} = 1 \qquad \liminf_{t\to\infty} \frac{(N_t - t)}{\sqrt{2\log\log t}} = -1 \qquad \text{a. s.} \tag{8.1}$$

Let $N_0(x)$ denote the number of points of N_0 in the half plane $(x, \infty)\times\mathbb{R}$, and $N_0(x, y)$ the number in $(x, \infty)\times(-\infty, y]$. Assume F^* is continuous in y. Then $N_0(x, y)/N_0(x) \to F^*(y)$ a.s. for $x \to 0+$. This also holds for N_a. If F^* is continuous, the limit relation holds almost surely for all rational y and for all integers a. Hence there is a null set Ω_0 such that

$$N_a(x, y)(\omega)/N_a(x)(\omega) \to F^*(y) \qquad a, y \in \mathbb{R}, \quad \omega \in \Omega_0^c. \tag{8.2}$$

8.2 Equivalence of the distributions π_a for $\xi > 1/2$

For many smooth strictly positive error densities $f^* = e^{-\varphi}$ for $\xi > 1/2$ almost no realization of N_a determines a. The distributions of N_a, $a \in \mathbb{R}$, are equivalent. Since equivalence holds for the restrictions of N_a to $\{x > 1\}$ for positive error densities f^* it suffices to prove equivalence for the restrictions of N_a to the vertical strip $(0, 1) \times \mathbb{R}$. We apply the third criterium of the theorem above with $M = N_0$ and $N = N_a$. Then $g = e^\gamma$ with $\gamma(x, y) = \varphi(y) - \varphi(y - ax)$. Set $\Delta(y) = \varphi(y) - \varphi(y - t)$. If we can prove that $J(0) = \int \Delta^2(y)f^*(y)dy$ and $J(1) = \int |e^\Delta - 1 - \Delta|f^*(y)dy$ are $O(t^2)$ for $t \to 0+$ the two integrals $\int \gamma^2 d\mu$ and $\int |g - 1 - \gamma|d\mu$ are finite for $\xi > 1/2$ and $d\pi_a = hd\pi_0$. By symmetry the distributions π_a of all the point processes N_a are equivalent.

We shall formulate simple criteria on the error density which ensure that the distributions of the point processes N_a are equivalent. The basic condition is that $f^* = e^{-\varphi}$

is strictly positive and continuous and that φ is the integral of a function φ' which is bounded on bounded intervals. The function φ' need not be continuous. If φ' is bounded equivalence holds. If $\varphi'(y)$ tends to ∞ for $y \to \infty$ or to $-\infty$ for $y \to -\infty$ extra conditions are needed. We then assume a second derivative φ'' which is bounded on bounded intervals and which satisfies some extra conditions.

Set $\Delta(y) = \varphi(y) - \varphi(y-t)$ where $f^* = e^{-\varphi}$. We consider the two integrals

$$J(0) = \int \Delta^2(y) f^*(y)dy \qquad J(1) = \int |e^{\Delta} - 1 - \Delta|(y) f^*(y)dy.$$

We want to show that the two integrals are $O(t^2)$ for $t \to 0$. Write $J(i)_a^b$ for the integral over the interval (a,b).

Proposition 8.2. *Suppose φ is the integral of a bounded function φ'. Then J_0 and J_1 are $O(t^2)$ for $t \to 0$ and the distributions π_a, $a \in \mathbb{R}$, are equivalent for $\xi > 1/2$.*

Proof Let $|u| \leq C_0$. There exists a constant C such that $1/C \leq (e^u - 1 - u)/u^2 \leq C$. Hence it suffices to prove that $J(0)$ is $O(t^2)$. This follows since $|\Delta(y)| = |\varphi(y) - \varphi(y-t)| \leq C_1|t|$ where C_1 is a bound for $|\varphi'|$. ¶

The set of densities described in Proposition 8.2 is closed for shifts, scaling, reflection, exponential tilting and powers. If f^* satisfies the conditions then so do $f^*(y-y_0)$, $cf^*(cy)$ for $c > 0$, $f^*(-y)$, $e^{\lambda y} f^*(y)/M(\lambda)$ provided the mgf $M(\lambda) = \mathbb{E}e^{\lambda Y^*}$ is finite at λ, and $(f^*)^q/C(q)$ for $q > 0$ provided the integral $C(q)$ of the power is finite.

If the density f^* is logconcave φ' is increasing. If the limits at $\pm\infty$ are finite the distributions π_a are equivalent. This is the case for Laplace densities $(e^{-x/a} \wedge e^{x/b})/(a+b)$ with $a, b > 0$, and for the EGBP densities. If the derivative of f^* varies regularly at ∞ with exponent $\alpha \leq -1$ then $y\varphi'(y) \to \alpha + 1$ and the distributions of the Poisson point processes N_a are equivalent for $\xi > 1/2$. Student densities satisfy the conditions of Proposition 8.2, and so do the continuous unimodal Pareto densities

$$f^*(y) = \Big(\frac{1_{[0,\infty)}(y)}{(1+y/a)^{\alpha+1}} + \frac{1_{(-\infty,0)}(y)}{(1-y/b)^{\beta+1}}\Big)\Big/\Big(\frac{a}{\alpha} + \frac{b}{\beta}\Big) \qquad a, b, \alpha, \beta > 0. \tag{8.3}$$

For the normal density φ' is not bounded. The situation then is less simple. We assume that φ' is locally bounded. The integrals $J(0)_a^b$ and $J(1)_a^b$ are $O(t^2)$ for bounded intervals

(a, b). We shall introduce extra conditions on the behaviour of φ at $+\infty$ which ensure that the integrals over (b, ∞) are $O(t^2)$ too. Results for the left tail are similar.

First note that $\int e^{\Delta(y)} f^*(y)dy = 1 = \int f^*(y)dy$, and $\int \varphi'(y)f^*(y)dy = 0$. Write

$$J(1) = \int (e^{\Delta} - 1 - \Delta)(y)f^*(y)dy - 2\int_{\Delta<0} (e^{\Delta} - 1 - \Delta)(y)f^*(y)dy.$$

The first integral equals $\int (t\varphi'(y) - \Delta(y))f^*(y)dy$ by the remarks above and the second is bounded by $J(0)/2$ since $e^{-u} - 1 + u \le u^2/2$ on $(0, \infty)$. Hence it suffices to give conditions which ensure that $J(0)$ is $O(t^2)$ for $t \to 0$ and also $J(2) = \int |\Delta(y) - t\varphi'(y)|f^*(y)dy$.

Proposition 8.3. *Suppose φ' is continuous and is the integral of φ''. Assume φ'' is bounded. Assume φ' is bounded on $[0, \infty)$ or $\varphi'(y) \to \infty$ for $y \to \infty$, and similarly for $|\varphi'|$ on $(-\infty, 0]$. Then $J(0)$ and $J(2)$ are $O(t^2)$ for $t \to 0$ and the distributions of the Poisson point processes N_a, $a \in \mathbb{R}$, are equivalent for $\xi > 1/2$.*

Proof Let C be a bound for $|\varphi''|$. Then $|\Delta(y) - t\varphi'(y)| \le (t^2/2)C$. This shows that $J(2)$ is $O(t^2)$. For any $c > 0$ the integral $J(0)$ over the interval $(-c, c)$ is $O(t^2)$. So consider the integral over (c, ∞). If φ' is bounded the proof of the previous proposition applies. So assume $\varphi'(y) \to \infty$ for $y \to \infty$. Observe that $\varphi(y) \to \infty$ and that $\varphi'(y + t)/\varphi'(y) \to 1$ uniformly on bounded t-intervals since φ'' is bounded. Hence $\Delta(y) \sim t\varphi'(y)$ for $y \to \infty$ and $\varphi'(y)/\varphi(y) \to 0$ and also $\varphi'(y)/e^{\varphi(y)/2}$. It follows that $\Delta^2(y) \le 2t^2(\varphi'(y))^2 \le t^2\varphi'(y)e^{\varphi(y)/2}$ for $y > b$, and for b sufficiently large

$$\int_b^{\infty} \Delta^2(y)f^*(y)dy \le t^2 \int_b^{\infty} \varphi'(y)e^{-\varphi(y)/2} = 2t^2e^{-\varphi(b)/2}.$$

A similar argument works for the integral over $(-\infty, -a)$. ¶

Proposition 8.4. *Suppose φ' is continuous and is the integral of a locally bounded function φ''. Assume $\varphi''(y) \to \infty$ for $y \to \infty$ and $\varphi''(y + t)/\varphi''(y) \to 1$ uniformly on bounded t-intervals. There exists a constant b such that the integrals $J(0)_b^{\infty}$ and $J(2)_b^{\infty}$ are $O(t^2)$.*

Proof Observe that $\varphi'(y) \to \infty$ and that $\varphi''(y)/\varphi'(y) \to 0$ and that $\varphi'(y + t)/\varphi'(y) \to 1$ uniformly on bounded t intervals, and similarly $\varphi(y) \to \infty$, $\varphi'(y)/\varphi(y) \to 0$ and $\varphi(y + t)/\varphi(y) \to 1$ uniformly on bounded t-intervals for $y \to \infty$. Hence $J(0)_b^{\infty} = O(t^2)$ by the same argument as above. For $J(2)_b^\infty$ we find $|\Delta - t\varphi'(y)| \leq t^2|\varphi''(y)|$. Now observe that $\varphi''(y) < \varphi'(y)$ implies $\int_b^\infty \varphi''(y) f^*(y)dy \leq f^*(b)$. ¶

The last proposition shows that for errors with positive smooth Weibull densities the Poisson point processes N_a are equivalent. A Weibull density has tail $ce^{-q(y+b)^r}$ for c, q, r positive, or more generally $e^{-R(y)}$ for a function R which varies regularly with exponent $r > 0$. Here we need to assume that R'' varies regularly. The exponents for the left and right tails may differ. The density of the double exponential Gumbel distribution is logconcave but φ' increases too fast for the conditions of the propositions above to apply.

8.3 Error densities with local irregularities

For errors with a Student or Gaussian distribution and $\xi > 1/2$ no realization of N_a determines a. Sometimes local irregularities in the error distribution may help to determine a. For errors with a Pareto distribution and $\xi \leq 1$ almost every realization of N_a determines a. We shall look at the effect of irregularities below.

Lines L with slope a through $(0, y_0)$, where y_0 is a discontinuity of the df F^*, contain infinitely many points of N_a almost surely. With probability one no other line contains more than two points. For discontinuous error distributions N_a determines a almost surely, whatever the value of $\xi > 0$.

Henceforth we again assume a continuous error distribution. There may still be local irregularities in the density which reflect in the Poisson point process N_a. The density may have a zero or become infinite at some point. It may have a jump, a vertex or a cusp. It may vanish on an interval or a half line.

The exponential density and shifted Pareto densities are positive on $[0, \infty)$ and vanish on $(-\infty, 0)$. The points of N_a lie above the ray with slope a. For certain values of the tail index ξ this boundary is sharp. The sector S_0^θ bounded by the horizontal axis and the ray with slope $\theta > 0$ has mean measure

$$\mu_0(S_0^\theta) = \int_0^\infty F^*(\theta x)d\rho_0(x) = \int_0^\infty F^*(\theta x)\lambda dx/x^{\lambda+1} \qquad \lambda = 1/\xi.$$

Since $F^*(x) \sim xf^*(0)$ for $x \to 0+$ and $f^*(0) > 0$ we see that $\mu_0(S_0^\theta) = \infty$ for all $\theta > 0$ for

$\xi \leq 1$.

Proposition 8.5. *Let the error distribution have a finite lower endpoint y_0 and suppose $F^*(y_0+t) \sim ct$ for $t \to 0+$ for a constant $c > 0$. Then N_a determines a almost surely for $\xi \leq 1$: a is the maximal slope for which there are no points of N_a below the line $y_0 + ax$.*

A similar argument shows that N_a determines a almost surely for $\xi \leq 1/\gamma$ if the error has a Gamma distribution with shape parameter $\gamma > 0$.

If F^* has an irregularity at the origin it is the restriction of the point process to sectors S_b^c bounded by two rays with slope $b < c$ which determine whether one can distinguish the point processes N_a. Assume

$$F^*(y) - F^*(-y) \sim c_0 y^\alpha, \qquad \frac{F^*(y) - F^*(0)}{F^*(y) - F^*(-y)} \to p \in [0,1] \qquad y \to 0+. \tag{8.4}$$

For $\xi \leq 1/\alpha$ the mean measure μ_0 of the sector S_{-1}^1 is infinite since $\int_0^1 x^\alpha \lambda dx/x^{\lambda+1} = \infty$ for $\alpha \leq \lambda = 1/\xi$. Let $M(\delta)$ denote the mean measure of the truncated sector $S_{-1}^1 \cap \{x > \delta\}$. For $\xi \leq 1/\alpha$

$$\mu_0(S_b^c \cap \{x > \delta\})/M(\delta) \to H(c) - H(b) \qquad \delta \to 0+$$

where $H(t) = pt^\alpha$ for $t > 0$ and $-(1-p)|t|^\alpha$ for $t < 0$. For μ_a the limit is $H(b+a, c+a)$. The limit relation holds almost surely if one replaces μ_a by N_a. Hence if one can distinguish the functions $t \mapsto H_a(t) = H(t+a)$ one can distinguish the point processes N_a almost surely for $\xi \leq 1/\alpha$. The functions H_a can be distinguished unless $\alpha = 1$ and $p = 1/2$. Then F^* has a positive derivative at the origin:

Proposition 8.6. *Let the df of the error satisfy (8.4). Then N_a determines a almost surely for $\alpha\xi \leq 1$ unless $\alpha = 1$ and $p = 1/2$.*

In particular N_a determines a almost surely for $\xi \leq 1$ if the error density has a jump.

8.4 Plots

The two figures in Figure 10 give a good description of the performance in two specific situations: Least Squares for Gaussian errors in the upper figure and Right Median for Cauchy errors in the lower figure.

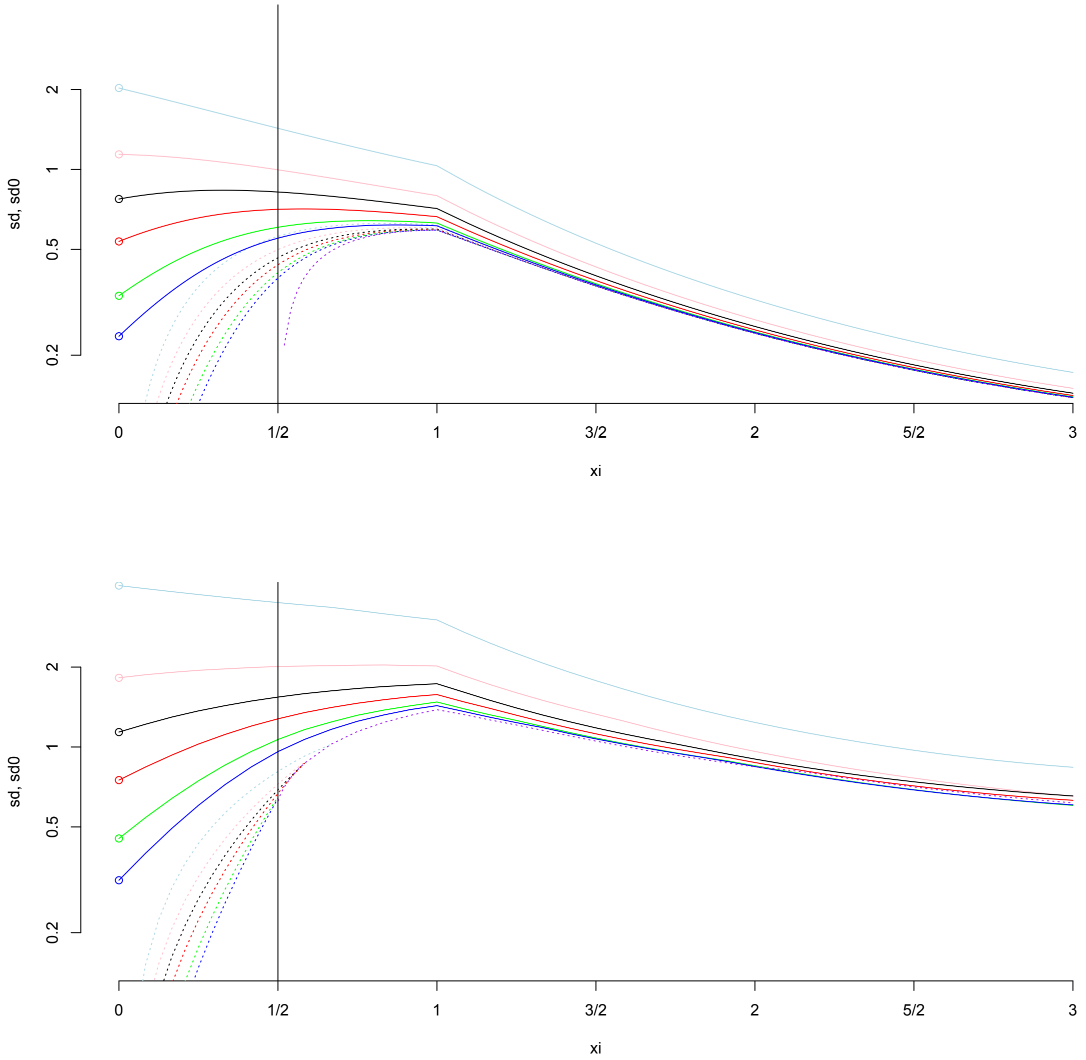


Figure 10: The upper figure shows the sd of the slopes $\hat{a}_n(\xi)$, $\xi \in [0,3]$, (and $\hat{a}_n^0(\xi)$, $\xi \in (0,3]$, dotted) of the LS estimates of the regression line $y = y^* + ax + b$ (and the regression ray $y = y^* + ax$) for the $n = 20, 50, 100, 200, 500, 1000$ (azure, pink, black, red, green, blue) and ∞ (purple dotted) rightmost points of the Poisson point process N_0 for normal errors Y_i^* with IQD $= 1$. The sd's have been scaled by $10^{\xi-1}$. The lower figure shows the scaled empirical sd for the RM-estimates for errors with a Cauchy distribution scaled by its IQD, based on a million simulations. In both figures the explanatory variables have the form $X_i = 1/U_i^\xi$ for $\xi \in [1,3]$ where $U_1 < U_2 < \ldots$ are the points of the standard Poisson point process on $(0,\infty)$. For $\hat{a}_n(\xi)$ $X_i = -\log U_i$ for $\xi = 0$ and $X_i = (1/U_i^\xi - 1)/\xi + \xi$ for $0 < \xi < 1$. For $\hat{a}_n^0$ $X_i = 1/(\xi U_i^\xi)$ for $0 < \xi < 1$. This renormalization is standard in extreme value theory. Because of the geometric nature of the estimators the effect on $\hat{a}_n$, $\hat{a}_n^0$ and on sd_n, sd_n^0 is simple. The values for the sequence $1/U_i^\xi$ are multiplied by ξ to obtain the values plotted in Figure 10. The extra factor ξ on $(0,1)$ explains the kink in the curves at $\xi = 1$. This is the price we pay for continuity of $\hat{a}_n$ and sd_n at $\xi = 0$.

We consider estimates $\hat{a}$ of the slope of the regression line $y = y^* + ax + b$ based on the rightmost points of the Poisson point process N_0 and estimates $\hat{a}^0$ of the slope of the regression ray $y = y^* + ax$. For $\xi \in [1,3]$ the Poisson point process $N_0 = N_0(\xi)$ has intensity $f^*(y)\lambda dx/x^{\lambda+1}$, $\lambda = 1/\xi$. For $\xi \in (0,1)$ the intensity is adapted as described in the caption in order to ensure continuity at $\xi = 0$ for the estimates $\hat{a}$. For $\xi = 0$ the intensity is $f^*(y)e^{-x}$ on $\mathbb{R}^2$.

The sd depends on the error distribution and on the estimator. In the upper figure we plot $\mathrm{sd}_n(\xi)$ and $\mathrm{sd}_n^0(\xi)$ for the Least Squares estimator with Gaussian errors scaled to have IQD $= 1$; in the lower figure the errors have a Cauchy distribution and the Right Median estimator with parameter r is used. Here r is an odd integer, the number of "red" points. It depends on n and ξ and is chosen to yield the minimal average loss over a million simulations. The curves plot the empirical sd, the square root of the average loss over a million simulations.

The similarity between the two figures is striking. For sample size $n = 20, 50, 100, 200, 500, 1000$ the full curves form a decreasing sequence. So do the dotted curves. These lie below the full curves. Knowledge of the abscissa gives a substantial improvement of the estimate. For $\xi \in [1,3]$ the full black curve, sd_{100}, can scarcely be distinguished from the purple dotted curve sd_∞^0. This is not a defect of the estimators LS and RM. For ξ large the points of the explanatory variable tend to zero so fast that the value of (X_i, Y_i) for $i > 100$ is of little use in estimating the slope of the regression line. For ξ closer to $1/2$ we see the same phenomenon in a less extreme form. For $\xi > 1/2$ the asymptotics of sd_n and sd_n^0 for $n \to \infty$ is trivial. Convergence to a positive limit holds without any normalization. That also is the case for the distribution of $\hat{a}_n$ and $\hat{a}_n^0$. Details are given below.

For $\xi > 1/2$ we could have listed in Section 6 the empirical sd for $n = \infty$ rather than $n = 100$. There are practical drawbacks. It is not always clear how the limit variable, the estimate $\hat{a}_\infty$ should be defined. For HB0 and HB40 one can show that truncation of the weight at an appropriate index hardly affects the performance, and one may define $\hat{a}_\infty$ for the truncated weights. For LAD and the variations LADPC and LADHC, or for Theil's weighted estimator or Weighted Least Trimmed Squares, or the estimators TB1, TB2 and TB∞ it is not clear how the empirical sd for sample size $n = \infty$ should be determined. Since the intention of the paper is to introduce and describe a number of estimators which perform well for heavy tails the focus on the finite sample size $n = 100$ is a viable procedure. In principle it makes no difference whether one takes $n = 100$ or $n = \infty$ as the standard.

A closer look reveals several differences between the upper and lower figure. The curves for LS are lower than the corresponding curves for RM. The decrease over the interval $[1, 3]$ is stronger for LS. The dotted lower curves for RM vanish when ξ becomes large. The dotted purple curve for LS seems to decrease to zero for $\xi \to 1/2+0$. This behaviour is less marked for RM. Some of these differences have a simple explanation. LS is optimal for normal errors, the performance of RM is only fair for Cauchy errors. The estimator RM has been chosen not for its good performance but because of its intuitive simplicity and because its behaviour for $n \to \infty$ and for $n = \infty$ can be described by simulations. The estimate $\hat{a}^0_\infty(\xi)$ for RM(r) has a simple form. Draw the rays through the $r = 2r_0 + 1$ rightmost points of N_0. Then $\hat{a}^0_n$ is the median of the slopes of the r rays. If ξ is large r is small since one wants to make optimal use of the leverage effect of the largest values of X in the estimate of the slope. The slow rate of decrease of sd$_n$ and sd^0_n over the interval $1 \le \xi \le 3$ for RM is due to the heavy tails of the Cauchy error. Heavy tails of Y^* demand a conservative estimator. For $\xi > 2$ RM is the median of the slopes of the rays through the $r = 5$ or $r = 7$ rightmost points of N_0. It does not exploit the leverage effect of the rightmost point to the full. The dotted curves for sd^0_n for $n = 20, 50, 100, 200, 500, 1000$ coincide with the purple dotted curve for sd$^0_\infty$ if ξ is so large that the optimal value of the parameter r for $\hat{a}^0_\infty$ is less than twenty.

It should be pointed out that IQD is a good normalization if one wants to compare errors with different tail indices η, but an unnatural normalization for LS. One can construct bounded errors Y^* with IQD $= 1$ and with a symmetric unimodal density for which the sd is very large. A normal error scaled by its IQD has sd 0.74. If one takes a bounded error Y^* with IQD $= 1$ and sd=1.5 the curves in the upper figure will all be shifted upwards over the same distance, corresponding to an increase in the sd's by a factor two.

The plots in the lower figure are random. They depend on the seed with which we start our sequence of a million simulations. This randomness is more pronounced for large values of ξ where r is small. It may account for the anomalous behaviour of the dotted purple curve for $\xi \to 3$.

8.5 Convergence for the LS-estimates

If the explanatory variables have finite second moment the Least Squares estimators $\hat{a}_n^0$ and $\hat{a}_n$ are consistent [5] and asymptotically normal, see [16]. Figure 10 suggest that sd_∞^0 vanishes on $(0, 1/2]$ and is positive on $(1/2, 3]$, and that sd_n^0 and sd_n converge to sd_∞^0 for $n \to \infty$. We shall prove this and show that $\hat{a}_n^0$ and $\hat{a}_n$ converge almost surely to $\hat{a}_\infty^0$ for $\xi > 1/2$. For $\hat{a}_n^0$ convergence in distribution was established in a more general setting in [24].

The sd d_n of the slope $\hat{a}_n$ of the LS estimate of the regression line $y = y^* + b + ax$ is $d_n = \sqrt{\mathbb{E}(1/V_n)}$ and the sd d_n^0 of the slope of the LS regression ray is $d_n^0 = \sqrt{\mathbb{E}(1/Q_n)}$, where

$$Q_n = X_1^2 + \cdots + X_n^2 \qquad V_n = (X_1 - M_n)^2 + \cdots + (X_n - M_n)^2 \tag{8.5}$$

with M_n the mean of $X_1, \ldots, X_n$. Here $X_1 > X_2 > \ldots$ are the points of a Poisson point process on $(0, \infty)$ with mean measure $\rho(x, \infty) = 1/x^\lambda$, $\lambda = 1/\xi$. We prove that d_n and d_n^0 decrease to the same positive limit $d_\infty = d_\infty^0$ for $\xi > 1/2$. The variables $1/V_n$ and $1/Q_n$ converge monotonically and in $\mathbf{L}^1$ to the same finite positive limit $1/V_\infty = 1/Q_\infty$ for $\xi > 1/2$. The variables V_n and Q_n converge monotonically and in $\mathbf{L}^1$ to the finite positive limit $V_\infty = Q_\infty$. The equation $V_n = Q_n - S_n^2/n$ then implies that $S_n/\sqrt{n}$ vanishes in $\mathbf{L}^2$ for $n \to \infty$.

For a finite sequence of real numbers $t_1, \ldots, t_n$ set $v_n = q_n - s_n^2/n$ where

$$q_n = t_1^2 + \cdots + t_n^2 \qquad s_n = t_1 + \cdots + t_n.$$

We do not assume that the t_i are ordered.

Lemma 8.7. *Replacing t_i by $t_i - c$ for $i = 1, \ldots, n$ has no influence on v_n.*

Lemma 8.8. *The sequence v_m, $m = 1, \dots, n$, is increasing.*

Proof Let $m < n$ and set $t_{m+1} = t$. Assume $s_m = 0$, see Lemma 8.7. Then $v_m = q_m = t_1^2 + \cdots + t_m^2 \leq q_m + t^2 - s_{m+1}^2/(m+1) = v_{m+1}$ since $s_{m+1} = t$. ¶

Write $X_i = 1/U_i^\xi$ where $U_1 < U_2 < \cdots$ are the points of a standard Poisson point process on $(0, \infty)$. Then $V_1 = 0$ almost surely and

$$\mathbb{E}(1/Q_n) \leq \mathbb{E}(1/X_1^2) = \mathbb{E}U_1^{2\xi} < \infty \qquad n = 1, 2, \dots, \xi > 0.$$

Lemma 8.9. $\mathbb{E}(1/V_4)$ *is finite.*

Proof First observe that $V_4 \geq X_1^2 + X_4^2 - (X_1 + X_4)^2/2 = (X_1 - X_4)^2/2$ by Lemma 8.7. Write $U_1 = U, U_4 = U + W$ where W is Gamma(3) and independent of the standard exponential variable U. Now observe

$$1/u^\xi - 1/(u+w)^\xi \geq \begin{cases} (1 - 1/2^\xi)/u^\xi & w \geq u \\ \xi w/(2u)^{\xi+1} & 0 < w < u. \end{cases}$$

Hence $\mathbb{E}(1/V_4) \leq \mathbb{E}(2/(X_1 - X_4)^2$ and $1/(X_1 - X_4)^2 \leq U^{2\xi}/(1 - 1/2^\xi)^2$ on $W \geq U$ and $\leq (2U)^{2\xi+2}/\xi^2/W^2$ on $W < U$. Since $\mathbb{P}\{W < t\} \sim t^3/2$ the expectation of $1/W^2$ is finite and so are $\mathbb{E}(1/(X_1 - X_4)^2)$ and $\mathbb{E}(1/V_4)$. ¶

Similar arguments show that $\mathbb{E}(1/V_2)$ is infinite.

The series $Q_\infty = \sum X_i^2$ is almost surely finite for $\xi > 1/2$. It may be expressed as $\int_0^\infty 1/u^{2\xi} dN$ where N is the standard Poisson point process on $(0, \infty)$. Write this as $Q(0,1) + Q(1, \infty)$ where $Q(0,1) = \int_0^1 u^{-2\xi} dN$ is finite as the sum of the squares X_i^2 with $X_i > 1$, and $\mathbb{E}Q(1, \infty) = \int_1^\infty u^{-2\xi} du = 1/(2\xi - 1)$.

Proposition 8.10. *For $\xi > 1/2$ the variable $1/Q_\infty$ is almost surely positive and finite. Its expectation is finite and $1/Q_n \to 1/Q_\infty$ in $\mathbf{L}^1$.*

Proposition 8.11. *For $\xi > 1/2$ the variable $1/V_\infty$ is almost surely finite and positive. Its expectation is finite and $1/V_n \to 1/V_\infty$ in $\mathbf{L}^1$.*

Proof The inequalities $(X_1 - X_4)^2/2 \leq V_4 \leq V_\infty \leq Q_\infty$ prove that V_∞ is finite and positive a. s. Then $V_n \uparrow V_\infty$, together with $\mathbb{E}1/V_4 < \infty$ imply convergence in $\mathbf{L}^1$ by dominated convergence. ¶

Lemma 8.12. *$V_\infty = Q_\infty$ for $\xi > 1/2$.*

Proof We have to prove that $S_n/\sqrt{n} \to 0$ in probability. Set $S_n = A_n + B_n$ where A_n is the sum of the terms $X_i \geq 1$. Then $A_n/\sqrt{n} \to 0$ almost surely. Now B_n may be compared to a stochastic integral. Set $J_t = \int_1^t (1/x^\lambda) dN$, $\lambda = 1/\xi < 2$ where N is the standard Poisson point process on $(0, \infty)$. Then for $\xi \in (1/2, 1)$

$$\mathbb{E}(J_t) = \int_1^t dx/x^\xi = (t^{1-\xi} - 1)/(1-\xi).$$

Hence $J_{2n}/\sqrt{n} \to 0$ in $\mathbf{L}^1$. The nth point U_n has a Gamma(n) distribution. Hence $\mathbb{P}\{U_n > 2n\} \to 0$ and $B_n \leq J_{2n}$ on $\{U_n \leq 2n\}$. It follows that $B_n/\sqrt{n} \to 0$ in probability if $\xi \in (1/2, 1)$. If ξ increases then $1/U_n^\xi$ decreases for $U_n > 1$. Hence $B_n(\xi) \leq B_n(3/4)$ for $\xi > 3/4$ and hence $B_n(\xi)/\sqrt{n} \to 0$ in probability for $\xi > 3/4$. ¶

Corollary 8.13. *Suppose $\xi > 1/2$. Then $S_n/\sqrt{n} \to 0$ in $\mathbf{L}^2$.*

Proof Q_n and $V_n = Q_n - S_n^2/n$ converge in $\mathbf{L}^1$ and the limits agree. ¶

For $\xi = 1/2$ the second moment of $1/U^\xi$ is infinite. The conditions for consistency of the estimators $\hat{a}_n^0$ and $\hat{a}_n$ in [5] do not apply. We shall prove that sd_n^0 and sd_n vanish almost surely for $n \to \infty$. The same arguments work for $\xi < 1/2$.

Proposition 8.14. *Suppose Y^* is centered and has finite variance. Let $X_i = 1/U_i^\xi$ for $\xi = 1/2$. The sd's sd_n^0 and sd_n vanish for $n \to \infty$.*

Proof Introduce the random integrals $S(t) = \int_0^t 1/u^\xi dN$ where N is the standard Poisson point process on $(0, \infty)$ with points $U_1 < U_2 < \ldots$. Similarly we define $Q(t) = \int_0^t 1/u^{2\xi} dN$ and $N(t) = \int_0^t dN$. We shall also consider integrals $S(s,t)$ and $Q(s,t)$ over intervals (s,t). Set $\xi = 1/2$. Note that $\mathbb{E}S(t) = \int_0^t du/u^\xi = 2\sqrt{t}$ and $\mathbb{E}Q(1,t) = \operatorname{var} S(1,t) = \int_1^t du/u^{2\xi} = \log t$. The variance of $Q(1,t)$ is $\int_1^t du/u^2 = 1 - 1/t$. Since $Q(0,1)$ is a finite sum of variables $1/U_i$ we see that $Q(t)/\log t \xrightarrow{\mathbb{P}} 1$. Hence $Q(t) \xrightarrow{\mathbb{P}} \infty$ and by monotonicity $Q(t) \to \infty$ a. s., which implies $Q_n \to \infty$ a. s. and $1/Q_n \to 0$ a. s.. Dominated convergence by Lemma 8.9 implies $\mathbb{E}(1/Q_n) \to 0$. Similarly $S(t)/t \xrightarrow{\mathbb{P}} 2$ implies $S^2(t)/N(t) \xrightarrow{\mathbb{P}} 4$. Set $V(t) = Q(t) - S^2(t)/N(t)$. Then $V(t)/\log t \xrightarrow{\mathbb{P}} 1$ and as above $\mathbb{E}(1/V_n) \to 0$. ¶

We now turn to convergence of the estimators $\hat{a}_n^0$ and $\hat{a}_n$ for $\xi > 1/2$.

Suppose $\xi > 1/2$. The estimate for the slope of the regression ray $y = y^* + ax$ based on the n rightmost points (X_i, Y_i) with $X_i = 1/U_i^\xi$ has the form $\hat{a}_n^0 = \omega_1 Y_1 + \cdots + \omega_n Y_n$ with $\omega_i = X_i/Q_n$ and Q_n as defined in (8.5). We assume that $Y = Y^*$ is centered normal scaled by its IQD. The series $\sum x_n Y_n^*$ converges in $\mathbf{L}^2$ and almost surely if $\sum x_n^2$ is finite. Hence

$$Z_n = X_1 Y_1^* + \cdots + X_n Y_n^* \to Z_\infty = \sum X_n Y_n^* \quad as$$

Then $Q_n \to Q_\infty$ almost surely implies $\hat{a}_n^0 \to \hat{a}_\infty^0 = Z_\infty/Q_\infty$ almost surely, hence in distribution. A more general result on convergence in distribution is given in [24]. The authors observe that the limit distribution may be expressed in terms of the points $U_1 < U_2 < \dots$ of the standard Poisson point process on $(0, \infty)$ as

$$\hat{a}_\infty^0 = (\sum Y_n^*/U_n^\xi)/Q_\infty \qquad Q_\infty = \sum 1/U_n^{2\xi}. \tag{8.6}$$

Almost sure equality holds if one expresses the explanatory variables as $X_i = 1/U_i^\xi$.

The same limit distribution holds for the estimate $\hat{a}_n = \tilde{\omega}_1 Y_1^* + \cdots + \tilde{\omega}_n Y_n^*$ where $\tilde{\omega}_i = \tilde{X}_i/V_n$ and $\tilde{X}_i = X_i - M_n$ for the mean M_n of $X_1, \dots, X_n$. Set $A_n = (\tilde{X}_1 Y_1^* + \cdots + \tilde{X}_n Y_n^*)/Q_n$. Then $Q_n \sim V_n$ a.s. implies $A_n - \hat{a}_n \to 0$ a.s. Now observe that $\hat{a}_n^0 - A_n = M_n(Y_1^* + \cdots + Y_n^*)/Q_n \xrightarrow{\mathbb{P}} 0$ since $\sqrt{n} M_n = S_n/\sqrt{n} \to 0$ in $\mathbf{L}^2$ by Corollary 8.13 and $(Y_1^* + \cdots + Y_n^*)/\sqrt{n}$ converges in distribution to a normal variable. For $\xi > 1/2$:

Theorem 8.15. *Let $U_1 < U_2 < \dots$ denote the points of the standard Poisson point process on $(0, \infty)$ and let (Y_n^*) be an iid sequence of centered variables with finite second moment which is independent of the Poisson point process. Let $\hat{a}_n$ denote the slope of the* LS *estimate of the regression line for the points $(X_1, Y_1^*), \dots, (X_n, Y_n^*)$ where $X_i = 1/U_i^\lambda$, $1/\lambda = \xi > 1/2$. Then*

$$\hat{a}_n \to \hat{a}_\infty^0 = (\sum X_n Y_n^*)/(\sum X_n^2) \qquad \text{a.s.}$$

8.6 Convergence for the RM estimates

The results for the Right Median estimator are slightly different and less complete than for Least Squares.

Recall the limit relation (8.2), $r^{\lambda}N_0\{x > r, y \leq y_0 + ax\} \to F^*(y_0)$, $r \to 0+$. Call a realization of N_0 *unexceptional* if no four points lie on two parallel lines and if the limit relation holds for all y_0 and a. At the end of Section 8.1 we saw that almost every realization of N_0 is unexceptional.

Assume $F^*(0) = 1/2$ and the median is unique. Given a weight, a decreasing sequence $w_i \geq 0$ with finite sum Ω, a *ray of balance* for an unexceptional realization is a ray $L : y = ax$ such that the weight of the points below L does not exceed $\Omega/2$ and neither does the weight of the points above L. Exact balance holds if one of these sets has weight $\Omega/2$. One of the attractive features of the weighted balance estimators is the possibility to define estimators not only for a finite set of rightmost points of the point process N_0 but for the set of all points. The line of balance L_n based on the n rightmost points depends on the weight $\mathbf{w}_n$. For the Right Median estimator for $\xi > 1/2$ simulations suggest that there exists an optimal value $r = 2r_0 + 1$ depending on ξ such that the slope $\hat{a}_{\mathrm{RM}}$ of the ray which divides these rightmost r (red) points fairly has minimal average loss. For the hyperbolic weights $w_i = 1/(d - 1 + i)$ for fixed parameter d and $\xi > 1/2$ there also exists such an optimal truncation. For truncated weights there is a simple continuity result.

Proposition 8.16. *Let $\mathbf{w}_n$ be weights of total weight Ω_n and suppose there exists an index r such that the components w_{ni} vanish for $i > r$. Assume $\mathbf{w}_n \to \mathbf{w}$, where $\mathbf{w}$ has the property that no subset A of $\{1, \dots, r\}$ has weight $w(A) = \Omega/2$ where Ω is the total weight of $\mathbf{w}$. Consider an unexceptional realization $N_0(\omega)$. Let L_n be a line of balance for the rightmost n points of $N_0(\omega)$. Assume the df of the error satisfies $F^*(0) = 1/2$ and the median is unique. Then L_n converges to the line of balance L_0 through the origin for the weight $\mathbf{w}$.*

Proof Let L_0 pass through the point $\mathbf{z}_0$ of the unexceptional realization $N_0(\omega)$. We claim that almost all lines L_n pass through $\mathbf{z}_0$. This implies $L_n \to L_0$ by (8.1). The weights $w_1 \geq \cdots \geq w_r \geq 0$ with total weight $\Omega = 1$ for which there is a subset A in $\{1, \dots, r\}$ of weight $1/2$ lie in a finite union of hyperplanes in $\mathbb{R}^r$. Hence there exists an index n_0 such that for the weights $\mathbf{w}_n$, $n \geq n_0$, no such subset A exists. For these weights the line of balance is unique. Colour the r rightmost points red. There is an interval $J = [-\delta, \delta]$

such that for $y \in J$ the line through $(y, 0)$ and $\mathbf{z}_0$ divides the red points fairly with respect to the eight $\mathbf{w}$. This then also holds with respect to the weights $\mathbf{w}_n$ for $n \geq n_1$. Let L be a line through $(0, y)$ for $y \in J$ and assume $\mathbf{z}_0$ lies below L. Then by monotonicity L does not divide the red points fairly for $\mathbf{w}_n$ for any $n \geq n_1$. So too if $\mathbf{z}_0$ lies above L. ¶

For $\xi < 1/2$ the RM estimate $\hat{a}_n^0$ based on the n rightmost points of the Poisson point process N_a is consistent if the error has a density which is continuous and positive in the median.

Proposition 8.17. *Assume $F^*(0) = 1/2$. Let $\hat{a}_n^0$ denote the median of the slopes of the rays through the rightmost n points of the point process N_a. If Y^* has a density which is positive and continuous at the origin and $\rho_0(x, \infty) = 1/x^{1/\xi}$ with $\xi \in (0, 1/2)$ then $\hat{a}_n \to a$ almost surely.*

Proof We may assume that $a = 0$. Set $\lambda = 1/\xi$. If there exists $c_0 > 1/\sqrt{2}$ and $\delta_0 \in (0, 1)$ such that the truncated sector $S_0^\theta(\delta) = \{x > \delta, 0 < y < \theta x\}$ satisfies

$$\mu_0(S_0^\theta(\delta)) > c_0 \delta^{-\lambda/2} \sqrt{\log\log(1/\delta)} \qquad \delta \in (0, \delta_0) \tag{8.7}$$

then by the Law of the Iterated Logarithm for almost every realization of N_0 the number of points below the ray through $(1, \theta)$ will eventually exceed the number of points above the ray, and hence $\hat{a}(\delta) < \theta$ where $\hat{a}(\delta)$ is the median of the slopes over the points in $(\delta, \infty) \times \mathbb{R}$. Now observe that $\mu(S_0^\theta(\delta)) = \int_\delta^\infty F^*(\theta x) - F^*(0) \lambda dx / x^{\lambda+1}$. Since $F^*(\theta x) - F^*(0) \sim f^*(0)\theta x$ for $x \to 0+$ we find

$$\mu(S_0^\theta(\delta)) \sim f^*(0)\theta\lambda/((\lambda - 1)\delta^{\lambda-1} \qquad \delta \to 0+ .$$

If $\xi < 1/2$ then $\lambda - 1 > \lambda/2$ and (8.7) holds. ¶

If n is odd the ray with the median slope will pass through a point (X_{K_n}, Y_{K_n}) and $K_n \to \infty$ almost surely since Y_i^* is non-zero for $i = 1, 2, \ldots$. The leverage effect of the horizontal coordinate decreases as K_n increases. Convergence is slow as may be seen in Figure 10.

Remark 1. If Y^* has a symmetric Pareto distribution with density $f^*(x) = 1/(2x^{1/\eta})$ on the complement of $(-1, 1)$ the df of the RM estimate $\hat{a}_n$ will converge to the defective df $G \equiv 1/2$. ◊

Proposition 8.18. *Suppose $\hat{a}_n^0 \to \hat{a}_\infty^0$ almost surely. Then the empirical sd's for the estimates $\hat{a}_n^0$ based on a million simulations converge almost surely to the empirical sd's of the estimates $\hat{a}_\infty^0$.*

Proof The average of a finite number of variables which converge almost surely converges almost surely to the average of the limit variables. ¶

Corollary 8.19. *For $\xi > 1/2$ the empirical sd's for the* RM *estimates converge:* $\mathrm{sd}_n^0 \to \mathrm{sd}_\infty^0$. *Similarly* $\mathrm{sd}_n \to \mathrm{sd}_\infty^0$ *almost surely. For* $0 < \xi < 1/2$, $\mathrm{sd}_n^0 \to 0$ *almost surely.*

9 Appendix 2. The left tail of D_n

With a sample of n points associate the uniform distribution on these n points and the mean m and sd d associated with this distribution. With the sample $X_1, \dots, X_n$ associate the variables S, Q, V, M, D:

$$S_n = X_1 + \cdots + X_n \qquad Q_n = X_1^2 + \cdots + X_n^2 \qquad V_n = Q_n - S_n^2/n \tag{9.1}$$

and

$$M_n = S_n/n \qquad D_n^2 = V_n/n = ((X_1 - M_n)^2 + \cdots + (X_n - M_n)^2)/n. \tag{9.2}$$

Proposition 9.1. *Let $M = (X_1, \dots, X_n)/n$ be the average of a sample of size $n > 1$ from a df F with a bounded density and let $D > 0$ be the square root of*

$$D^2 = \Big((X_1 - M)^2 + \cdots + (X_n - M)^2\Big)/n.$$

There exists a constant $A = A_n$ such that

$$\mathbb{P}\{D \le s\} < As^m \qquad s > 0 \qquad m = [n/2].$$

Proof The inequality $D^2 < s^2$ implies that the sample clusters around the average M: more than m points X_i satisfy $|X_i - M| \le \sqrt{2}d$. (If $(X_i - M)^2 > 2s^2$ holds for $n-m$ indices then $D^2 > 2(n-m)s^2/n \ge s^2$.) Clustering implies that the interval $(M - \sqrt{2}s, M + \sqrt{2}s)$

contains more than m points. In terms of the order statistics $X_{(1)} < \cdots < X_{(n)}$: There exists an index $i = 1, \ldots, n-m$ such that

$$X_{(i+m)} - X_{(i)} < 2\sqrt{2}s. \tag{9.3}$$

We shall first derive bounds on the events $E_i(r) = \{U_{i+m} - U_i < r\}$ for uniform order statistics U_i. By symmetry $\mathbb{P}E_{n+1-m-i}(r) = \mathbb{P}E_i(r)$. The order statistic U_k has density $f_{k,n}$ and

$$f_{k,n}(u) = \binom{n-1}{k-1} u^{k-1}(1-u)^{n-k} \quad \Rightarrow \quad \mathbb{P}\{U_k \leq u\} \leq c_{k,n}u^k \qquad c_{k,n} = \frac{1}{k}\binom{n-1}{k-1}.$$

Conditional on $U_i = u$ the difference $U_{i+m} - u$ is distributed like $(1-u)V$ where V is the mth order statistic from a sample of size $n-i$ from the uniform distribution on $(0,1)$. Hence

$$\mathbb{P}(U_{i+m} - U_i \leq r \mid U_i = u) = \mathbb{P}\{V \leq r/(1-u)\} \leq c_{m,n-i}(r/(1-u))^m.$$

We may restrict attention to $i \leq [([n/2]+1)/2]+1$ by symmetry. Then $n-i \geq m$ and $\mathbb{P}E_i(r) \leq c_{m,n-i}r^m\mathbb{E}(1/(1-U_i)^m)$ and

$$\mathbb{E}(1-U_i)^{-m} = \int_0^1 f_{i,n}(u)/(1-u)^m du = C_i = \binom{n-1}{i-1} \Big/ \binom{n-m-1}{i-1}.$$

Finally write $X = F^{\leftarrow}(U)$. Then $X_{(i)} = F^{\leftarrow}(U_i)$ and

$$X_{(i+m)} - X_{(i)} = \int_{U_i}^{U_{i+m}} dF^{\leftarrow}(u) \geq (U_{i+m} - U_i)/\|f\|_\infty$$

where $\|f\|_\infty$ is the bound on the density f of X. This yields the desired inequality with $A_n = (\sqrt{8}\|f\|_\infty)^m C$, where C is a sum of products $c_{i,m}C_i$. ¶

10 Appendix 3. The EGBP distributions

EGBP is the acronym of Exponential Generalized Beta Prime. The EGBP densities form a four dimensional family of logconcave functions $f = e^{-\psi}$ where ψ is a smooth function with asymptotes which have finite non-zero slope. We are interested in the class EGBP since the characteristic shape of the loglog frequency plots of $|\hat{a}_n(E)|$ for many estimators

E is a concave function with non-zero asymptotic slopes. The fit with a function of the form $y_0 - \psi$ is good.

There is a relation with heavy tailed distributions on $(0, \infty)$ such as the Gamma distributions and the Snedecor F-distributions and with symmetric heavy-tailed distributions such as the Student distributions. The variable $S = aS_0 + b$ has an EGBP distribution precisely if $X_0 = \exp(S_0)$ has a Generalized Beta Prime distribution. A Beta Prime distribution is a Beta distribution transformed to live on the positive half line rather than the interval $(0, 1)$. If U has a Beta distribution on $(0, 1)$ with parameters a, b then $X = U/(1 - U)$ lives on the positive half line. It has a strictly positive density and its df tends to ∞ like $1/x^a$ and to zero like x^b. X has a Beta Prime distribution.

We shall first give a description of the class EGBP, then show the relation with heavy tailed distributions, and finally discuss the remarkable fit to the loglog frequency plots of the estimators $\hat{a}$ considered in this paper.

10.1 The EGBP densities

This section contains information about EGBP, the set of Exponential Generalized Beta Prime distributions, their densities, the role of exponential tilting and powers.

We begin with a simple result on logconcave densities $f = e^{-\psi}$. Let ψ' denote the derivative of ψ. It is increasing since ψ is convex. We may assume that it is right-continuous. The derivative ψ' determines the logconcave density f. If ψ_0 is convex with derivative ψ' one may write $f = C_0 e^{-\psi_0}$ where C_0 is the constant which ensures that $\int f(s)ds = 1$. This results in a one to one correspondence between the class LCA of logconcave densities with non-zero finite asymptotes on the one hand and a product space on the other

$$\text{LCA} \quad \leftrightarrow \quad (0, 1) \times (0, \infty) \times \text{DF}$$

where DF is the space of all dfs on $\mathbb{R}$. Write $\psi' = (H - p)c$ with $p \in (0, 1)$ and $c > 0$ for some df H. Every non-degenerate df H generates a four parameter family of log concave densities $f = e^{-\psi}$ where

$$\psi'(x) = c(H(ax+b) - p) \qquad p \in (0,1), a, c > 0, b \in \mathbb{R}. \tag{10.1}$$

These families are disjoint unless H_1 and H_2 are of the same type, in which case the families coincide. The degenerate distribution H concentrated at $b \in \mathbb{R}$ generates the three parameter family of shifted Laplace densities $f(x) = f_0(x+b)$ where

$$f_0(x) = (e^{x/\alpha} \wedge e^{-x/\beta})/(\alpha+\beta) \qquad \alpha, \beta > 0. \tag{10.2}$$

An increasing function ψ_0' which assumes both positive and negative values has a unique primitive ψ_0 such that $f_0 = e^{-\psi_0}$ is a probability density. The probability density associated with $\psi_0'(ax+b)$ is $af_0(ax+b)$. The probability densities associated with $c\psi_0'(x)$, $c > 0$, are powers $f = C_c f_0^c$. The probability densities associated with $\psi_0' + d$ for $d \in (\psi'(-\infty), \psi'(\infty))$ form the exponential family generated by f_0. EGBP is the four dimensional set of distributions with logconcave densities generated by the logistic df $H_0(x) = 1/(1+e^{-x})$ as we shall see presently. Here we want to stress that in terms of the derivative ψ' the four parameters consist of two parameters which determine an affine transformation on the vertical axis and two parameters which determine an affine transformation on the horizontal axis. This holds for any df H by (10.1). To describe the four dimensional class of functions ψ associated with the df H_0 we may use the group $\mathcal{G}$ of transformations $(x', y') = \gamma(x, y) = (px+q, ax+b+cy)$, see (1.2).

Example 4. Note that the density $f_0(s) = c/\cosh(s)$ is logconcave, $\psi_0(s) = c_0 + \log\cosh(s)$ and

$$\psi_0'(s) = \tanh(s) = (e^s - e^{-s})/(e^{-s} + e^s) = 2(H_0(2s) - 1/2).$$

The variable S with density $c/\cosh(s)$ is EGBP. The variable $X = e^S$ has density $2c/(x + 1/x)/x = 2c/(1+x^2)$. We see that $c = 1/\pi$ and X is the absolute value of a standard Cauchy variable. The question which interests us is whether the loglog frequency plot of $\hat{a}_n$ for a given estimator E can be approximated well by the graph of $y_0 - q(\psi_0(t) + pt)$, $t = as + b$ for appropriate constants a, b, p, q and y_0. ♢

The density $f_0(s) = (1/\pi)/\cosh(s) = e^{\log\cosh(s)}/\pi = e^{-\psi_0(s)}/\pi$ in the example above generates the class EGBP. Every density f in EGBP has the form $f = ce^{-\psi}$ where $\psi(s) = q(\psi_0(as+b)+p(as+b))$ with $q > 0$ and $p \in (-1,1)$ to ensure that ψ' is positive at ∞ and negative at $-\infty$. We shall now first look at the distributions of the heavy tailed positive variables $X = e^S$ associated with the variable S with an EGBP density.

10.2 Basic formulas

The Beta Prime distribution with parameters (a,b) has density

$$g(x) = \frac{1}{B(a,b)}\frac{x^a}{(1+x)^c}\frac{1}{x} \qquad B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(c)} \qquad a > 0, b > 0, c = a+b. \tag{10.3}$$

The distribution has power tails with exponent a at zero and $-b$ at infinity. If X has a Beta Prime distribution on $(0,\infty)$ then $X/(X+1)$ has a Beta distribution on $(0,1)$ with the same parameters. The variable X may also be written as the quotient of two independent Gamma variables $X = X(a)/X(b)$ where $X(\lambda)$ has density $x^{\lambda-1}e^{-x}/\Gamma(\lambda)$. The variable $X(a)$ yields the power tail of X at zero, the heavy tailed variable $1/X(b)$ the power tail at infinity.

The variable $T = \log X$ has density

$$\frac{1}{B(a,b)}\frac{e^{at}}{(1+e^t)^c}.$$

The moment generating function of T has a simple form:

$$\int \frac{e^{\xi t}e^{at}}{(1+e^t)^c}dt = \frac{\Gamma(a+\xi)\Gamma(b-\xi)}{\Gamma(c)} \qquad \Rightarrow \qquad \mathbb{E}e^{\xi T} = \frac{\Gamma(a+\xi)}{\Gamma(a)}\frac{\Gamma(b-\xi)}{\Gamma(b)} \qquad a+\xi > 0, b-\xi > 0. \tag{10.4}$$

We shall compute the density and mgf of the normalized variable

$$S = \log((X(a)/a)/(X(b)/b)) = T - t_0 \qquad e^{t_0} = a/b.$$

Proposition 10.1. *The normalized variable S above has density $f(s)$ and mgf $M(\xi)$ given by*

$$f(s) = Ce^{-r\psi_p(s)} \qquad C = \frac{\Delta(c)}{\Delta(a)\Delta(b)} \qquad \Delta(x) = e^x\Gamma(x)/x^x \qquad p = a/c, q = 1-p = b/c, r = ab/c$$

$$M(\xi) = \mathbb{E}e^{\xi S} = \Big(\frac{b}{a}\Big)^{\xi}\mathbb{E}e^{\xi T} = \frac{\Gamma(a+\xi)}{a^{\xi}\Gamma(a)}\frac{\Gamma(b-\xi)}{\Gamma(b)/b^{\xi}} \qquad -a<\xi<b \qquad a=r/q, b=r/p, c=a+b.$$

The functions

$$\psi_p(s) = (\log A(s))/pq \qquad A(s) = pe^{qs} + qe^{-ps} \qquad p+q=1 \tag{10.5}$$

with increasing derivative

$$\psi_p'(s) = \frac{1}{p+1/(e^s-1)} \tag{10.6}$$

are standardized (see Figure 11):

$$\psi_p(0) = \psi_p'(0) = 0 \qquad \psi_p''(0) = 1 \qquad \psi_p'(-\infty) = -1/q \qquad \psi_p'(\infty) = 1/p. \tag{10.7}$$

Proof Set $s = t - t_0$. Then

$$\int \frac{e^{at}}{(1+e^t)^c}dt = \Big(\frac{p}{q}\Big)\int\frac{e^{as}}{(1+pe^s/q)^c}ds = \frac{a^ab^b}{c^c}\int\frac{ds}{(pe^{qs}+qe^{-ps})^c}$$

and hence

$$\int e^{-r\psi_p(s)}ds = \int\frac{ds}{(pe^{qs}+qe^{-ps})^c} = \frac{c^c}{a^ab^b}\frac{\Gamma(a)\Gamma(b)}{\Gamma(c)} = \frac{\Delta(a)\Delta(b)}{\Delta(c)}$$

with $c = r/pq$. ¶

The functions ψ_p satisfy the symmetry relation:

$$\psi_q(s) = \psi_p(-s) \qquad q = 1-p.$$

Let h_p, $p \in (0,1)$, be one of the families:

$$h_p(t) = 1/(pe^{qt} + qe^{-pt});$$

$$h_p(t) = e^{\theta t}/\cosh(t) \qquad \theta = 2p-1;$$

$$h_p(t) = e^{pt}/(1+e^t).$$

Theorem 10.2. *Let $\psi_p = -\log h_p$ for one of the three families h_p, $p \in (0,1)$, above. The distributions with logconcave densities of the form $f = e^{-\psi}$ where*

$$\psi(t) = d + c\psi_p(at+b) \qquad p\in(0,1), a,c>0, b\in\mathbb{R}, d = -\log\Big(a\int e^{-c\psi_p(s)}ds\Big)$$

are the EGBP*-distributions.*

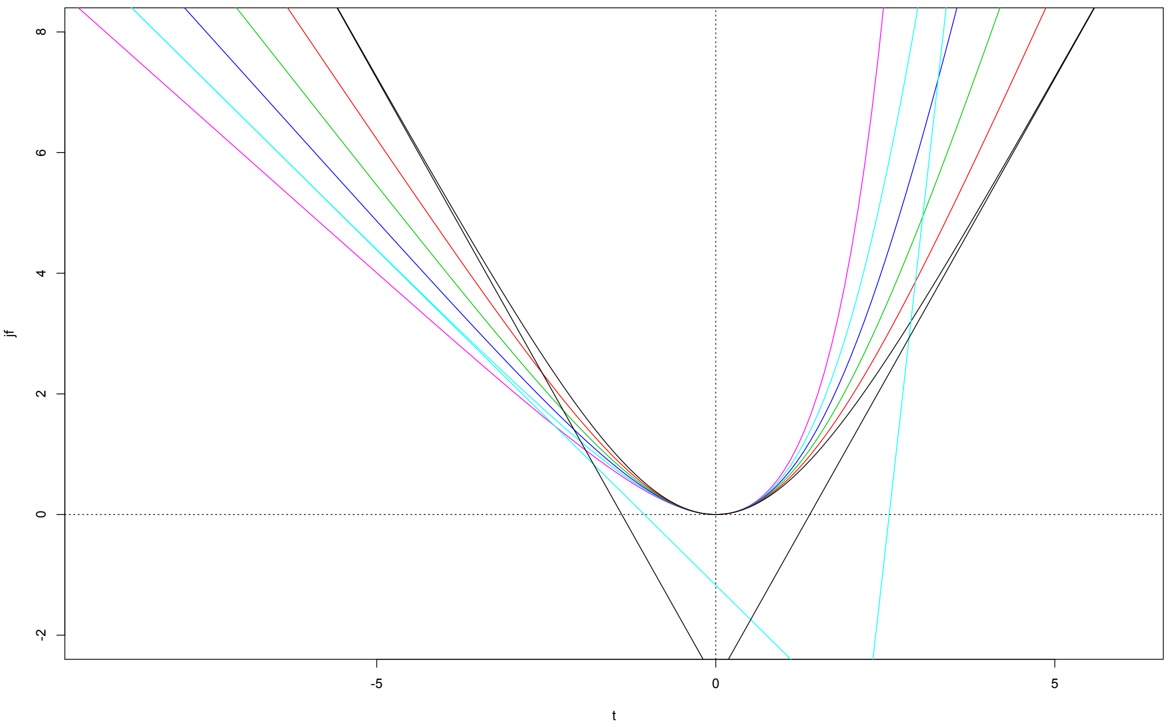


Figure 11: The convex functions ψ_p for $p = 0.5, 0.4, 0.3, 0.2, 0.1, 0$ (black, red, green, blue, azure, purple) and the asymptotes for $p = 1/2$ and $p = 0.1$.

Corollary 10.3. EGBP *is the set of dfs with logconcave densities of the form* $ce^{-\psi}$ *where*

$$\psi'(t) = a_0 H_0(c_0 t + c_1) + a_1 \qquad |a_1| < a_0, c_0 > 0,$$

for the logistic df $H_0(t) = 1/(1 + e^{-t})$.

10.3 The closure of EGBP

The EGBP distributions form a four dimensional set in the space of all non-degenerate dfs on $\mathbb{R}$ with the topology of weak convergence. The closure of this set contains the normal distributions, the exponential and the Laplace distributions, but also the Gumbel distribution for maxima and the corresponding limit distribution for minima. We shall show that there is a closed triangle of dfs F_{θ_1,θ_2} such that EGBP distributions are the dfs $F_{\theta_1,\theta_2}(ax + b)$ where (θ_1, θ_2) are interior points of the triangle. We shall first consider the effect of scaling, both in the horizontal and in the vertical direction, on the convex functions ψ_p.

- Zoom out. The transforms $c\psi_p(s/c)$ have the same asymptotic slopes as ψ_p. For $c \to \infty$ the functions $c\psi_p(s/c)$ converge to the wedge $-s/q \vee s/p$ corresponding to the Laplace density $e^{s/q} \wedge e^{-s/p}$.
- Zoom in. The transforms $c^2\psi_p(x/c)$ have the same curvature as ψ_p at the origin. For $c \to 0$ we obtain the parabola $y = s^2/2$ corresponding to the standard normal density.

Loosely speaking the EGBP distributions form a bridge between the normal distributions and the (shifted asymmetric) Laplace distributions.

The density $\psi_p'(s) = 1/(p + 1/(e^s - 1))$ tends to $e^s - 1$ for $p \to 0$. The limits ψ_1 and ψ_0 exist. They correspond to the Gumbel distribution and the corresponding limit law for minima. The limit relations

- $\psi_p(t) \to \psi_0(t) := e^t - 1 + t$ for $p \to 0$,
- $c_n\psi_{p_n}(t/c_n) \to \varphi_p(t) = t/p \vee (-t/q)$ for $c_n \to 0$, $p_n \to p \in (0, 1)$,

- $c_n^2\psi_{p_n}(t/c_n) \to \varphi(t) = t^2/2$ for $c_n \to \infty$, $p_n \in (0,1)$,

follow from the corresponding limit relations for the derivatives.

Theorem 10.4. *The closure of this four-dimensional set of* EGBP-*dfs in the space of non-degenerate dfs is the set of dfs*

$$F(t) = F_{\theta_1,\theta_2}(at+b) \qquad (\theta_1,\theta_2) \in \Theta, a > 0, b \in \mathbb{R}$$

where Θ *denotes the closed triangle with vertices* $(0,0)$ *and* $(\pm 1, 1)$ *and* F_{θ_1,θ_2} *denotes the df with density* $f = C(r,p)e^{-\psi_{r,p}}$ *for* $\theta_2 = e^{-r}$ *and* $\theta_1 = (2p-1)/e^r$.

Proof Let $\psi_{r,p}(t) = (r \vee r^2)\psi_p(t/r)$ for $r > 0$ and $p \in [0,1]$. If $(r_n, p_n) \to (0,0)$ and $r_n p_n > 0$ then

$$\psi'_{r_n,p_n}(t) = \psi'_{p_n}(t/r_n) = \frac{e^{t/c_n} - 1}{p_n e^{t/c_n} + q_n} \to \begin{cases} \infty & t > 0 \\ -1 & t < 0. \end{cases}$$

The corresponding dfs converge to the standard exponential df. This also holds if $r_n = 0$ or $p_n = 0$. Similarly for $r_n \to \infty$ the functions $\psi_n = \psi_{r_n,0}$ satisfy

$$\psi'_n(t) = r_n\psi'_0(t/r_n) = r_n(e^{t/r_n} - 1) \to t.$$

We conclude that the closure of the set of functions $\psi_{r,p}$ and of the corresponding dfs $F_{r,p}$ is a triangle.

Suppose Z_n has df $F_n = F_{\theta(n)}(a_n z + b_n)$ and $Z_n \Rightarrow Z$. Then $X_n = a_n Z_n + b_n$ has df $F_{\theta(n)}(x)$. Since Θ is compact there is a subsequence $\theta(k_n) \to \theta(0)$ and $X_n \Rightarrow X_0$. By the Convergence of Types Theorem $a_{k_n} \to a_0 > 0$ and $b_{k_n} \to b_0$ and

$$Z_{k_n} = (X_{k_n} - b_{k_n})/a_{k_n} \Rightarrow (X - b_0)/a_0 = Z.$$

Hence the limit of F_n has the form $F_{\theta(0)}(a_0 t + b_0)$. ¶

10.4 SGBP

The estimator $\hat{a}$ is symmetric if Y^* is. The distribution of $\log|\hat{a}|$ may be approximated by an EGBP distribution; the distribution of $\hat{a}$ by the corresponding Symmetric Generalized

Beta Prime distribution. If X has a Symmetric Beta Prime distribution then $T = \log|X|$ has density $Ce^{at}/(1+e^t)^c$ as we saw above. The symmetric variable $\tilde{X}$ corresponding to $\tilde{T} = rT + d$ is $\tilde{X} = e^d|X|^r \operatorname{sign}(X)$, a power transform of X. Thus the SGBP distributions have three shape parameters, the EGBP densities $f = e^{-\psi}$ two, the exponents ψ one, and their derivative ψ' none.

The class EGBP is closed under certain operations. Variables may be scaled, translated and their sign may be changed; densities are closed for powers and exponential tilting. For the class SGBP of Symmetric Generalized Beta Primes distributions there exist related results. Let c be positive.

- if X is a SGBP variable then so are cX, $|X|^c \operatorname{sign}(X)$ and $1/X$;
- if f is a SGBP density and $J = \int f^c(x)dx$ is finite then f^c/J is a SGBP density;
- if f is a SGBP density and $J = \int |x|^c f(x)dx$ is finite then $|x|^c f(x)/J$ is a SGBP density.

Proposition 10.5. *SGBP is the smallest set of dfs which satisfies the three closure properties above for $c > 0$ and which contains the Cauchy distribution.*

Proposition 10.6. *The set SGBP is the smallest set of distributions which contains the symmetric Student t distributions and satisfies for $c > 0$:*

- *if X is a SGBP variable then so is $|X|^c \operatorname{sign}(X)$;*
- *if f is a SGBP density and $J = \int |x|^c f(x)dx$ is finite then $|x|^c f(x)/J$ is a SGBP density.*

Multiplication by a power of x for the density of a positive variable X corresponds to exponential tilting for the densitiy of $\log X$. A good example is the family of Gamma densities. Powers of densities do not have a simple probabilistic interpretation. Families of densities which are closed for powers include the symmetric normal densities, the symmetric and the asymmetric Laplace densities and the symmetric Student t densities.

The closure of the set of Generalized Beta Prime distributions (or Beta distributions of the second kind) is rich. It contains the Beta Prime distributions, Student t distributions, the F-distributions, the gamma distributions, the Weibull distributions with densities $Cx^{a-1}e^{-bx^c}$ for a, b, c positive, the lognormal, log-Laplace and loglogistic distributions. The notes above exhibit the simple underlying structure of this set of distributions on $(0, \infty)$.

10.5 The parameters

The set EGBP has four parameters. The mode x_0 is unambiguous. For the functions ψ with the top at the origin there are different parametric descriptions:

1) geometric: (L, R, D). The absolute values L and R of the inverse of the left and right asymptotic slope and the absolute value D of the curvature $\psi''(0)$. The shape parameter is $p = R/M$ where $M = L + R$.

2) algebraic: (p, u_0, v_0). One may write $\psi(x) = \psi_p(u_0 x)/v0$. Hence $D = \psi''(0) = u_0^2/v_0$ and $M = v0/u0$.

3) stochastic: (a, b, s_0). Let X have a Prime Beta distribution with parameters (a, b) and set $S = \log(X/(a/b))$. The rv $Z = S/s_0$ has density $e^{-\psi}$ with $\psi(z) = y_0 + r_0\psi_p(s_0 z)$ with $r_0 = ab/c$.

Set $L + R = M$, $a + b = c$, $p + q = 1$ and $r_0 = ab/c$, $c = r_0/pq$. Then

$$p = \frac{a}{c} = \frac{R}{M}, u_0 = s_0 = DM, v_0 = \frac{1}{r_0} = \frac{c}{ab} = DM^2; \qquad a = \frac{r_0}{q} = \frac{1}{v_0 q}, b = \frac{1}{v_0 p}.$$

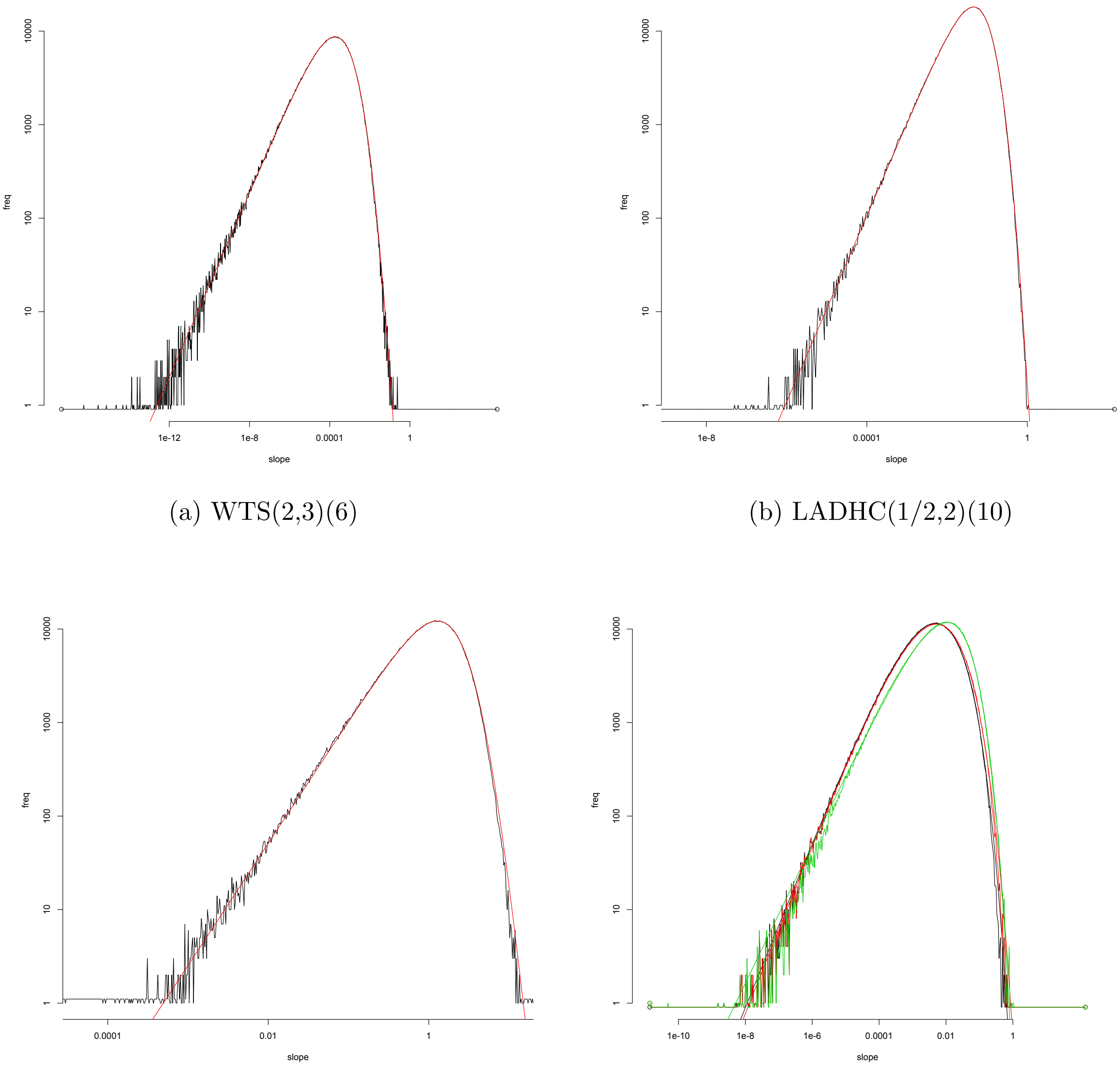
10.6 Fitting EGBP distributions to $\log|\hat{a}_n(E)|$

Ten batches of a hundred thousand simulations yield a useful estimate of the empirical sd. The million simulations also yield good frequency plots. We restrict attention to errors with a symmetric distribution. Then so has $\hat{a}$ given that the true regression line is the horizontal axis. It suffices to plot the frequencies for the absolute values. The right tail is heavy. It decreases like a power of $1/x$. The exponent of the tail of the density will be roughly -3 since we try to minimize the average quadratic loss, less if the second moment of $\hat{a}$ is finite, more if the second moment is infinite. For $x \to 0+$ the density will tend to infinity like $x^{1/\xi-1}$ for $\xi > 1$ as we shall see below. The frequency plot for $|\hat{a}|$ shows the central part of the density, a part which is of little interest. The loglog frequency plot which plots the log of the frequency of $\log|\hat{a}|$ is more interesting.

There is empirical evidence that EGBP densities $g = e^{-\psi}$ give a good fit for the frequency plots of $\log|\hat{a}_E|$ for good estimators E, in particular if the error has a symmetric distribution. The intuition behind this good fit is vague. The density $g_E = e^{-\psi_E}$ of $\log|\hat{a}_E|$ is smooth even when the density of the error has discontinuities since the value of $\hat{a}_E$ depends continuously on a hundred independent sample points. The absolute slope of the left asymptote of ψ_E corresponds to the power of x which describes the df of $|\hat{a}_E|$ at the origin. The power is $\lambda = 1/\xi$ for $\xi \geq 1$ since the most accurate estimates are due to large values of X_1. The right asymptotic slope of ψ_E determines the tail behaviour of $|\hat{a}_E|$. For estimators E with a parameter like RM(r) the loglog frequency plot often exhibits a number of isolated large values of $|\hat{a}_E|$ if the parameter is not optimal. These large values are due to outliers of the vertical coordinate Y_i^* for the rightmost points. For the optimal value of the parameter r these large values are eliminated by giving less weight to the extreme rightmost points.

Of the four parameters of the EGBP distribution there is one, the slope of the left asymptote of ψ_E which we can link to the tail indices (ξ, η). It would be of interest to know how the remaining three parameters depend on the tail indices, the shape of the error density and the estimator. In the text below we define the optimal fit. Let S denote the variable with density $e^{-\psi}$ where $y_0 - \psi$ yields the optimal fit. We consider two issues:

- How close is the theoretical sd $\sqrt{\mathbb{E}e^{2S}}$ to the empirical sd listed in the tables in Section 6?
- How does the distance between the loglog frequency plot of $|\hat{a}|$ and $e^{-\psi}$ compare to the distance between the log of the frequency plot of S and $e^{-\psi}$.



(a) WTS(2,3)(6)

(b) LADHC(1/2,2)(10)

(c) RM$_\infty$(3/2,1)(7)

(d) TB1, TB2, TB∞ (1,2)

Figure 12: EGBP approximations to the loglog frequency plots for various values of (ξ, η). Empirical sd and theoretical sd:

(a) Weighted Theil-Sen: 0.0018[1], 0.00173;

(b) LAD with Hyperbolic Correction: 0.0711[2], 0.07118;

(c) Right Median for the Poisson point process, $n = \infty$: 5.1[1], 5.03;

(d) Trimmed about the Bisector: 1 (black) 0.0197[2], 0.02044; 2 (red) 0.0237[2], 0.02460; ∞ (green) 0.0316[5], 0.03197.

Our loglog frequency plots are random piecewise linear functions with twenty bins per unit. We round off to an integer value m the random value $20 * \log|\hat{a}|$, count the number, n, of occurrences of m in the million simulations and connect the points $(m/20, \log n)$. The random fluctuations in the resulting plot $\mathbf{fr}$ are clearly visible, in particular in the tails. We choose the base line at level $-1/10$ so as to see unique occurrences, $n = 1$. The global shape is a concave curve like a parabola but with asymptotes with finite slopes. Let $g = e^{-\psi}$ be a logconcave density. Define the distance

$$d(\mathbf{fr}, g) = \sqrt{\mathbb{E}(\mathbf{fr} - (y0 - \psi))^2} \qquad y_0 = \log(1e6) - \log 20, \quad g = e^{-\psi}. \tag{10.8}$$

where we take the expectation with respect to the probability measure with atoms of size $n/10^6$ in $m/20$. In order to see whether the fit is good we simulate a million samples from the density g, construct $\mathbf{gr}$, the log of the corresponding frequency plot, and compute the distance $d(\mathbf{gr}, g)$. Actually we simulate twenty batches of a million samples and write down the average and sd of the distances $d(\mathbf{gr}, g)$ in the notation (1.3) introduced in the Introduction.

What logconcave density $g = e^{-\psi}$ should one choose to obtain a good fit? The density $g(s) = c/\cosh(s)$ is symmetric. The function $\varphi_0(s) = \log\cosh s$ has asymptotes with slope ± 1. It satisfies $\varphi_0(0) = \varphi_0'(0) = 0$ and $\varphi_0''(0) = 1$. A scale transformation in the vertical and the horizontal direction will transform φ_0 into the function $\varphi(s) = c\varphi_0(s/a)$ which satisfies $\varphi(0) = \varphi'(0) = 0$ and $\varphi''(0) = c/a^2$. The asymptotes of φ have slope $\pm c/a$. We still need to modify the function to have asymptotes whose absolute slopes assume different values. This may be achieved by replacing $1/\cosh(s)$ by one of the functions below:

$$e^{\theta s}/\cosh(s) \qquad 1/(e^{s/p} + e^{s/q}) \qquad e^{r_0 s}/(1 + e^{r_1} s).$$

Note that the scale transformation of φ in the vertical direction corresponds to a power transformation of the density. We now have a three-dimensional family of analytic convex functions φ which are determined by the absolute values of the asymptotic slopes and the curvature $\varphi''(0)$ at the origin. We still need a horizontal translation to fit the densities $g = e^{-\psi}$ to the frequency plot $\mathbf{fr}$.

Let Φ denote this four-dimensional space of convex analytic functions φ. Each function

$\varphi \in \Phi$ is determined by four parameters: $x_0, \varphi''(x_0), -\varphi'(-\infty), \varphi'(\infty)$ in $\mathbb{R} \times (0,\infty)^3$. The corresponding space of dfs is EGBP. For a given loglog frequency plot **fr** choose the density $g_0 = e^{-\varphi_0}$ with $\varphi_0 \in \Phi$ to minimize the distance $d(\mathbf{fr}, g_0)$. Let

$$d = d(\mathbf{fr}, g_0) = d_\Phi(\mathbf{fr})$$

denote this minimal distance. Let **gr** denote the log of the frequency plot for a sample of size $n = 10^6$ from g_0 and set$d_0 = d(\mathbf{gr}, g_0)$. It may happen that $d_0 > d$. Perhaps one should compare d not to d_0, the distance between **gr** and g_0, but to $d_1 = d_\Phi(\mathbf{gr})$, the minimum of $d(\mathbf{gr}, g)$ over all densities $g = e^{-\varphi}$, $\varphi \in \Phi$. Do that for twenty batches of a million simulations from the density g_0.

Example 5. For the estimators HB40(3), HB0(3) and RM(9) applied to samples of size $n = 100$ of points (X_i, Y_i) where the $X_1 > \cdots > X_{100}$ have a Pareto(1) distribution, $X_i = 1/U_i$, for the order statistics U_i of a sample from the uniform distribution on $(0,1)$ and errors Y_i from a Cauchy distribution scaled by its IQD we obtain $(d, d_0, d_1) = (0.01965, 0.019[1], 0.019[1])$, $(0.02036, 0.020[1], 0.019[1])$ and $(0.02077, 0.028[1], 0.020[1])$.

Here is a crude estimate of the distance $d(\mathbf{gr}, g)$ between the log of the frequency plot **gr** for a million simulations from the density g and the density. Let b_0 denote the number of non-empty bins. Then

$$d(\mathbf{gr}, g) \approx \sqrt{b_0}/1000.$$

In our case $b_0 \approx 340$ which gives $d \approx 0.017$.

Indeed the number N_k of sample points in the kth bin B_k is binomial(n, p_k) where $n = 10^6$ is the number of simulations and p_k is the integral of g over the bin B_k. Hence $N_k = np_k + \sigma_k U_k$ where $\sigma_k = \sqrt{np_k q_k}$ with $q_k = 1 - p_k$ close to one and U_k asymptotically standard normal for $np_k \to \infty$. Hence $\log N_k = \log np_k + U'_k/\sigma_k$ and

$$nd^2 = \sum N_k |\log N_k - \log(np_k)|^2 = \sum N_k (U'_k)^2/\sigma_k^2 \approx \sum (U'_k)^2 \approx b_0.$$

The condition that np_k be large does not hold for the bins in the tails. The loglog frequency plots are based on a million values of $\hat{a}$. Most of these occur in the center. The tails, say the part where the frequency, the number of entries N_k in a bin, is less than a hundred is less than 0.5% of the total. It is this part which determines the tail behaviour. The distance between the smooth EGBP-fit $y_0 - \psi$ and the loglog frequency plot **fr** is determined by the middle part. A good fit in the tails is a bonus.

10.7 Variations in the error density at the origin

What happens if one replaces the Student density by a symmetric density which is constant on a neigbourhood of the origin, or which has a vertex at the origin, or a zero, or a pole? We introduce three error variables with symmetric densities and with tail index $\eta = 1$. Start with variables Z_s, Z_u, Z_p with IQD $= 2$. These variables have symmetric densities and satisfy $\mathbb{P}\{Z > 1\} = 1/4$. The variable Z_s has a Student distribution. It is the Cauchy variable with density $(1/\pi)/(1+z^2)$. The variable Z_u has a density which is constant with value $1/4$ on the interval $(-1, 1)$ and has Pareto tails: $\mathbb{P}\{Z_u > z\} = 1/(4z)$, $z > 1$. The variable Z_p is the symmetric version of a shifted Pareto variable $\mathbb{P}\{Z_p > z\} = 1/(2+2z)$, $z > 0$.

Define the corresponding error variables $Y_t = Z_t/2$ for $t = s, u, p$. The explanatory variables are $X_i = 1/U_i$ where $U_1, \ldots, U_n$ are the increasing order statistics from the uniform distribution on $(0, 1)$. We shall use the estimators HB0[3] and HB40[3] to determine the empirical sd and to construct loglog frequency plots **fr**. One may expect HB40 to be less sensitive to the precise form of the density at the origin since it is based on the behaviour of the df at the 0.4 and 0.6 quantiles of the error distribution. Determine the EGBP approximation $g_0 = e^{-\psi_0}$ in these six cases, and the theoretical sd, the square root of $\int e^{2s} g_0(s)ds$. We also compute the distance $d = d(\mathbf{fr}, g_0)$. We shall then construct twenty plots **gr** corresponding to twenty batches of a million simulations from g_0, and compute the distances $d_0 = d(\mathbf{gr}, g_0)$ and $d_1 = d(\mathbf{gr}, g)$ where $y - \log g$ is the best EGBP-approximation to **gr**.

If U is uniformly distributed on the interval $(0, 1)$ then U^2 has density $1/(2\sqrt{u})$ and $\sqrt{U}$ has density $2u^2$ on $(0, 1)$. In general $\mathbb{P}\{U^r > x\} = x^{1/r}$ for $r > 0$. For the three variables Z_t introduced above define samples of $Z_t^{[r]}$ by replacing Z_t by $\text{sign}(Z_t)|Z_t|^r$ when Z_t lies in the interval $(-1, 1)$. The density $f_t^{[r]}$ of $Y_t^{[r]} = Z_t^{[r]}/2$ is asymptotic to $c_t|y|$ at the origin if $r = 1/2$ and asymptotic to $c'_t/\sqrt{|y|}$ if $r = 2$. We shall also check how good the EGBP-fit is for errors with these densities.

The final plot below shows the EGBP-fit to the loglog frequency plot for $\hat{a}_{\mathrm{LS}}$. The error has a Cauchy distribution scaled by its IQD, the explanatory variables are Pareto with tail index $\xi = 1$. The figure shows two things: LS is not a good estimator for $(\xi, \eta) = (1, 1)$. The right tail of the log log frequency plot extends beyond 10000. The EGBP fit is not good. The loglog frequency plots suggests a smooth logconcave density $e^{-\varphi}$ with asymptotes with finite non-zero slopes, although one sees a certain amount of rubble at the extreme right tail. The EGBP fit ignores the tails. It only looks at the central part around the top. This central part clearly does not fit with the tails of the loglog frequency plot.

density	emp sd	theor sd	$d = d(\mathbf{fr}, \psi_0)$	$d_0 = d(\mathbf{gr}, \psi_0)$	$d_1 = d(\mathbf{gr}, \psi)$
f_s	0.01158[5]	0.01158	0.01965	0.019[1]	0.019[1]
f_u	0.01105[5]	0.01106	0.01790	0.019[1]	0.019[1]
f_p	0.0122[2]	0.01223	0.02012	0.0200[5]	0.0192[5]

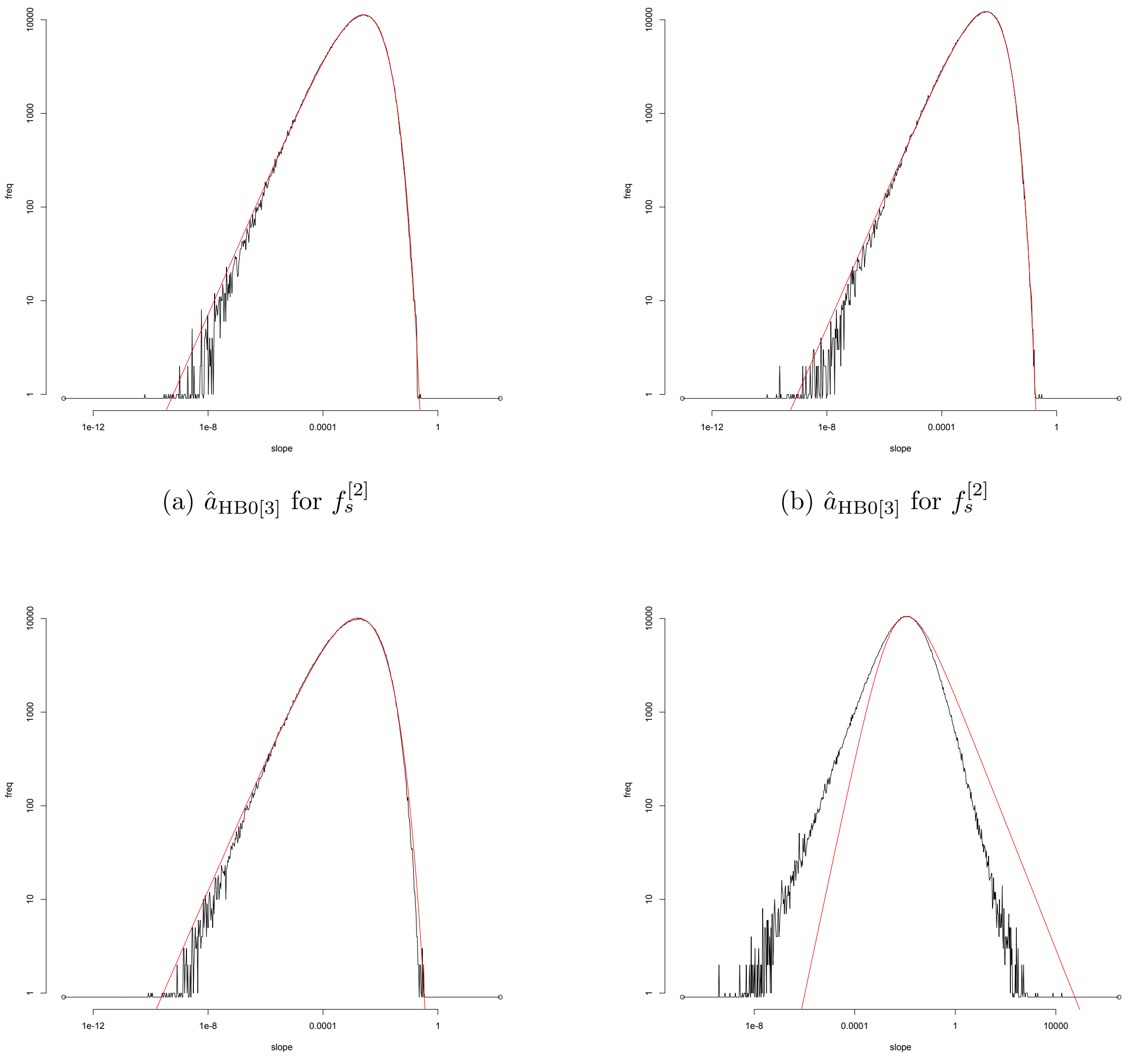
HB40(1, 1)[3]

density	emp sd	theor sd	$d = d(\mathbf{fr}, \psi_0)$	$d_0 = d(\mathbf{gr}, \psi_0)$	$d_1 = d(\mathbf{gr}, \psi)$
f_s	0.01201[5]	0.01200	0.02036	0.020[1]	0.019[1]
f_u	0.01205[5]	0.01218	0.01893	0.0192[5]	0.0188[5]
f_p	0.0120[1]	0.01217	0.02050	0.0193[5]	0.0189[5]

HB0(1, 1)[3]

density	emp sd	theor sd	$d = d(\mathbf{fr}, \psi_0)$	$d_0 = d(\mathbf{gr}, \psi_0)$	$d_1 = d(\mathbf{gr}, \psi)$
$f_s^{[2]}$	0.00952[5]	0.009545	0.02022	0.0194[5]	0.0193[5]
$f_u^{[2]}$	0.00963[5]	0.009643	0.02100	0.019[1]	0.019[1]
$f_p^{[2]}$	0.0096[2]	0.009824	0.02295	0.0203[5]	0.0201[5]

HB40(1, 1)[3]



(a) $\hat{a}_{\mathrm{HB0[3]}}$ for $f_s^{[2]}$

(b) $\hat{a}_{\mathrm{HB0[3]}}$ for $f_s^{[2]}$

(c) $\hat{a}_{\mathrm{HB0[3]}}$ for $f_s^{[2]}$

(d) $\hat{a}_{\mathrm{LS}}$ for f_s

Figure 13: EGBP approximations to three loglog frequency plots for $\hat{a}_{\mathrm{HB0[3]}}$ and one for $\hat{a}_{\mathrm{LS}}$ The error densities $f_s^{[2]}$, $f_u^{[2]}$, $f_p^{[2]}$ in the first three are asymptotic to $c/\sqrt{|y|}$ at the origin. In the fourth plot the error is Cauchy scaled by its IQD.

density	emp sd	theor sd	$d = d(\mathbf{fr}, \psi_0)$	$d_0 = d(\mathbf{gr}, \psi_0)$	$d_1 = d(\mathbf{gr}, \psi)$
$f_s^{[2]}$	0.00847[5]	0.008634	0.03120	0.0209[5]	0.0208[5]
$f_u^{[2]}$	0.00881[5]	0.008888	0.03094	0.0200[5]	0.0199[5]
$f_p^{[2]}$	0.0083[1]	0.008916	0.03863	0.0210[5]	0.0209[5]

$\mathrm{HB0}(1,1)[3]$

density	emp sd	theor sd	$d = d(\mathbf{fr}, \psi_0)$	$d_0 = d(\mathbf{gr}, \psi_0)$	$d_1 = d(\mathbf{gr}, \psi)$
$f_s^{[1/2]}$	0.01209[5]	0.01226	0.01971	0.0194[5]	0.0190[5]
$f_u^{[1/2]}$	0.01103[5]	0.01115	0.01892	0.0193[5]	0.0190[5]
$f_p^{[1/2]}$	0.0134[2]	0.01338	0.02144	0.020[1]	0.019[1]

$\mathrm{HB40}(1,1)[3]$

density	emp sd	theor sd	$d = d(\mathbf{fr}, \psi_0)$	$d_0 = d(\mathbf{gr}, \psi_0)$	$d_1 = d(\mathbf{gr}, \psi)$
$f_s^{[1/2]}$	0.01518[5]	0.01525	0.01727	0.019[1]	0.019[1]
$f_u^{[1/2]}$	0.0148[1]	0.01482	0.01876	0.018[1]	0.018[1]
$f_p^{[1/2]}$	0.0157[2]	0.01585	0.02003	0.0191[5]	0.0186[5]

$\mathrm{HB0}(1,1)[3]$

It is the distribution of the estimator $\hat{a}$ which determines how good it is, in particular the right tail of the distribution of the absolute value. This right tail determines the risk associated with the estimate. The loglog frequency plot gives a good description of the tail behaviour. A steep decrease on the right, indicating an asymptote with large absolute slope, is ideal. A good EGBP fit means that the logconcave density of the EGBP variable S agrees well with the distribution of the variable $\log|\hat{a}|$. If moreover the theoretical sd $\sqrt{\mathbb{E}e^{2S}}$ is close to the empirical sd associated with the loglog frequency plot that indicates that the good fit extends to the right tail. The simulations on which this paper is based suggest that minimizing the average quadratic loss leads to loglog frequency plots which may be fitted accurately by the exponent of the logconcave EGBP densities, and that this good fit extends to the right tail. These empirical results vindicate the choice of average quadratic loss of $\hat{a}$ as a measure of the performance of estimators in linear regression for heavy tails.

Acknowledgements

The authors would like to thank Holger Drees who read an earlier version of the first half of the MS for his valuable comments.

References

[1] A.A.Balkema (2019) Least Absolute Deviation and balance. To appear.

[2] A.A. Balkema and P. Embrechts (2007) *High Risk Scenarios and Extremes. A geometric Approach.* Zurich Lectures in Advanced Mathematics. European Mathematical Society, Zurich.

[3] G. Bassett Jr and R. Koenker (1978) Asymptotic theory of Least Absolute Error regression. *J of the Amer. Statist. Assn* **73**, pp 618–622.

[4] T. E. Dielman (2005) Least absolute value regression: recent contributions. *J of statist. computation and simulation* **75**, pp 263–286.

[5] H. Drygas (1976) Weak and strong consistency of least squares estimates in regression models. *Z. Wahrsch* **34**, pp 119–127.

[6] A. Eddington (1914) *Stellar Movements and the Structure of the Universe.* Macmillan, London.

[7] F.Y. Edgeworth (1887) A new method of reducing observations relating to several quantities. *Philos. Mag. (5th Ser)* **24**, pp 222–223.

[8] P. Embrechts, C. Klüppelberg and T. Mikosch (1997) *Modeling Extremal Events for Insurance and finance.* Springer-Verlag, Berlin.

[9] S.A. van de Geer (1988) Asymptotic normality of minimum $\mathbf{L}^1$-norm estimators in linear regression. *Report* **MS-R8806** 5pp. CWI, Amsterdam.

[10] L. de Haan and A. Ferreira (2006) *Extreme Value Theory: An Introduction.* Springer.

[11] J. E. Heffernan and J. A. Tawn (2004) A conditional approach for multivariate extreme values. *J. of the Royal Statist. Soc.* **B66**, pp 497–546.

[12] J. E. Heffernan and S.I. Resnick (2007) Limit laws for random vectors with an extreme component. *Annals of Appl. Probab.* **17**, pp 537–571.

[13] L. A. Jaeckel (1972) Estimating regression coefficients by minimizing the dispersion of the residuals. Ann. Math. Statist. **43**, pp 1449–1458.

[14] R. Koenker (2000) Galton, Edgeworth, Frisch, and the prospects for quantile regression in econometrics. *J of Econometrics* **95**, pp 347–374.

[15] R. Koenker and G. Bassett (1985) On Boscovich's estimator. *Ann. of Statistics***13** pp 1625–1628.

[16] Chung-Ming Kuan (2007) Asymptotic Least Squares Theory: Part I. `http://homepage.ntu.edu.tw/~ckuan/pdf/et01/et_Ch6.pdf`

[17] E.L. Lehmann (1983) *Theory of Point Estimation.* Wiley, New York.

[18] T. Mikosch and C.G. de Vries (2013) Heavy tails of OLS. *Journal of Econometrics*, **172**, 205-221.

[19] J. P. Nolan and D. Ojeda-Revah (2013) Linear and non-linear regression with stable errors. *Journal of Econometrics*, **172**, 186–194.

[20] E. B. Postnikov and I. M. Sokolov (2015) Robust linear regression with broad distributions of errors. *Physica A***434**, pp 257–267.

[21] P.J. Rousseeuw (1984) Least median of squares regression. *J. of the Amer. Statist. Assn* **79**, pp 871–880.

[22] P.J. Rousseeuw (1991) Tutorial to robust statistics. *J. of Chemometrics* **5**, pp 1–20.

[23] D. Ruppert and R.J. Carroll (1980) Trimmed Least Squares estimation in the Linear Model. *J. of the Amer. Statist. Assn* **75**, pp 828–838.

[24] G. Samorodnitsky, S.T. Rachev, J.-R. Kurz-Kim and S.V. Stoyanov (2007) Asymptotic distribution of unbiased linear estimators in the presence of heavy-tailed stochastic regressors and residuals. *Probab. and Mathem. Statistics* **27**, pp 275–302.

[25] P. K. Sen (1968) Estimates of the regression coefficient based on Kendall's tau. *J. of the Amer. Statist. Assoc.* **63**, pp 1379–1389.

[26] A.F. Siegel (1982) Robust regression using repeated medians. *Biometrika*

[27] G.L. Sievers (1978) Weighted rank statistics for simple linear regression. *J. of the Amer. Statist. Assoc.* **73**, pp 628–631.

[28] H. Theil (1950) A rank-invariant method of linear and polynomial regression analysis. *Proceedings of the KNAW* **53**, pp 386–392, 521–525, 1397–1412.

[29] H.H. Turner (1887) On Mr. Edgeworth's *Method of reducing observations relating to several quantities. Philos. Mag. (5th Series)* **24**, pp 466–470.