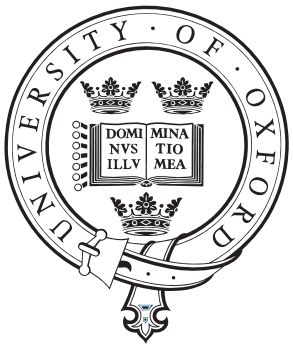
Dimension reduction of streaming data via random projections

Ioana Ada Cosma

Department of Statistics

University of Oxford

Trinity Term, 2009



Thesis submitted for the degree of Doctor of Philosophy at the University of Oxford.

Dimension reduction of streaming data via random projections

Ioana Ada Cosma, Jesus College

DPhil, Trinity Term, 2009

A data stream is a transiently observed sequence of data elements that arrive unordered, with repetitions, and at very high rate of transmission. Examples include Internet traffic data, networks of banking and credit transactions, and radar derived meteorological data. Computer science and engineering communities have developed randomised, probabilistic algorithms to estimate statistics of interest over streaming data on the fly, with small computational complexity and storage requirements, by constructing low dimensional representations of the stream known as data sketches.

This thesis combines techniques of statistical inference with algorithmic approaches, such as hashing and random projections, to derive efficient estimators for cardinality, l_α distance and quasi-distance, and entropy over streaming data. I demonstrate an unexpected connection between two approaches to cardinality estimation that involve indirect record keeping: the first using pseudo-random variates and storing selected order statistics, and the second using random projections. I show that l_α distances and quasi-distances between data streams, and entropy, can be recovered from random projections that exploit properties of α-stable distributions with full statistical efficiency. This is achieved by the method of L-estimation in a single-pass algorithm with modest computational requirements. The proposed estimators have good small sample performance, improved by the methods of trimming and winsorising; in other words, the value of these summary statistics can be approximated with high accuracy from data sketches of low dimension.

Finally, I consider the problem of convergence assessment of Markov Chain Monte Carlo methods for simulating from complex, high dimensional, discrete distributions. I argue that online, fast, and efficient computation of summary statistics such as cardinality, entropy, and l_α distances may be a useful qualitative tool for detecting lack of convergence, and illustrate this with simulations of the posterior distribution of a decomposable Gaussian graphical model via the Metropolis-Hastings algorithm.

Acknowledgments

I am immensely grateful to my supervisor, Dr. Peter Clifford, for showing me the path of rigorous research, for his dedicated mentoring, and many words of encouragement along the way, to Professor Steffen Lauritzen for insightful discussions, suggestions, and advice, and finally to my family and friends for their love and support.

Contents

Chapter 1

Introduction

Over the last three decades, increasingly large data sets have appeared in a wide range of online applications in science and commerce, for example, Internet traffic on routers (Akella et al., 2003; Cormode and Muthukrishnan, 2005b), time series data such as stock levels (Indyk et al., 2000), database applications (Flajolet and Martin, 1985; Whang et al., 1990), and clustering of music files (Yang, 2001) or Web documents (Haveliwala et al., 2000). The need to develop techniques capable of handling such extensive data continues to be the driving force behind much of the ongoing research in computer science and engineering. This thesis investigates new methodologies for analysing high throughput streaming data using random projection methods (Indyk, 2006; Vempala, 2004).

A *data stream* is a transiently observed sequence of data elements that arrive unordered, with repetitions, and at very high rate of transmission, such that it may be hard to transmit the entire input, compute functions on all or part of it, or store it temporarily for subsequent access (Aggarwal, 2007). Examples include data streams generated by automatic, highly detailed data feeds from applications that monitor financial, atmospheric, astronomical, and networking activity in real time, for the purpose of detecting fraud, extreme events, and unusual, or anomalous activity (Muthukrishnan, 2005). Data streams encountered in real-life applications are typically massive, having possibly thousands of millions of elements of

many different types. The complete data set is typically too large to fit in main computer memory and, if stored on disk, random accesses to the data would considerably slow down the performance of many statistical algorithms.

In comparison to the traditional situation in which all the data are stored and available for access, stream data are observed within a narrow time window. At the current time point, a small portion of the entire data set is observed, processed, and discarded. The retained information must accurately represent important features of all the data observed up to the current time point and be of relevance for estimating the query of interest; this process is known as *synopsis construction* (Aggarwal, 2007). In contrast, the sliding window technique evaluates the query on the most recent data observed in the stream, under the assumption that in real-life applications, recent data is more important and relevant than old data (Babcock et al., 2002). Moreover, algorithms for processing streaming data must have low storage and time complexity (usually by allowing only one pass over the data), and be insensitive to the pattern of repetitions in the stream.

Aggarwal (2007) surveys a variety of techniques for synopsis construction of data streams, e.g., sampling, wavelet-based techniques, and histogram representations. The problems encountered by sampling methods well exemplify the difficulty in handling data streams. Since no distributional assumptions are allowed regarding the nature of the data, and the size of the data stream is not known in advance, the probability of including a point in the sample must change as the stream evolves; see reservoir-based sampling described in Aggarwal (2007). Moreover, for real-life applications, interest centres on representing infrequent and anomalous behaviour, i.e., the sample must include points that lie far out in the tails of the evolving distribution of the data, and that are very difficult to capture.

We are interested in a method of synopsis construction known as *data sketching* (Indyk, 2000) that summarises the data stream by a relatively low dimensional vector, updated on-the-fly as the stream evolves. We will focus on two different approaches: hashing to

independent copies of random variables and storing some function of these, e.g., order statistics (Giroire, 2005), or generating vectors of random variables from the α-stable distribution and storing linear combinations of these (Indyk, 2006). Data sketching methods offer space-efficient representations particularly suited for applications that require estimates of the distribution of the observed elements, such as the number of distinct elements (cardinality) or frequency moments.

A data stream can be viewed as a high-dimensional data point if represented by a vector of totals of each element type. There is a long history in the statistics literature of methods aiming to find low dimensional representations of high-dimensional data that highlight interesting features of the original data; these methods are globally termed *dimension reduction*. Examples include projection pursuit (Huber, 1985), principal components, discriminant analysis, and multidimensional scaling (Mardia et al., 1997; Hastie et al., 2001). Diaconis and Freedman (1984) show that most low-dimensional projections of high-dimensional points are approximately normal, and argue that the most interesting one-dimensional projections are those that are far from Gaussian. As the dimension increases, these methods become too computationally expensive to be of practical value.

We are interested in the recovery of l_α distances (quasi-distances) between data streams from data sketches based on α-stable random projections (Indyk, 2006). Distance-preserving projection methods have important applications in clustering and classification of high-dimensional data sets. For streaming data, they provide Hamming distance approximations and other measures of distributional dissimilarity (Li et al., 2007; Li and Hastie, 2008).

The present thesis is organised as follows. Chapter 2 introduces the stable law distribution that lies at the heart of the random projection method. In particular, we present parameterisations, properties, and sampling algorithms, as well as a novel result on the influence of the maximal term on the sum of i.i.d. positive, strictly stable random variables as the index of stability tends to zero, i.e., as the stable law approaches a degenerate distribution

at zero. Furthermore, we investigate and correct a numerical instability in the behaviour of algorithms for density, distribution, and quantile approximation of the symmetric, strictly stable distribution provided in the R contributed package fBasics.

Chapter 3 is devoted to the problem of cardinality estimation of data streams, i.e., the problem of estimating the number of distinct elements. We focus on two approaches in the literature that involve indirect record keeping: the first using pseudo-random variates and storing selected order statistics, and the second using random projections. We propose recursively computable, maximum likelihood estimators of the cardinality derived from low-dimensional data sketches in both cases, and show that the estimators have comparable asymptotic efficiency. Moreover, we explain this result by demonstrating a previously unsuspected link between the two approaches.

Chapter 4 considers data sketches based on α-stable projections for $\alpha \in (0, 2]$ and proposes algorithms with modest computational requirements that recover l_α distances (quasi-distances) from such sketches with full statistical efficiency. We reduce the problem to that of estimating location or scale parameters of transformed stable random variables. For the latter problem, we propose simple, asymptotically efficient estimators based on the method of L-estimation (Chernoff et al., 1967) that outperform existing estimators.

Chapter 5 investigates projections based on maximally skewed stable random variables with application to the problem of entropy estimation over streaming data. Finally, we argue that online, fast, and efficient computation of cardinality, entropy, and l_α distance estimates may be a useful tool for qualitative assessment of convergence of Markov Chain Monte Carlo (MCMC) methods in high-dimensional problems, and illustrate this with simulations of the posterior distribution of a decomposable Gaussian graphical model.

Finally, Chapter 6 concludes the thesis with a review of the results presented therein, and discusses possible directions of future research.

Chapter 2

The stable law

The stable distribution (Lévy, 1924) has been extensively studied for modelling heavy tailed data (Nolan, 2007). It is the only limiting distribution of normalised sums of independent and identically distributed (i.i.d.) random variables (having possibly infinite variance); a result known as the Generalized Central Limit Theorem (Breiman, 1992). This chapter presents some properties of the stable law, algorithms for generating variables from this distribution, as well as a novel result on sums of i.i.d. stable random variables. Since the density of the stable distribution does not exist in closed form for most parameter values, it is necessary to estimate the density, distribution, and quantile functions. For this purpose, we employ the statistical software R and the contributed package fBasics; we identify a numerical instability in the density, distribution, and quantile estimation procedures in fBasics, in the case of the symmetric stable distribution, and propose numerically stable versions of these procedures.

2.1 Properties and parameterisations

The following stability property is the key element of the method of dimension reduction via random projections, that maps from a high-dimensional to a lower-dimensional space by weighted linear combinations of stable random variables.

Definition 2.1.1. *A random variable X with distribution F is said to be stable if for every $n > 0$, and independent variables $X_1, \ldots, X_n \sim F$, there exist constants $a_n > 0$ and b_n such that*

$$X_1 + \ldots + X_n \stackrel{\mathcal{D}}{=} a_n X + b_n, \tag{2.1}$$

where $\stackrel{\mathcal{D}}{=}$ denotes equality in distribution. If $b_n = 0$, then X is said to be strictly stable.

Feller (1971) proves that the only possible norming constants in (2.1) are $a_n = n^{1/\alpha}$, where $0 < \alpha \leq 2$. The parameter α is known as the *characteristic exponent* or *index of stability*. Stable distributions of index α exist for all values of α in $(0, 2]$; the density functions are smooth and approach a degenerate distribution as $\alpha \to 0$ (Feller, 1971; Zolotarev, 1986). Furthermore, if X is stable of index $\alpha < 2$, then the rth moment of X, $\mathbb{E}|X|^r$, is finite for $0 < r < \alpha$ and is infinite for $r \geq \alpha$ (Feller, 1971); it follows that the stable distribution has infinite variance $\forall \alpha < 2$.

Of particular interest is the following property: all linear combinations of independent, identically distributed stable random variables belong to the same type (Feller, 1971).

Theorem 2.1.1. *The distribution F is stable of index α if and only if for arbitrary constants a_1 and a_2, and random variables $X_1, X_2 \sim F$, there exist constants a and b such that*

$$a_1 X_1 + a_2 X_2 \stackrel{\mathcal{D}}{=} aX + b, \quad \text{where } X \sim F. \tag{2.2}$$

In particular, if the distribution F is strictly stable, and $a_1, a_2 > 0$, then

$$a_1 X_1 + a_2 X_2 \stackrel{\mathcal{D}}{=} \left(|a_1|^\alpha + |a_2|^\alpha\right)^{1/\alpha} X. \tag{2.3}$$

If, in addition, the distribution F is symmetric, then (2.3) holds for all $a_1, a_2 \in \mathbb{R}$.

2.1.1 Common parameterisations

We present several of the most commonly encountered parameterisations of the stable law existing in the literature (Chambers et al., 1976; Cheng and Liu, 1997; Kolokoltsov et al.,

2001; Nolan, 2007; Samorodnitsky and Taqqu, 1994; Zolotarev, 1986). We explicitly state the connections between them, lacking in current literature. The stable law is parameterised by four parameters: index $\alpha \in (0, 2]$, skewness parameter $\beta \in [-1, 1]$, location parameter $\delta \in \mathbb{R}$, and scale parameter $\gamma > 0$. The parameters δ and γ do not have the interpretation of location and scale in all parameterisations. We present the different parameterisations in terms of the characteristic function (c.f.) $\phi(t) = \mathbb{E}\exp(itX)$, $t \in \mathbb{R}$.

Zolotarev (1986) introduces the following three parameterisations. The c.f. of a stable random variable under parameterisation (A) is given by

$$\phi(t) = \exp\big(\gamma[it\delta - |t|^\alpha + it\omega_A(t, \alpha, \beta)]\big),$$

where

$$\omega_A(t, \alpha, \beta) = \begin{cases} |t|^{\alpha-1}\beta\tan(\pi\alpha/2) & \text{if } \alpha \neq 1, \\ -\beta(2/\pi)\log|t| & \text{if } \alpha = 1. \end{cases}$$

This parameterisation is not continuous in all 4 parameters; in particular, it has discontinuities at all points of the form $\alpha = 1$, $\beta \neq 0$. Moreover, the parameters δ and γ lack an intuitive interpretation.

A slight modification of Zolotarev's parameterisation (A) results in the parameterisation of Samorodnitsky and Taqqu (1994). Let $\delta = \delta_A\gamma_A$, and $\gamma = \gamma_A^{1/\alpha}$, where δ_A and γ_A are the parameters in (A); α and β remain unchanged. The c.f. has the following form

$$\phi(t) = \exp\big(\gamma^\alpha[-|t|^\alpha + it\omega_A(t, \alpha, \beta)] + i\delta t\big). \tag{2.4}$$

This is the S1 parameterisation in the R package fBasics, and also appears in Nolan (2007) as the $S(\alpha, \beta, \gamma, \delta; 1)$ parameterisation introduced in Chapter 1. Under this parameterisation, for $\alpha \neq 1$, δ and γ have the intuitive interpretation of location and scale parameters.

A modification of Zolotarev's parameterisation (A) gives parameterisation (M) with c.f.

$$\phi(t) = \exp\big(\gamma[it\delta - |t|^\alpha + it\omega_M(t, \alpha, \beta)]\big),$$

where

$$\omega_M(t,\alpha,\beta) = \begin{cases} (|t|^{\alpha-1}-1)\beta\tan(\pi\alpha/2) & \text{if } \alpha \neq 1, \\ -\beta(2/\pi)\log|t| & \text{if } \alpha = 1. \end{cases}$$

Let γ_1 and δ_1 denote the parameters under parameterisation S1 above; α and β remain unchanged. Then, $\gamma = \gamma_1^\alpha$, and $\delta = \delta_1\gamma_1^{-1}$ if $\alpha = 1$, and $\delta = \delta\gamma^{-\alpha} + \beta \times \tan\big(\pi\alpha/2\big)$ if $\alpha \neq 1$. This parameterisation is continous in all 4 parameters, i.e., $\lim_{\alpha\to 1}\omega_M(t,\alpha,\beta) = \omega_M(t,1,\beta)$. Furthermore, for δ and γ to have the interpretation of location and scale, define

$$\gamma = \gamma_M^{1/\alpha} = \gamma_1, \tag{2.5}$$

$$\begin{aligned} \delta &= -\gamma_M^{1/\alpha}\omega_M(t,\alpha,\beta) + \gamma_M\delta_M + \gamma_M\omega_M(t/\gamma_M^{1/\alpha},\alpha,\beta) \qquad (2.6)\\ &= \begin{cases} \delta_1 + \beta 2/\pi\gamma_1\log\gamma_1 & \text{if } \alpha = 1, \\ \delta_1 + \gamma_1\beta\tan(\pi\alpha/2) & \text{if } \alpha \neq 1, \end{cases} \end{aligned}$$

under the constraint that $\lim_{(\delta,\gamma)\to(0,1)}(\delta_M,\gamma_M) = (0,1)$, where δ_M, γ_M come from parameterisation (M), and δ_1, γ_1 from parameterisation S1. Again, α and β are unchanged. Then the c.f. of the new parameterisation is

$$\phi(t) = \exp\big(\gamma^\alpha[-|t|^\alpha + it(\delta\gamma^{-\alpha} + (\gamma^{1-\alpha}+1)\omega_M(t,\alpha,\beta) - \omega_M(t/\gamma,\alpha,\beta))]\big),$$

continuous in all 4 parameters. This is the S0 parameterisation in fBasics.

The third parameterisation introduced in Zolotarev (1986) is known as parameterisation (B) with c.f.

$$\phi(t) = \exp\big(\gamma[it\delta - |t|^\alpha\omega_B(t,\alpha,\beta)]\big),$$

where

$$\omega_B(t,\alpha,\beta) = \begin{cases} \exp\big(-i(\pi/2)\beta K(\alpha)\operatorname{sign} t)\big) & \text{if } \alpha \neq 1, \\ \pi/2 + i\beta\log|t|\operatorname{sign} t & \text{if } \alpha = 1, \end{cases}$$

and $K(\alpha) = \alpha - 1 + \operatorname{sign}(1-\alpha)$. Again, this parameterisation is not continuous in all 4 parameters. The connection betweeen parameterisations (A), (M) and (B) is explicitly stated in Zolotarev (1986). In the sequel, we cite many results from Zolotarev (1986), particularly

on the asymptotic and integral representations of stable density and distribution functions; most of these results appear in the book under parameterisation (B), which is connected to parameterisation (A) as follows. Let β_B, γ_B, and δ_B denote the parameters under the former parameterisation. Then, if $\alpha = 1$,

$$\beta_A = \beta_B, \quad \gamma_A = \frac{\pi}{2}\gamma_B, \quad \delta_A = \frac{2}{\pi}\delta_B,$$

and if $\alpha \neq 1$,

$$\begin{aligned} \beta_A &= \cot\big(\frac{\pi\alpha}{2}\big)\tan\big(\frac{\pi}{2}\beta_B K(\alpha)\big) \\ \gamma_A &= \gamma_B \cos\big(\frac{\pi}{2}\beta_B K(\alpha)\big) \\ \delta_A &= \delta_B\big(\cos\big(\frac{\pi}{2}\beta_B K(\alpha)\big)\big)^{-1}. \end{aligned}$$

We employ the parameterisation in Samorodnitsky and Taqqu (1994) given by equation (2.4). Let $f(x;\alpha,\beta,\gamma,\delta)$, and $F(x;\alpha,\beta,\gamma,\delta)$ denote the corresponding probability density function (p.d.f.) and cumulative distribution function (c.d.f.). If $\gamma = 1$ and $\delta = 0$, the distribution is said to be standardised and the symbol $F(x;\alpha,\beta)$ will be used as an abbreviation for $F(x;\alpha,\beta,1,0)$; we define $f(x;\alpha,\beta)$ similarly.

2.1.2 Properties of the stable distribution

The parameterisation in Samorodnitsky and Taqqu (1994) has some useful properties. First, it can easily be shown that if $X \sim F(x;\alpha,\beta,\gamma,\delta)$, then

$$\begin{aligned} \frac{X-\delta}{\gamma} &\sim F(x;\alpha,\beta), \quad \text{if } \alpha \neq 1, \\ \frac{X-\delta-2\beta\gamma\log\gamma/\pi}{\gamma} &\sim F(x;1,\beta), \quad \text{if } \alpha = 1. \end{aligned}$$

Moreover, it follows from results in Nolan (2007) that the distribution is strictly stable if $\delta = 0$ ($\alpha \neq 1$) or $\beta = 0$ ($\alpha = 1$). The distribution is symmetric around 0 if $\beta = 0$ and $\delta = 0$, in which case the characteristic function simplifies to $\phi(t) = \exp(-\gamma^\alpha|t|^\alpha)$.

The stable distribution is defined on the half line if $\alpha < 1$ and $\beta = 1$ or $\beta = -1$; in particular, the support of the density is (δ, ∞) if $\beta = 1$, $(-\infty, \delta)$ if $\beta = -1$, when $0 < \alpha < 1$, and $(-\infty, \infty)$ otherwise. If the distribution is strictly stable, then the support is $(0, \infty)$ for $\beta = 1$, and $(-\infty, 0)$ for $\beta = -1$, provided $0 < \alpha < 1$. In other words, there is no strictly stable, positive law of index $\alpha \geq 1$ (Feller, 1971). Figures 2.1 and 2.2 display the density functions of the positive strictly stable distribution and the symmetric, strictly stable distribution, respectively; these distributions are of particular interest in the sequel.

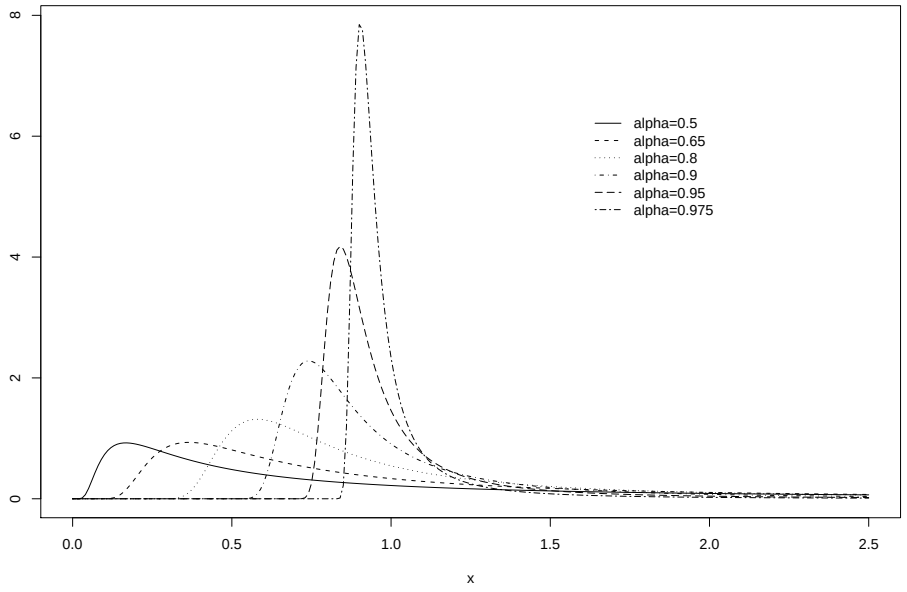


Figure 2.1: The positive, strictly stable density of index α and $\beta = 1$ (command *dstable* in fBasics).

Non-degenerate stable distributions are continuous with infinitely differentiable density functions; moreover, the derivatives of the density functions are uniformly bounded. Following Zolotarev (1986) (p. 13), it can be shown that $|f^{(n)}(x; \alpha, \beta, \gamma, \delta)| \leq (\pi\alpha)^{-1}\Gamma((n + 1)/\alpha)\gamma^{-(n+1)}$, for $n \geq 0$. Furthermore, the density and distribution functions satisfy several reflection properties: $f(x; \alpha, \beta) = f(-x; \alpha, -\beta)$, and $F(x; \alpha, \beta) = 1 - F(-x; \alpha, -\beta)$.

The stable distribution can be expressed in closed form for only three distinct values of

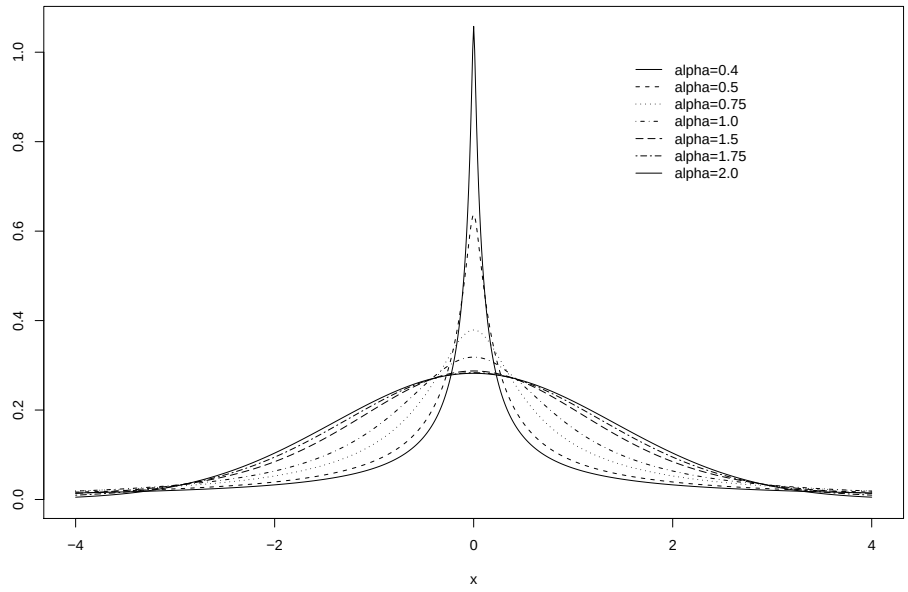


Figure 2.2: The symmetric, strictly stable density function $f(x;\alpha,0,1,0)$ (*dstable*, parameterisation S1, fBasics).

α: $F(x;2,0,\gamma,\delta)$ is the Normal distribution with mean δ and variance $2\gamma^2$, $F(x;1,0,\gamma,\delta)$ is the Cauchy distribution with location parameter δ and scale parameter γ, and the stable law with $\alpha = 0.5$ and skewness parameter $\beta = 1$, $F(x;0.5,1,\gamma,\delta)$, is the Lévy distribution with location and scale parameters δ and γ, respectively.

Formulae for generating stable random variables are a good starting point in deriving results about the stable distribution. Chambers et al. (1976) present a method for generating stable random variables according to Zolotarev's parameterisation (B), proved by Weron (1996). Let $X \sim F(x;\alpha,\beta,1,0)$. Let $L \sim \text{Exp}(1)$, and $U \sim \text{Unif}[-\pi/2,\pi/2]$, independently. Define $U_0 = \alpha^{-1}\arctan\big(\beta\tan(\pi\alpha/2)\big)$. Then,

$$X \stackrel{\mathcal{D}}{=} \frac{\sin\big(\alpha(U_0+U)\big)}{(\cos(\alpha U_0)\cos U)^{1/\alpha}} \times \left(\frac{\cos\big(\alpha U_0 + (\alpha-1)U\big)}{L}\right)^{(1-\alpha)/\alpha}, \tag{2.7}$$

if $\alpha \neq 1$, and,

$$X \stackrel{\mathcal{D}}{=} \frac{2}{\pi}\left[\left(\frac{\pi}{2}+\beta U\right)\tan U - \beta\log\left(\frac{\pi/2 L\cos U}{\pi/2+\beta U}\right)\right],$$

if $\alpha = 1$. The fBasics command *rstable* for generating stable random variables is incorrect when $\alpha = \beta = 1$, computing an infinite value that is evaluated as $1.633178e + 16$, so in this case we use the above expression.

Lastly, we mention some results on the limiting distribution of normalised sums of i.i.d. random variables. Let $S_n = \sum_{i=1}^{n} X_i^{(n)}$, where $X_i^{(n)}$, $i = 1, \ldots, n$, are i.i.d., with distribution function F. With no assumption on the common variance of the $X_i^{(n)}$, Breiman (1992) proves that $S_n \xrightarrow{\mathcal{D}} X$ if and only if X has an infinitely divisible distribution, i.e., for every $m > 0$, there exist i.i.d. random variables $Y_1^{(m)}, \ldots, Y_m^{(m)}$ such that $X \stackrel{\mathcal{D}}{=} \sum_{j=1}^{m} Y_j^{(m)}$. Furthermore, the distribution of normalised sums $S_n/A_n - B_n$, for some B_n and $A_n > 0$, converges to a distribution G if and only if G is stable. This is the Generalised Central Limit Theorem.

A distribution F is in the domain of attraction of a non-degenerate distribution G if there exist constants $A_n > 0$ and B_n such that the distribution of $S_n/A_n - B_n$ converges to G as $n \to \infty$. Gnedenko and Kolmogorov (1954) prove that S_n, suitably normalised, has a nondegenerate limiting distribution as $n \to \infty$ if and only if the distribution F of the individual components X_i is in the domain of attraction of a stable law. We conclude that only stable laws have non-empty domains of attraction; in other words, it is necessary and sufficient that F belong to the domain of attraction of a stable law (Ibragimov and Linnik, 1971).

2.2 The positive, strictly stable law

We present the following result from Feller (1971) on the positive, strictly stable distribution, with distribution function denoted by $S(x; \alpha)$.

Theorem 2.2.1. *For fixed* $0 < \alpha < 1$, *the function* $\omega(\lambda) = e^{-\lambda^\alpha}$, $\lambda \geq 0$ *is the Laplace*

transform of the density of $S(x;\alpha)$ *defined on* $(0,\infty)$ *satisfying*

$$\begin{aligned} x^{\alpha}\big[1 - S(x;\alpha)\big] &\rightarrow \frac{1}{\Gamma(1-\alpha)}, \quad \textit{as } x \rightarrow \infty, \\ e^{x^{-\alpha}} S(x;\alpha) &\rightarrow 0, \quad \textit{as } x \rightarrow 0. \end{aligned} \tag{2.8}$$

The distribution $S(x;\alpha)$, $\alpha \in (0,1)$, has parameters $\beta = 1$, $\gamma = 1$, and $\delta = 0$ under Zolotarev's parameterisation (B) (Chambers et al., 1976). By comparing this parameterisation to that of Samorodnitsky and Taqqu (1994), it follows that if X is stable with parameters $\alpha \neq 1$, β, $\gamma = 1$, and $\delta = 0$ under parameterisation (B), then X is stable with parameters α, β_1, γ_1, and $\delta = 0$ under the latter, where $\tan(\pi\alpha/2)\beta_1 = \tan(\pi\beta\kappa(\alpha)/2)$ and $\gamma_1^{\alpha} = \cos(\pi\beta\kappa(\alpha)/2)$. We then conclude that the $S(x;\alpha)$ distribution is the same as the $F\big(x;\alpha,1,\gamma = \big(\cos(\pi\alpha/2)\big)^{1/\alpha}, 0\big)$ distribution.

Zolotarev (see reference in Zolotarev (1986)), and later Kanter (1975), showed the following result used to simulate from this distribution. For fixed $\alpha \in (0,1)$, define

$$a(u) = \left(\frac{\sin(\alpha u)}{\sin(u)}\right)^{(1-\alpha)^{-1}} \left(\frac{\sin\big((1-\alpha)u\big)}{\sin(\alpha u)}\right).$$

Let $L \sim \text{Exp}(1)$, and $U \sim \text{Unif}[0,\pi]$, independently. Then, $\big(a(U)/L\big)^{(1-\alpha)/\alpha} \sim S(x;\alpha)$, i.e., $\big(a(U)/L\big)^{(1-\alpha)/\alpha} \stackrel{\mathcal{D}}{=} \big(\cos(\pi\alpha/2)\big)^{1/\alpha} Y$, where $Y \sim f(y;\alpha,1,1,0)$.

The distribution $S(x;\alpha)$ is of particular interest in this thesis, and appears in the main result of the present chapter, Theorem 2.3.1. Darling (1952) analyses the manner in which the largest term influences the sum of positive, i.i.d. random variables, when the sum has a limiting stable distribution. Feller (1971) presents a result on the influence of the maximal term on the sum of i.i.d. random variables with common distribution belonging to the domain of attraction of the $S(x;\alpha)$ distribution. The following theorem is a less general version of that result; refer to Appendix A for the proof.

Theorem 2.2.2. *Let* $X_1, X_2, \ldots, X_n \sim S(x;\alpha)$ *be i.i.d. random variables with* $0 < \alpha < 1$ *fixed. Let* $S_n = \sum_{i=1}^{n} X_i$ *and* $X_{(n)}$ *be the largest order statistic. Then the ratio* $S_n/X_{(n)}$

converges in distribution to a random variable W with Laplace transform

$$\frac{e^{-\lambda}}{1+\alpha\int_0^1(1-e^{-\lambda t})t^{-\alpha-1}dt}, \quad \lambda \geq 0, \tag{2.9}$$

as $n \to \infty$. Furthermore, $\mathbb{E}(W) = (1-\alpha)^{-1}$ and $\mathrm{var}(W) = \alpha/[(2-\alpha)(1-\alpha)^2]$.

2.3 The α-stable law as $\alpha \to 0$

Letting the index parameter α tend to 0, we obtain the following result (previously derived by Cressie (1975) from the integral representation of the density function of Zolotarev (1986)).

Lemma 2.3.1. *Let $X \sim S(x;\alpha)$. Then as $\alpha \to 0$, $X^\alpha \xrightarrow{\mathcal{D}} 1/L$, where $L \sim \mathrm{Exp}(1)$.*

Proof. For $\alpha \in (0,1)$,

$$X^\alpha \stackrel{\mathcal{D}}{=} \left(\frac{a(U)}{L}\right)^{1-\alpha} = \left(\frac{1}{L}\right)^{1-\alpha} \times \big(\sin(\alpha U)\big)^\alpha \times \frac{\sin((1-\alpha)U)^{1-\alpha}}{\sin(U)}$$

Letting $\alpha \to 0$, $L^{\alpha-1} \to 1/L$, $\sin((1-\alpha)U)^{1-\alpha} \to \sin(U)$, and $\lim_{\alpha\to 0}\big(\sin(\alpha U)\big)^\alpha = 1$ by two applications of l'Hôpital's Rule. □

Figure 2.3 displays the empirical c.d.f. of L^{-1}, $L \sim \mathrm{Exp}(1)$, against that of X^α, $X \sim S(x;\alpha)$ with $\alpha \in \{0.05, 0.1, 0.25, 0.5, 0.75, 0.95\}$, showing that the rate of convergence as $\alpha \to 0$ is fast. The same conclusion is drawn from the comparison of quantiles of the two distributions in Figure 2.4.

We obtain a similar result for the stable law $F(x;\alpha,\beta,1,0)$, $\alpha \neq 1$; for proof, see Appendix A.

Lemma 2.3.2. *Let $X \sim F(x;\alpha,\beta,1,0)$, for $0 < \alpha \leq 2$, $\alpha \neq 1$, and $\beta \in [-1,1]$. Then as $\alpha \to 0$, $|X|^\alpha \xrightarrow{\mathcal{D}} 1/L$, where $L \sim \mathrm{Exp}(1)$.*

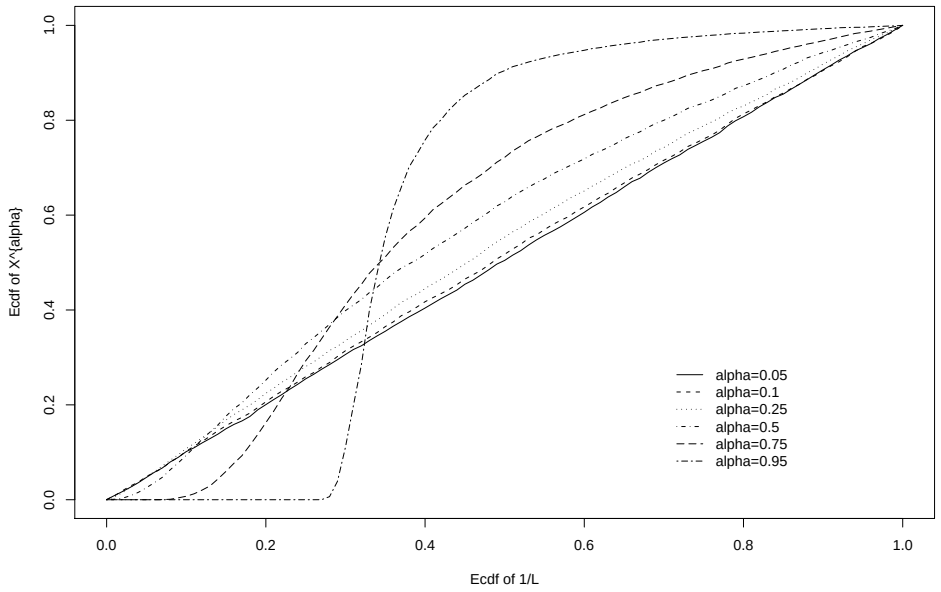


Figure 2.3: Comparison of empirical c.d.f. of X^α (vertical axis) against L^{-1} (horizontal axis), where $X \sim S(x;\alpha)$, $\alpha \in \{0.05, 0.1, 0.25, 0.5, 0.75, 0.95\}$, and $L \sim \text{Exp}(1)$.

The main result of this chapter states that the sum of n i.i.d. random variables with distribution $S(x;\alpha_n)$ is completely determined by the largest summand as $n \to \infty$, provided that $\alpha_n \in o(1/\log(n))$, i.e., $1/\log(n)$ grows much faster than α_n as $n \to \infty$. A one line proof is presented below and a more complicated one appears in Appendix A.

Theorem 2.3.1. *Let $X_i \sim S(x;\alpha_n)$, $i = 1, \ldots, n$, be i.i.d. Then $\left(X_{(n)}/S_n\right)^{\alpha_n} \xrightarrow{\mathcal{P}} 1$ as $n \to \infty$ provided that $\alpha_n \log(n) \to 0$ (condition A).*

Proof.

$$\left(\frac{S_n}{n}\right)^{\alpha_n} \leq X_{(n)}^{\alpha_n} \leq S_n^{\alpha_n},$$

so dividing both sides by $S_n^{\alpha_n}$ gives the inequality $n^{-\alpha_n} \leq \left(X_{(n)}/S_n\right)^{\alpha_n} \leq 1$, where the left hand side converges to 1 by condition A. □

Claim 2.3.1. *Under condition A, $\left(X_{(n)}\right)^{\alpha_n}/n \xrightarrow{\mathcal{D}} 1/L$, where $L \sim \text{Exp}(1)$.*

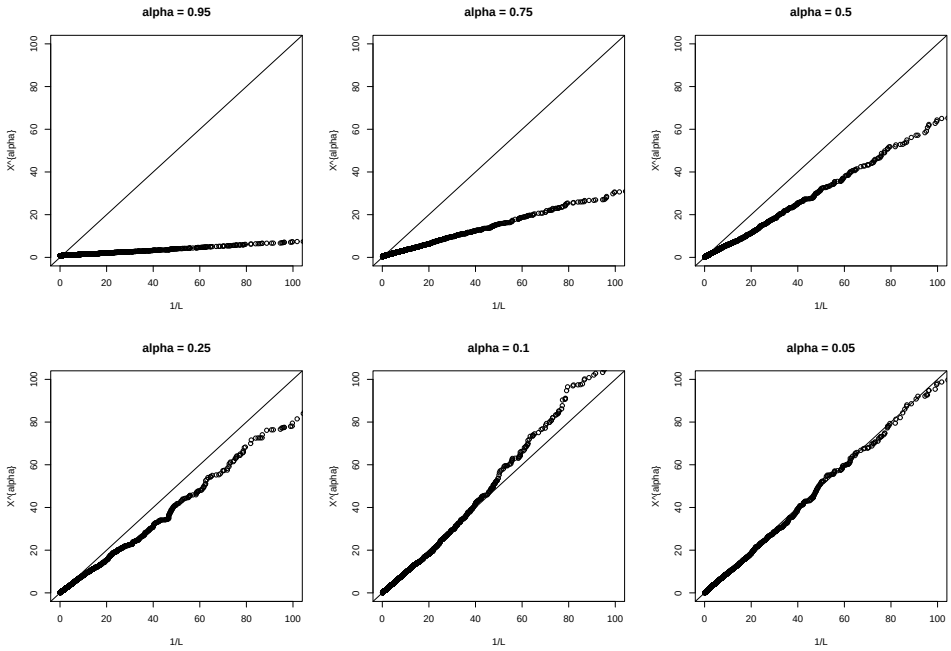


Figure 2.4: Comparison of QQ plots of X^α, $\alpha \in \{0.05, 0.1, 0.25, 0.5, 0.75, 0.95\}$ against L^{-1}, where $X \sim S(x;\alpha)$, and $L \sim \text{Exp}(1)$.

Proof. By Theorem 2.1.1, $S_n^{\alpha_n} \stackrel{\mathcal{D}}{=} nX^{\alpha_n}$, where $X \sim S(x;\alpha_n)$; it follows from Lemma 2.3.1 that $X^{\alpha_n} \stackrel{\mathcal{D}}{\to} L^{-1}$ as $n \to \infty$. By Theorem 2.3.1, $n^{-1}\big(X_{(n)}\big)^{\alpha_n} \stackrel{\mathcal{D}}{\to} L^{-1}$, provided condition A holds. □

2.4 Estimation of parameters of the stable law

Parameter estimation is particularly challenging due to the fact that the density function does not exist in closed form for most values of $\alpha \in (0, 2]$. For references on the extensively studied cases $\alpha = 1$ and $\alpha = 2$, see, for example, Li et al. (2007). DuMouchel (1973) shows that maximum likelihood estimators (MLEs) of the parameters, subject to the restriction $\alpha \geq \epsilon > 0$, for ϵ arbitrarily small, are both consistent and asymptotically normal, and DuMouchel (1975) computes estimates of the asymptotic standard deviations and correlations. Matsui and Takemura (2006) improve upon these estimates by providing accurate approximations

to the first and second derivatives of the stable densities. Nolan (2001) proposes an iterative approach to maximum likelihood estimation of the parameters, implemented in his software package STABLE, available at http://www.robustanalysis.com/. Furthermore, the package STABLE implements five additional methods for estimating stable parameters, including the empirical characteristic function method and fractional moments; see references therein.

The symmetric stable law ($\beta = 0$) has received great attention in the literature. Among the first estimators of the parameters γ and δ of the symmetric stable law with index $\alpha > 1$ are the semi-interquartile range, and the 0.5 truncated mean, respectively, proposed by Fama and Roll (1968). Fama and Roll (1971) analyse properties of the estimator of γ based on sample quantiles, and propose an estimator of α based on order statistics. McCulloch (1986) proposes simple, consistent, asymptotically normal estimators of the four stable parameters based on functions of five pre-determined sample quantiles. More recently, for $\alpha = 1$, $\beta = 0$, Besbeas and Morgan (2001) propose robust estimators of scale and location parameters based on the integrated squared error method with joint asymptotic relative efficiency of 96% compared to the maximum likelihood estimators. The following estimators of the scale parameter in the symmetric case are discussed in detail in Chapter 4: the geometric mean, and harmonic mean estimators (Li, 2008b), the fractional power estimator (Li and Hastie, 2008), and the optimal quantile estimator (Li, 2008a).

Press (1972) and Paulson et al. (1975) are the first to tackle the problem of parameter estimation for the skewed stable distribution by using the known form of the characteristic function. Feuerverger and McDunnough (1981a) and Feuerverger and McDunnough (1981b) provide consistent estimators based on the asymptotic distribution at r points, $r > 1$, of the empirical characteristic function of a random sample of fixed size. By sufficiently increasing r, the procedures attain arbitrarily high asymptotic efficiency; in numerical examples, the efficiency is estimated using the results of DuMouchel (1975). The latter estimating procedures are general and apply to the problem of parameter estimation for any distribution

having a well-defined characteristic function. The idea of employing the known form of the characteristic function is further developed by Koutrouvelis (1980), Koutrouvelis (1981), Kogon and Williams (1998), and Besbeas and Morgan (2008).

For estimating the scale parameter in the case $\alpha = 0.5$, Besbeas and Morgan (2004) derive a robust estimator based on the empirical Laplace transform having asymptotic relative efficiency of 94% compared to the maximum likelihood estimator. For the maximally skewed case ($\beta = 1$), the following estimators of the scale parameter are discussed in Chapter 4: the geometric mean, and harmonic mean estimators (Li, 2009), the fractional power estimator (Li, 2008c), and the optimal quantile estimator (Li, 2008d).

2.5 Estimation of density, distribution, and quantile functions in R

The R contributed package fBasics implements the method of Nolan (1997) for numerical calculations of density, distribution, and quantile functions. The method uses integral representations of the density and distribution functions derived by Zolotarev (1986), and employs adaptive quadrature for estimating the integrals.

For the symmetric strictly stable distribution, the c.d.f. for $x > 0$ and $\alpha \neq 1$ is given by

$$F(x;\alpha,0,1,0) = 1 - \frac{1}{4}\big(1+\epsilon(\alpha)\big) + \frac{\epsilon(\alpha)}{\pi}\int_0^{\pi/2} \exp\Big(-x^{\alpha/(\alpha-1)}U_\alpha(y,0)\Big)dy, \qquad (2.10)$$

where $\epsilon(\alpha) = \text{sign}(1-\alpha)$, and

$$U_\alpha(y,0) = \left(\frac{\sin(\alpha y)}{\cos y}\right)^{\frac{\alpha}{1-\alpha}} \times \frac{\cos(y(1-\alpha))}{\cos y}.$$

The p.d.f. is obtained by differentiating expression (2.10),

$$f(x;\alpha,0,1,0) = \frac{\alpha\epsilon(\alpha)}{\pi(\alpha-1)}x^{1/(\alpha-1)}\int_0^{\pi/2} U_\alpha(y,0)\exp\Big(-x^{\alpha/(\alpha-1)}U_\alpha(y,0)\Big)dy. \qquad (2.11)$$

The algorithms in fBasics work well for most values of α and β, but numerical difficulties are encountered for α very close to 0 and 1. In these situations, the integrand functions are highly peaked, so the adaptive quadrature algorithm may miss the spike, and underestimate the value of the integrals (Nolan, 1997). These estimation procedures are also included in the commercial software program STABLE; Nolan (1998) claims that the relative accuracy of the calculations performed by this program is 10^{-6}.

Matsui and Takemura (2006) provide reliable approximations to the first and second derivatives of the stable densities $f(x;\alpha,0,1,0)$ with respect to x, improving the approximation of Nolan (1997) to the symmetric density function at the boundary cases $x \to 0$, $x \to \infty$, and α close to 1. Matsui and Takemura (2006) avoid the unstable behaviour of the adaptive quadrature algorithm by using Taylor expansions of the density function. We write R functions to estimate the density $f(x;\alpha,0,1,0)$ (for $\alpha \geq 1$), and the first and second derivatives, $f'(x;\alpha,0,1,0)$ and $f''(x;\alpha,0,1,0)$, with respect to x (for $\alpha \geq 0.2$), following the method of Matsui and Takemura (2006) who claim that their numerical integration approach works well for most values of x and α.

Alternatively, we estimate $f'(x;\alpha,0,1,0)$ and $f''(x;\alpha,0,1,0)$ by a finite difference scheme with width $h = 0.01$, and approximate the density using *dstable* (fBasics); that is,

$$\begin{aligned} f''(x;\alpha,0,1,0) &\approx \frac{f'(x+h/2;\alpha,0) - f'(x-h/2;\alpha,0)}{h} \\ &\approx \frac{1}{h}\left[\frac{f(x+h;\alpha,0) - f(x;\alpha,0)}{h} - \frac{f(x;\alpha,0) - f(x-h;\alpha,0)}{h}\right] \\ &= \frac{1}{h^2}\big[f(x+h;\alpha,0) - 2f(x;\alpha,0) + f(x-h;\alpha,0)\big]. \end{aligned}$$

Figure 2.5 displays these approximations, with exact second derivative functions for $\alpha = 1, 2$. We obtain similar estimates using the expressions of Matsui and Takemura (2006).

We discover the numerical instability of the fBasics commands when computing the weights for the L-estimator of the location parameter based on a sample of transformed stable random variables, to be discussed in detail in Section 4.4.2. The weight function

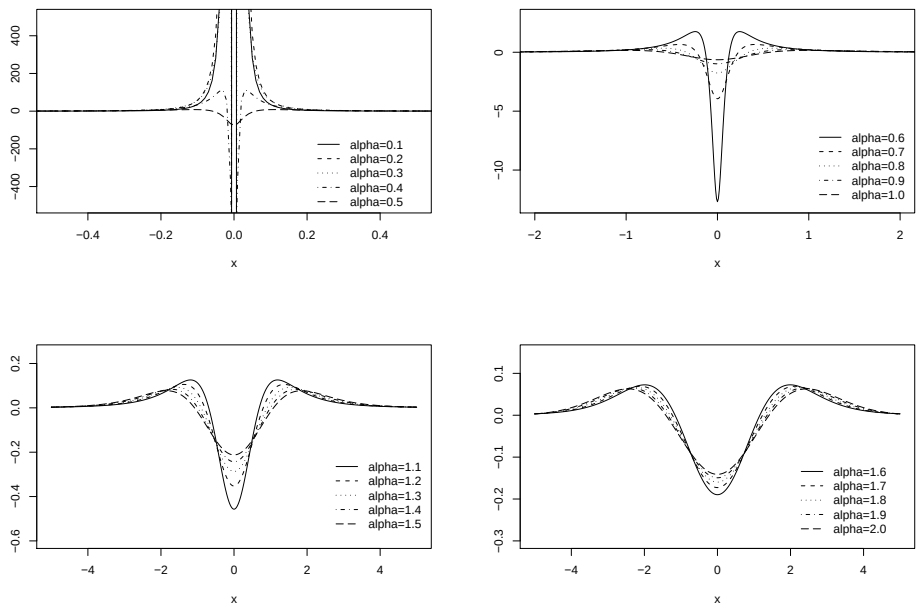


Figure 2.5: Approximations to $f''(x;\alpha,0,1,0)$ for $\alpha \in [0.1, 2]$ using a second order finite difference scheme.

involves the computation of the quantile function $F^{-1}((1+x)/2;\alpha,0,1,0)$, for $x \in (0,1)$ and $\alpha \in (0,2]$. Although we expected a smooth weight function, we observed small jumps in the weight function for x near 0, and $\alpha = 0.2, 0.3, 1$, which led us to investigate the behaviour of the quantile command *qstable* near 0.5.

Figure 2.6 (left) displays the 0.5001 (blue), and 0.501 (red) quantiles of the symmetric, strictly stable distribution computed with *qstable* (fBasics), for $\alpha \in [0.1, 2]$. We expect the quantile function to be a smooth, continuous, strictly increasing function of α, but remark that it is unstable for small values, and in the vicinity of $\alpha = 1$; the instability is particularly striking for the 0.5001 quantile around $\alpha = 1$. At $\alpha = 1.66$, the 0.501 quantile function jumps from 0.003483876 to 0.003544911, and remains constant for $\alpha > 1.66$. For $\alpha \in [1.3, 2]$, the 0.5001 quantile function is constant at 3.544908×10^{-4}, whereas we see in Figure 2.6 (right) that $F(3.544908 \times 10^{-4};\alpha,0,1,0) \geq 0.5001$ estimated with *pstable* (fBasics).

The function *qstable* returns an approximation x to the p-quantile by solving the equation

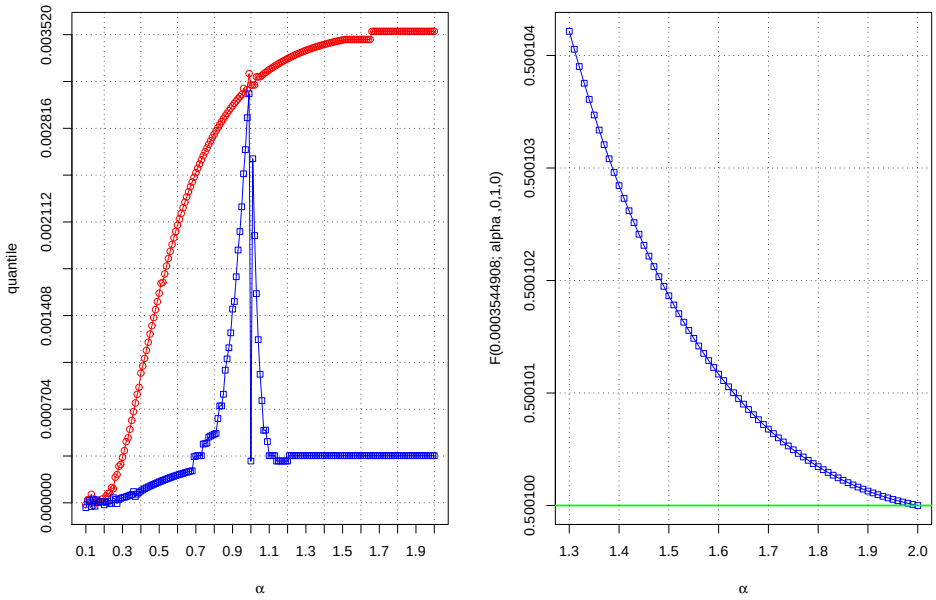


Figure 2.6: Left: Approximate $F^{-1}(x;\alpha,0,1,0)$ for $x = 0.5001$ (blue) and $x = 0.501$ (red), $\alpha \in [0.1, 2]$, computed using *qstable*. Right: Approximate $F(3.544908 \times 10^{-4};\alpha,0,1,0)$, $\alpha \in [1.3, 2]$, computed using *pstable*; the horizontal green line is drawn at 0.5001.

$F(x;\alpha,\beta,\gamma,\delta) - p = 0$ for x, calling *pstable* to estimate the c.d.f., and *uniroot* to find the root. The latter searches in steps, starting from a given interval, and enlarging it at every step until the exact root is found, or until the change in the estimate of the root found at consecutive steps falls below a specified tolerance level. Figure 2.7 displays the 0.5001 quantile function approximation for $\alpha \in [0.1, 2]$, returned by *qstable*; varying the tolerance parameter improves the approximation, particularly for $\alpha < 0.4$ and $\alpha \geq 1.5$.

The command *pstable* estimates the c.d.f. in (2.10) by invoking the command *integrate* to implement globally adaptive interval subdivision and extrapolation. *integrate* performs a series of subdivisions that adapt globally to the behaviour of the integrand. At each stage, the subinterval with largest error estimate is bisected, and a new extrapolated integral approximation with error estimate is produced. The process continues until the maximum number of subdivisions is reached, or until the error estimate falls below a specified tolerance

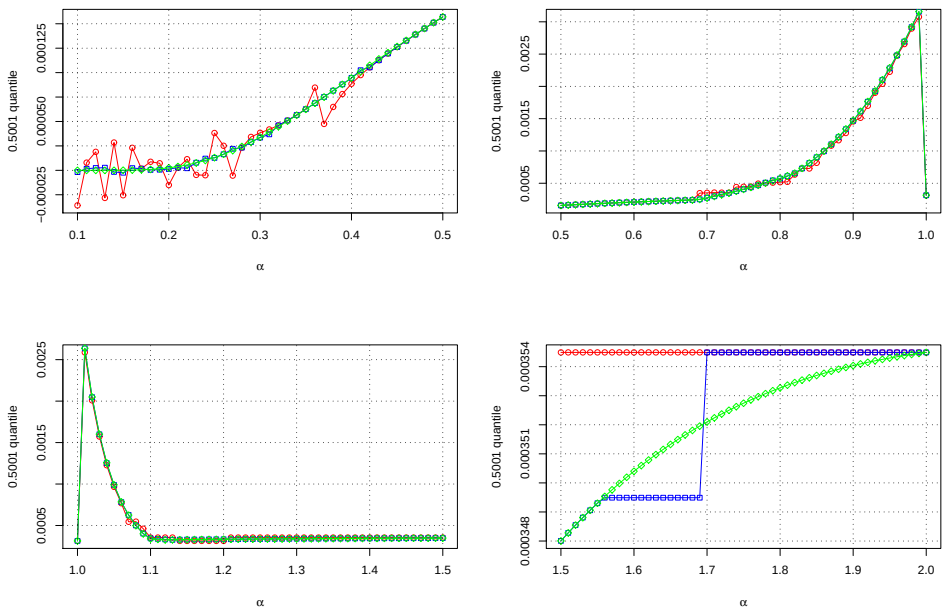


Figure 2.7: Approximate $F^{-1}(0.5001; \alpha, 0, 1, 0)$, $\alpha \in [0.1, 2]$, computed using *qstable* with tolerance levels: 1.220703×10^{-4} (red), 10^{-5} (blue), and 10^{-8} (green).

level; in the latter case, the corresponding integral approximation is returned. In contrast to quadrature rule that evaluates the integrand at closely spaced points covering the entire region of integration, adaptive subdivision may miss regions of significant mass where the integrand function changes very rapidly.

We consider two sources of instability: one arising from the behaviour of the integrand in the vicinity of $x = 0$ as $\alpha \to 1$, and $\alpha \to 0$, and one arising from the value of the tolerance parameter. Consider

$$x^{\alpha/(\alpha-1)} U_\alpha(y, 0) = \left(\frac{x \cos y}{\sin(\alpha y)} \right)^{\alpha/(\alpha-1)} \times \frac{\cos(y(1-\alpha))}{\cos y}.$$

We begin by analysing the behaviour of the integrand in the vicinity of $x = 0$ as $\alpha \to 1$. If y is away from the endpoints $y = 0$, and $y = \pi/2$, then

$$\frac{\cos(y(1-\alpha))}{\cos y} \to \frac{1}{\cos y} \in (1, \infty), \text{ as } \alpha \to 1,$$

and for x close to 0,

$$\frac{x\cos y}{\sin(\alpha y)} \approx 0.$$

So,

$$\left(\frac{x\cos y}{\sin(\alpha y)}\right)^{\alpha/(\alpha-1)} \to \begin{cases} 0 & \text{if } \alpha \to 1^+, \\ \infty & \text{if } \alpha \to 1^-, \end{cases}$$

and

$$\exp\Big(-x^{\alpha/(\alpha-1)}U_\alpha(y,0)\Big) \to \begin{cases} 1 & \text{if } \alpha \to 1^+, \\ 0 & \text{if } \alpha \to 1^-, \end{cases}$$

provided $x \approx 0$, and y is away from the endpoints $y = 0$ and $y = \pi/2$.

If $x \approx 0$ is fixed, and $y \to 0$, then the second and third terms in the product

$$\frac{x\cos y}{\sin(\alpha y)} = \frac{x}{\alpha y} \times \cos y \times \frac{\alpha y}{\sin(\alpha y)}$$

tend to 1, whereas the first one tends to ∞. Hence, letting $\alpha \to 1$, we obtain the following limiting behaviour

$$\exp\Big(-x^{\alpha/(\alpha-1)}U_\alpha(y,0)\Big) \to \begin{cases} 0 & \text{if } \alpha \to 1^+, \\ 1 & \text{if } \alpha \to 1^-. \end{cases}$$

Similarly, it can be shown that if $x \approx 0$ is fixed, and $y \to \pi/2$, then

$$\exp\Big(-x^{\alpha/(\alpha-1)}U_\alpha(y,0)\Big) \to \begin{cases} 1 & \text{if } \alpha \to 1^+, \\ 0 & \text{if } \alpha \to 1^-. \end{cases}$$

Approximate integration techniques based on adaptive subdivision of $(0, \pi/2)$ are likely to miss the sharp increase to 1 towards the endpoint $y = 0$ as $\alpha \to 1^-$, or the sharp decrease to 0 towards $y = 0$ as $\alpha \to 1^+$, evaluating the integral to 0 or $\pi/2$, respectively.

The command *dstable* for density estimation (see (2.11)) also displays a numerically unstable behaviour, e.g., it evaluates $f(0.001; 0.95, 0, 1, 0) = 0$, $f(0.01; 0.95, 0, 1, 0) = 0.325721$, $f(0.1; 0.95, 0, 1, 0) = 0.321945$, and returns an integration error for $f(0.001; 0.98, 0, 1, 0)$. The command appears highly numerically unstable for $\alpha \approx 1$, and for large values of x when $\alpha \in (1, 2]$; the latter situation is shown in Figure 2.8.

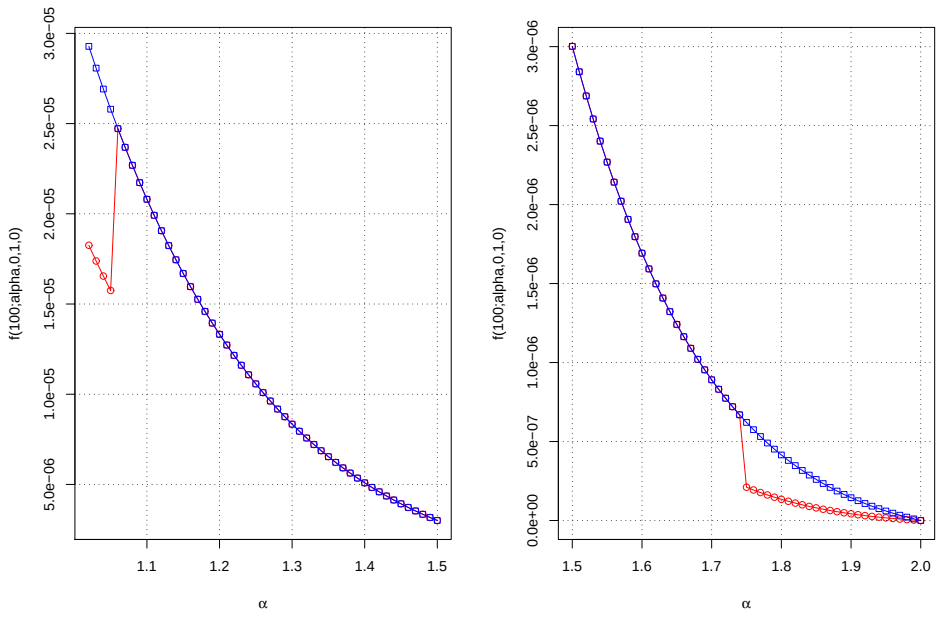


Figure 2.8: Approximate $f(100; \alpha, 0, 1, 0)$ for $\alpha \in [1.02, 2]$, obtained using *dstable* (red) and our algorithm with tolerance parameter 10^{-8} (blue).

2.6 Improved estimation of density, distribution, and quantile functions

To overcome the numerical instability in evaluating the integrals in (2.10) and (2.11), we propose an alternative method for estimating the density of the symmetric, strictly stable distribution ($\beta = 0$) based on the Fourier inversion formula (Feller, 1971).

For $\alpha \in (0, 2]$, and $\beta \in [-1, 1]$, Nolan (1999) uses the same method to evaluate the density function expressed in Zolotarev's parameterisation (M). He numerically evaluates the integral over regions of integration where the cosine term changes sign until the accuracy falls below a given tolerance level, and concludes that it is difficult to obtain accurate results when $\alpha < 0.75$, $\alpha \approx 1$ for $\beta \neq 0$, and when x is large.

When $\beta = 0$, the integration task is simpler, and our method corrects the numerical

instability of the fBasics commands. By the Fourier inversion formula,

$$\begin{aligned} f(x;\alpha,0,1,0) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi(t)dt \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}e^{-|t|^{\alpha}}dt \\ &= \frac{1}{2\pi}\left\{\int_{-\infty}^{\infty}\cos(tx)e^{-|t|^{\alpha}}dt - i\int_{-\infty}^{\infty}\sin(tx)e^{-|t|^{\alpha}}dt\right\} \\ &= \frac{1}{\pi}\int_{0}^{\infty}\cos(tx)e^{-t^{\alpha}}dt \\ &= \frac{1}{\pi}\sum_{i=1}^{\infty}\int_{(i-1)2\pi}^{i2\pi}\cos(tx)e^{-t^{\alpha}}dt, \end{aligned} \tag{2.12}$$

where the imaginary part disappears because the sine function is odd.

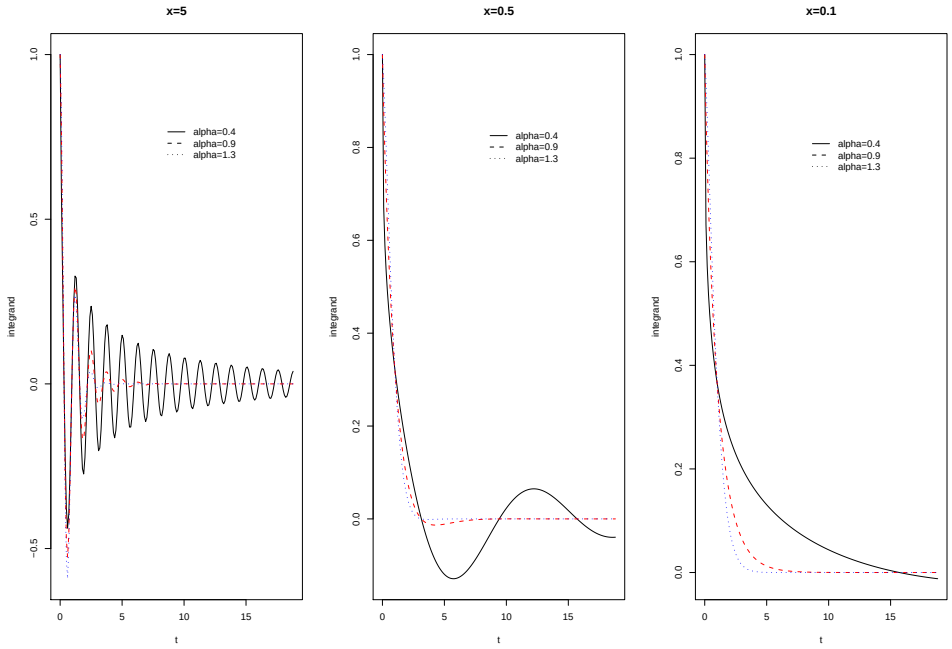


Figure 2.9: Plot of t versus $\cos(tx)e^{-t^{\alpha}}$ for $t \in (0, 30\pi)$, $x = 5, 0.5, 0.1$, and $\alpha = 0.4, 0.9, 1.3$.

Figure 2.9 shows the integrand in equation (2.12) as a function of t, for $t \in (0, 30\pi)$, and fixed x and α values. The cosine term in the integrand dominates (as shown by the case $x = 5$), and the smaller the value of α, the more undulating the integrand function is. As $x \to 0$, the integrand function tends to 0 quickly, and only the first few complete cycles of the cosine function make a significant contribution to the infinite summation in (2.12).

Next, we divide the integration bounds in (2.12) by x, for $x > 0$, improving the integration with respect to t as the function $\cos(tx)$ goes through one complete cycle for $t \in ((i-1)2\pi/x, i2\pi/x)$; this is particularly important when the product function is highly ondulating. However, dividing the computation bounds by x is not recommended when the integrand function decreases quickly to 0. In such instances, if x is large, then the integration bounds are very small, for small i, and many summands are required to obtain a good approximation to the density function.

The infinite sum in (2.12) is approximated by the first k terms, where k depends both α and x, and is chosen adaptively as explained below. For $x > 0$, we choose between the following approximations (note that if $x < 0$, $f(x;\alpha,0,1,0) = f(-x;\alpha,0,1,0)$):

$$f(x;\alpha,0,1,0) \approx \frac{1}{\pi}\sum_{i=1}^{k}\int_{(i-1)2\pi}^{i2\pi}\cos(tx)e^{-t^\alpha}dt, \tag{2.13}$$

and

$$f(x;\alpha,0,1,0) \approx \frac{1}{\pi}\sum_{i=1}^{k}\int_{(i-1)2\pi/x}^{i2\pi/x}\cos(tx)e^{-t^\alpha}dt.$$

The latter can be further simplified. Let $y = tx$.

$$f(x;\alpha,0,1,0) \approx \frac{1}{\pi}\sum_{i=1}^{k}\int_{(i-1)2\pi}^{i2\pi}\cos y e^{-(y/x)^\alpha}\frac{dy}{x},$$

and, make a second substitution $y = (i-1)2\pi + z$, giving

$$\begin{aligned} f(x;\alpha,0,1,0) &\approx \frac{1}{x\pi}\sum_{i=1}^{k}\int_0^{2\pi}\cos\big((i-1)2\pi+z\big)e^{-[x^{-1}((i-1)2\pi+z)]^\alpha}dz \\ &= \frac{1}{x\pi}\int_0^{2\pi}\cos(z)\sum_{i=1}^{k}e^{-[x^{-1}((i-1)2\pi+z)]^\alpha}dz \\ &= \frac{1}{x\pi}\int_0^{\pi}\cos(z)\left[\sum_{i=1}^{k}e^{-[\frac{1}{x}((i-1)2\pi+z)]^\alpha} - \sum_{i=1}^{k}e^{-[\frac{1}{x}((i-1)2\pi+z+\pi)]^\alpha}\right]dz \end{aligned} \tag{2.14}$$

To choose between (2.13) and (2.14), consider the product function $\cos(tx)\times e^{-t^\alpha}$ when $t = \pi/(2x)$, i.e., the cosine term first takes on the value 0. If $e^{-(\pi/(2x))^\alpha} \le 0.2$ and $x > 10$,

then the exponential term decreases quickly towards 0 and dominates the cosine term, so use expression (2.13); otherwise use expression (2.14).

To determine k, we add terms to the summation until the contribution falls below a prespecified threshold level that depends α. For expression (2.13), we start with $k_0 = 100$, and at step $j \geq 1$, let $k_j = k_{j-1} + 100$; if $|\sum_{i=k_{j-1}+1}^{k_j} \int_{(i-1)2\pi}^{i2\pi} \cos(tx)e^{-t^\alpha} dt| < \epsilon$, stop and return the approximation to the density with k_j terms, else continue. For expression (2.14), the procedure for determining k is analogous, with $k_0 = 5$ and $k_j = k_{j-1} + 10$. The difference in these two procedures rests with the starting value k_0 and the fact that in the former we add 100 terms at each step, while in the latter only 10; this is explained by the fact that expression (2.13) is used when the product term is mildly ondulating, so many terms must be considered at a given step to notice a significant contribution towards the total sum, whereas in the latter the product term is highly ondulating.

Integrating expression (2.12), we obtain for $x > 0$

$$\begin{aligned} F(x;\alpha,0,1,0) &= \frac{1}{2} + \frac{1}{\pi}\int_0^x \int_0^\infty \cos(tu)e^{-t^\alpha} dtdu \\ &= \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \frac{1}{t}\sin(tx)e^{-t^\alpha} dt \\ &= \frac{1}{2} + \frac{1}{\pi}\sum_{i=1}^{\infty}\int_{(i-1)2\pi}^{i2\pi} \frac{1}{t}\sin(tx)e^{-t^\alpha} dt. \end{aligned} \tag{2.15}$$

The approximations are

$$F(x;\alpha,0,1,0) \approx \frac{1}{2} + \frac{1}{\pi}\sum_{i=1}^{k}\int_{(i-1)2\pi}^{i2\pi} \frac{1}{t}\sin(tx)e^{-t^\alpha} dt, \tag{2.16}$$

and

$$F(x;\alpha,0,1,0) \approx \frac{1}{2} + \frac{1}{\pi}\int_0^\pi \sin(z)\left[\sum_{i=1}^{k} \frac{e^{-[x^{-1}((i-1)2\pi+z)]^\alpha}}{(i-1)2\pi+z} - \sum_{i=1}^{k} \frac{e^{-[x^{-1}((i-1)2\pi+z+\pi)]^\alpha}}{(i-1)2\pi+z+\pi}\right] dz, \tag{2.17}$$

where the latter is obtained by dividing the integration bounds by x when the product term $t^{-1}\sin(tx)e^{-t^\alpha}$ is highly ondulating. Since $\sin(t)/t \to 0$ as $t \to \infty$, the product term

in (2.15) has a faster rate of convergence to 0 than the product term in (2.12), and the integrand function in the former case is less ondulating than in the latter. Hence, we use expression (2.16) if $e^{-(\pi/x)^\alpha} \leq 0.00001$, and (2.17) otherwise; we determine k as previously.

We write improved versions for the fBasics commands *dstable*, *pstable*, and *qstable*, using the expressions via finite summations for density and distribution functions to correct the detected numerical instability. The tolerance parameter of the root finding algorithm in *qstable* is set to 10^{-8}, and the number of subdivisions in the integration procedure is left at the default value of 1000 (varying this number did not improve the approximations). Set the threshold parameter ϵ (determining k) as follows: $\epsilon = 10^{-9}$ for $\alpha < 0.2$ and $\alpha \in (1.0, 1.2)$, $\epsilon = 10^{-7}$ for $\alpha \in [0.2, 0.4)$ and $\alpha \in [1.2, 1.5)$, and $\epsilon = 10^{-6}$ for all other values of α. For $x \in [0, 1500)$, these ϵ values are small enough to give good approximations.

For comparison, we estimate the 0.5001 and 0.501 quantiles for various values of α, plotted in blue in Figures 2.10 and 2.11; the instability for small α and in the vicinity of $\alpha = 1$ is completely removed by our algorithm. For x close to 0, the behaviour of our algorithms appears numerically unstable in the range $\alpha < 0.2$; those of Matsui and Takemura (2006), and Nolan (2001) suffer from the same drawback.

2.7 Summary

This chapter introduces the stable distribution, presenting properties and parameterisations, as well as algorithms for simulating from this distribution. In particular, it discusses the positive, strictly stable distribution, and the α-stable distribution as $\alpha \to 0$. Moreover, a novel result on sums of i.i.d. positive, strictly stable random variables is presented, and will be used subsequently in Section 3.4.2 to prove an unexpected connection between data sketching and hashing. Finally, we correct a numerical instability in the fBasics commands for density, distribution, and quantile estimation of the symmetric, stable distribution.

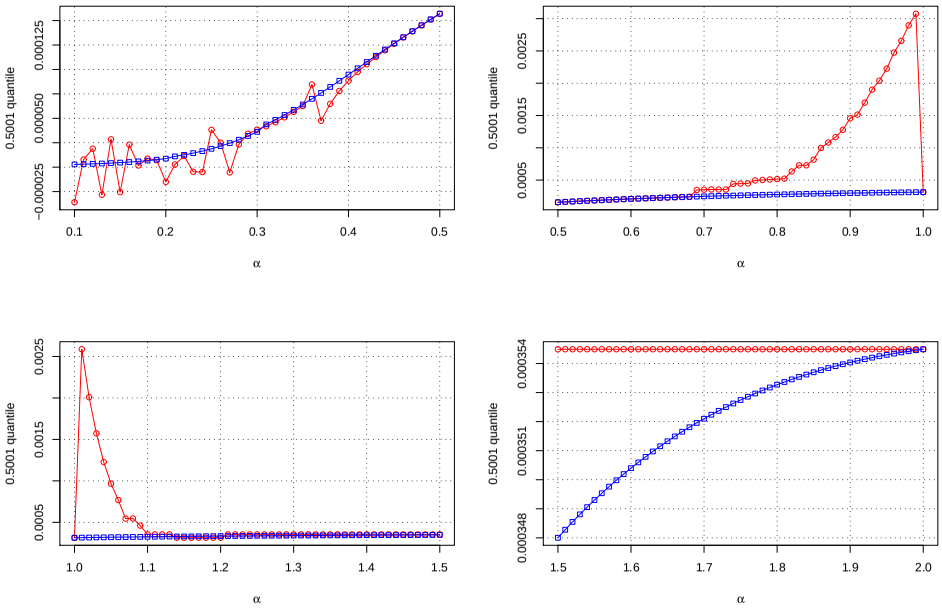


Figure 2.10: Approximate $F^{-1}(0.5001; \alpha, 0, 1, 0)$ for $\alpha \in [0.1, 2]$ computed by *qstable* (red), and our algorithm (blue).

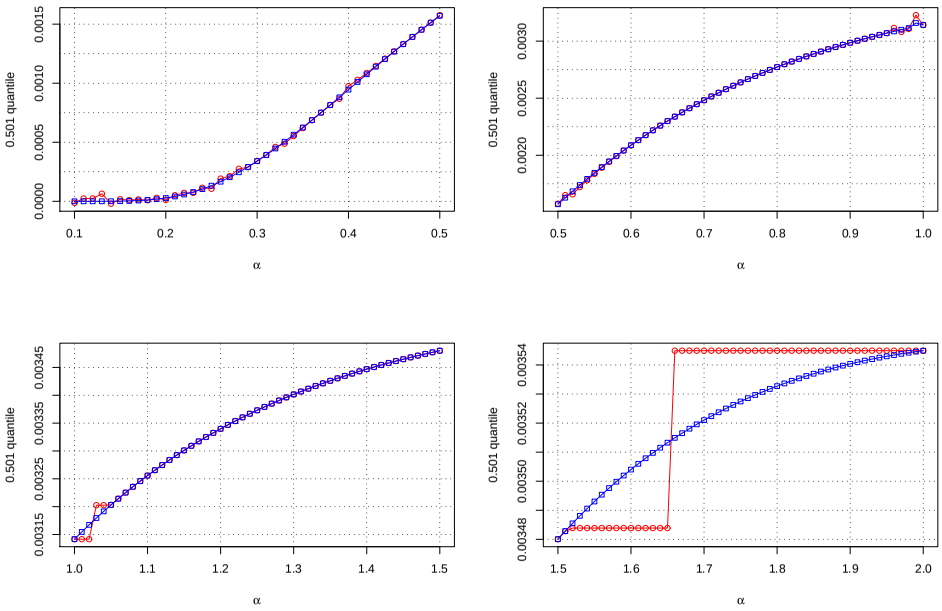


Figure 2.11: Approximate $F^{-1}(0.501; \alpha, 0, 1, 0)$ for $\alpha \in [0.1, 2]$, computed by *qstable* (red), and our algorithm (blue).

Chapter 3

Cardinality estimation in streaming data

The focus of this chapter is cardinality estimation in streaming data, i.e., counting the number of distinct elements observed, via fast algorithms that require space much smaller than the order of the cardinality, where the latter is assumed to be prohibitively large in real-life data streams. We distinguish between two methods: hashing and storing order statistics, and data sketching via projections to stable random variables. In both cases, we derive recursively computable, maximum likelihood estimators of the cardinality. We prove an unsuspected link between the two methods in Theorem 3.4.1.

3.1 Data streams

Observations in stream format present new challenges in the statistical analysis of massive data sets. Exact computation of simple properties of extensive streaming data, such as Hamming norm, l_α norm, and αth frequency moments, becomes infeasible in real time (Cormode and Muthukrishnan, 2005a). Hence, the need arises to approximate these values, and in order for the approximation to be meaningful, it must accompanied by a measure of

uncertainty. Let s denote a property of interest with approximation $\hat{s}$.

Definition 3.1.1. *$\hat{s}$ is an (ϵ, δ)-approximation to s, for some $\epsilon, \delta > 0$ arbitrarily small, if $\mathbb{P}\left(|\hat{s} - s| > \epsilon s\right) \leq \delta$. ϵ is the approximation parameter, and $1 - \delta$ the confidence parameter.*

Consider a data stream with elements drawn from a countable, possible infinite set of values $\mathcal{D}$. At the current time point t, a pair of the form (i_t, d_t) is observed, where $i_t \in \mathcal{D}$ is the type of data element, and d_t is the quantity. For certain kinds of data, it is possible for data elements to cancel each other out; for example, when an input (i_t, d_t) is followed by (i_s, d_s) such that $d_s = -d_t$ and $i_t = i_s$. This would be the case when (i_t, d_t) is a record of purchase or sale, and the objective is to monitor stock levels. It is assumed that data in stream format can come in an arbitrary order, and that the same element type can appear several times in the stream.

Definition 3.1.2. *The accumulation vector of a data stream over $\mathcal{D}$ observed up to finite time $T > 0$ is defined to be the vector $\mathbf{a}_T = \left(a_1, \ldots, a_i, \ldots\right)$, where $a_i = \sum_{t=1}^{T} d_t \mathbb{I}(i_t = i)$ is the cumulative quantity of elements of type i at time T, for $i \in \mathcal{D}$.*

Note that we are assuming that the types can be ordered by some convention, for example by the order of their first appearance. Following the terminology of Cormode et al. (2003), when $d_t > 0 \; \forall t > 0$ we refer to this as the *cash register* case, and when $d_t \in \mathbb{R}$ as the *turnstile* case. The latter is further divided into the *general* case when $a_i \in \mathbb{R}$, and the *non-negative* case when $a_i \geq 0 \; \forall i$. When $d_t = 1 \; \forall t$, we will call this a *simple data stream.*

Definition 3.1.3. *The Hamming norm of a vector $\mathbf{a} = \mathbf{a}_T$ over $\mathcal{D}$ is defined as*

$$|\mathbf{a}|_H = \left|\{i; a_i \neq 0, i \in \mathcal{D}\}\right|.$$

Definition 3.1.4. *The Hamming distance between two vectors $\mathbf{a} = \mathbf{a}_{T_1}$ and $\mathbf{b} = \mathbf{b}_{T_2}$ over $\mathcal{D}$ is defined as the Hamming norm of $\mathbf{a} - \mathbf{b}$, $|\mathbf{a} - \mathbf{b}|_H = \left|\{i; a_i - b_i \neq 0, i \in \mathcal{D}\}\right|$.*

Definition 3.1.5. *The l_α norm of a vector $\mathbf{a} = \mathbf{a}_T$ over $\mathcal{D}$ is defined for $\alpha \geq 1$ by*

$$l_\alpha(\mathbf{a}) = \Big(\sum_{i \in \mathcal{D}} |a_i|^\alpha \Big)^{1/\alpha}.$$

For $\alpha < 1$, it is called a quasi-norm because the triangle inequality fails to hold.

Definition 3.1.6. *The l_α distance between vectors $\mathbf{a} = \mathbf{a}_{T_1}$ and $\mathbf{b} = \mathbf{b}_{T_2}$ over $\mathcal{D}$ is defined for $\alpha \geq 1$ by*

$$d_\alpha(\mathbf{a}, \mathbf{b}) = \Big(\sum_{i \in \mathcal{D}} |a_i - b_i|^\alpha \Big)^{1/\alpha}.$$

For $\alpha < 1$, it is called a quasi-distance.

Definition 3.1.7. *The αth frequency moment of a vector $\mathbf{a} = \mathbf{a}_T$ is given by $F_\alpha(\mathbf{a}) = \big(l_\alpha(\mathbf{a})\big)^\alpha$. If $\alpha = 0$, we define $|a_i|^\alpha = 1$ if $a_i \neq 0$ and 0 otherwise.*

The statistic $F_1(\mathbf{a})$ is the cumulative quantity of all types in the data set, and the statistics $F_\alpha(\mathbf{a})$ for $\alpha \geq 2$ indicate the degree of non-uniformity of the data, where $F_2(\mathbf{a})$ is known as the *repeat rate* or *Gini's index of homogeneity*.

Definition 3.1.8. *The cardinality of a general data stream with accumulation vector $\mathbf{a} = \mathbf{a}_T$ is defined as the limit of $F_\alpha(\mathbf{a})$ as $\alpha \to 0$, and equals the Hamming norm $|\mathbf{a}|_H$.*

Referring to Cormode and Muthukrishnan (2005a), other quantities of interest in data sets are: the cumulative quantity of type i, a_i, the cumulative quantity of type i for $l \leq i \leq r$, $\sum_{i=l}^{r} a_i$, where $l, r \in \mathcal{D}$ such that $l < r$, and the inner product of the cumulative quantities of two streams $\mathbf{a} = \mathbf{a}_{T_1}$ and $\mathbf{b} = \mathbf{b}_{T_2}$, defined by $\mathbf{a} \odot \mathbf{b} := \sum_{i \in D} a_i b_i$. In the non-negative case, i.e., $a_i \geq 0 \ \forall i$, quantities of interest in summarising the distribution of the observed data approximately are ϕ-quantiles and ϕ-heavy hitters, defined for $0 \leq \phi \leq 1$; refer to Cormode and Muthukrishnan (2005a) for definitions. In the sequel, we will drop the subscript T and write $\mathbf{a}$ for $\mathbf{a}_T$.

Two radically different approaches exist for estimating the value of a given property: the sampling approach and the random record-keeping approach. In the former, the estimation

is based on information contained in a random sample of the elements in the data set, while in the latter, all elements of the set are inspected but only partial information is retained. Sampling-based algorithms for cardinality estimation have been investigated in the context of estimating the number of distinct species or classes in a population (Bunge and Fitzpatrick, 1993; Haas et al., 1995). The difficulty with these algorithms is that unless a specific model of the population is assumed *a priori*, there is no guarantee that the random sample is large enough to offer a good representation of the underlying population. Further difficulties arise in devising sampling strategies when data is acquired in stream format; this is particularly acute when the elements arrive in adversarial order, i.e., the pattern of repetitions is far from uniform. Random-record keeping algorithms proceed by scanning the entire data set and retaining only partial information. We identify in the literature two different methods of extracting meaningful information from observed data: hashing and stable law sketching.

Hashing is a basic tool used by many algorithms for handling data where the type of data element is identified by a complicated label. Hash functions were initially used for storing data in tables since data indexed by integer values can be more readily accessed; the hash function, h, maps element $i \in \mathcal{D}$ into slot $h(i)$ of the table. The function h is deterministic in the sense that given $i \in \mathcal{D}$, the same output $h(i)$ is produced, but h is constructed so that it has pseudo-random properties.

Definition 3.1.9. *A hash function $h : \mathcal{D} \mapsto \{1, \ldots, L\}$, where $L \ll |\mathcal{D}|$, is a deterministic, pseudo-random function with the property that $h(i) \sim \text{Unif}(\{1, \ldots, L\})$ independently.*

Since $L \ll |\mathcal{D}|$, there is a non-zero probability that for some $i, j \in \mathcal{D}, i \neq j$, the function produces $h(i) = h(j)$; this situation is called a collision. A good hash function is a pseudo-random function that maps uniformly and has low collision rate. To minimise the probability of collisions, an approach known as universal hashing is oftentimes implemented. The hash function h is chosen randomly from a finite collection $\mathcal{H}$ of functions satisfying the following property: the probability of the event $\left[h(i) = h(j), i \neq j\right]$ is at most $1/L$. In other words,

the probability of collision is at most equal to the probability of $\big[h(i) = h(j)\big]$ occurring if $h(i)$ and $h(j)$ were randomly chosen from the set $\{1, \ldots, L\}$. In the following, we assume that a hash function mapping to random, independent values can be adequately emulated by random number generators via the method of seeding, to be presented in Section 3.3. For a discussion on random number generators and hash functions, refer to Knuth (1998).

3.2 Cardinality estimation: existing methodology

The most obvious way of counting the number of types in a simple data stream, denoted by c, is as follows. Allocate $|\mathcal{D}|$ separate, unoccupied bins. For each distinct type $i \in \mathcal{D}$ observed, change the status of the corresponding bin from unoccupied to occupied (this can be achieved via a bit variable that switches from 0 to 1), and then count the number of occupied (or unoccupied) bins after the data stream has been processed. The disadvantage of this approach is two-fold: first, prior knowledge on the value of $|\mathcal{D}|$ is needed, and second, if this value is large, the number of bins required may exceed available computer storage. The simplest solution is to reduce the number of bins to L, where $L \ll c$, and to map distinct types to bins at random via a hash function. Care, however, must be taken that L is large enough such that the probability of occupying all bins is negligible. The algorithm is as follows.

Let h be a hash function mapping from $\mathcal{D}$ to uniformly distributed, independent values in $\{1, \ldots, L\}$. Allocate a vector v of length L with entries initially all equal to 0. At time t, observe (i_t, d_t), compute the hashed value $x = h(i_t)$, and set $v_x = 1$. Once the entire data stream has been observed, compute $\hat{N} = |\{v_x = 0,\ 1 \leq x \leq L\}|$, the number of remaining entries equal to zero in the vector v. The probability of $[v_x = 0]$ is the probability that none of the c distinct data elements hash to the value x, for $1 \leq x \leq L$, which equals $(1 - 1/L)^c$. As a random variable, N follows approximately a Binomial$\big(L, (1 - 1/L)^c\big)$ distribution with expected value $\mathbb{E}(N) = L(1 - 1/L)^c$. (Note that N cannot take the value 0.) By the

method of moments, estimate $\mathbb{E}(N)$ by $\hat{N}$, and solve for c to obtain the following estimator: $\hat{c} = \log(\hat{N}/L)/\log(1 - 1/L)$. The variance of $\hat{c}$ is approximately $Le^{c/L}$ for c large, obtained by expanding the function $\log(N/L)$ into a Taylor series about $\mathbb{E}(N)$ and truncating after the second term. In practical applications, when c is very large, L would need to be of order $c/\log c$ to produce a standard deviation of order c, i.e., storage of size $c/\log c$ needs to be allocated in order to use this algorithm. Hence, the storage reduction is relatively small and the method is impractical when c is large.

3.2.1 Probabilistic counting

One of the earliest suggestions for storage reduction in cardinality estimation of simple data streams is the probabilistic counting algorithm of Flajolet and Martin (1985). At time t, the pair (i_t, d_t) is observed, and the data element i_t is hashed to a random, independent, uniformly distributed number over a finite range (Flajolet and Martin, 1985), or equivalently to the interval $[0, 1]$ (Durand and Flajolet, 2003; Flajolet, 2004; Giroire, 2005). Let $\rho(i_t)$ be the position of the first 1 in the binary representation of $h(i_t)$. Flajolet and Martin (1985) propose the asymptotically unbiased estimator $\hat{c} = 2^R/\phi$, where $R = \max\left\{k : [1, \ldots, k] \subseteq \{\rho(i_t), t = 1, \ldots, T\}\right\}$, and ϕ is an appropriate constant; similar in spirit is Wegman's Adaptive Sampling algorithm (Flajolet, 1990). The estimator of Durand and Flajolet (2003) is a function of the statistic $\max\{\rho(i_t), t = 1, \ldots, T\}$, while Giroire (2005) derives estimators that are functions of order statistics of the hashed values.

To increase the precision of the estimates, the method of *stochastic averaging* (Flajolet and Martin, 1985) is employed. The interval $[0, 1]$ is divided into m distinct subintervals of size $1/m$, called buckets, and an estimate of c is obtained for each bucket. In particular, Giroire (2005) computes the kth order statistic of the hashed values for fixed k, denoted by $Y_i^{(k)}$ for the ith bucket. We point out that the resulting estimators are not well-defined; if k is large compared to c/m, then the kth order statistic of values in a given bucket may not

exist. Giroire (2005) proposes three families of estimators, the inverse family, ξ_1, the square root family, ξ_2, and the logarithm family, ξ_3, based on the sample $Y_1^{(k)} = y_1, \ldots, Y_m^{(k)} = y_m$.

$$\begin{aligned}
\xi_1 &:= (k-1)\sum_{i=1}^{m}\frac{1}{y_i} \\
\xi_2 &:= \frac{1}{\left(\frac{1}{(k-1)!}\Gamma(k-1) + \frac{m-1}{(k-1)!^2}\Gamma(k-1/2)^2\right)}\left(\sum_{i=1}^{m}\frac{1}{\sqrt{y_i}}\right)^2 \\
\xi_3 &:= m\cdot\left(\frac{\Gamma(k-1/m)}{\Gamma(k)}\right)^{-m}\cdot\exp\left\{-\frac{1}{m}\sum_{i=1}^{m}\ln y_i\right\}
\end{aligned}$$

As $c \to \infty$, the estimators are approximately asymptotically unbiased, with standard deviations (s.d.):

$$\begin{aligned}
s.d.(\xi_1) &\approx \frac{c}{\sqrt{k-2}}\cdot\frac{1}{\sqrt{m}} \\
s.d.(\xi_2) &\approx \frac{2c}{\sqrt{m}}\sqrt{\frac{1}{k-1}\left(\frac{\Gamma(k)}{\Gamma(k-1/2)}\right)^2 - 1}, \text{ for } m \text{ large} \\
s.d.(\xi_3) &\approx \sqrt{\Psi'(k)}\cdot\frac{c}{\sqrt{m}}, \text{ for } m \text{ large},
\end{aligned}$$

where Γ and Ψ denote the gamma and digamma functions with $\Psi(z) = d\ln\Gamma(z)/dz$.

Alon et al. (1999) modify the algorithm of Flajolet and Martin by requiring only pairwise independent hash values, uniform over a finite range. This seminal paper is the first attempt at obtaining tight lower bounds on the space complexity of approximating frequency moments $F_\alpha(\mathbf{a})$. For simple stream data, they show that for every $\alpha \geq 1$, $\epsilon > 0$, and $\delta > 0$, there exists a randomized algorithm that computes an (ϵ, δ)-approximation to $F_\alpha(\mathbf{a})$ in one pass over the data set and using $O\big(\alpha\epsilon^{-2}m^{1-1/\alpha}(\log m + \log T)\log(\delta^{-1})\big)$ memory bits, where $m = |\mathcal{D}|$ and T is the length of the data stream. Moreover, they prove that $O\big(\log m\big)$ bits suffice for estimating the cardinality of a simple data stream.

Bar Youssef et al. (2002) present three algorithms for cardinality estimation with different space/time tradeoffs that require pairwise independent hash functions and account for the probability of collisions, that previous algorithms based on hashing dismissed as negligible.

They obtain the best (ϵ, δ)-approximation for the cardinality of a simple data stream in terms of space requirements and processing time per element. The space requirement is of the order of $O\big(\big(\epsilon^{-2}\log\log m + \log m \log(\epsilon^{-1})\big)\log(\delta^{-1})\big)$. Indyk and Woodruff (2003) show that the dependence of the space requirement on ϵ through the factor $1/\epsilon^2$ cannot be reduced to $1/\epsilon$. In fact, they prove that any algorithm that (ϵ, δ)-approximates the cardinality must use $\Omega(1/\epsilon^2)$ space, provided that $\epsilon = \Omega\big(m^{-1/(9+k)}\big)$, for any $k > 0$.

3.2.2 Data sketching

Indyk (2006) (previously published in Indyk (2000)) introduces the expression *data sketching*; the idea is to represent important features of the data using small amount of memory.

Definition 3.2.1. *Let $F(S)$ be a given statistic of $S \in \mathbb{R}^d$. $C(S) \in \mathbb{R}^k$ is a good sketch representation of S for estimating $F(S)$ if $k \ll d$ (i.e., the sketch requires less space to store than the original vector), and $F(S)$ can be accurately approximated by applying some function G to $C(S)$.*

Indyk (2006) claims that "computing sketches of normed vectors enables to compress the data and speed-up computation". Cormode and Muthukrishnan (2005a) point out that "most sketches described in the literature are good for one single, pre-specified aggregate computation". Indyk (2006) proposes an algorithm for estimating the l_α norm of a data stream for $\alpha = 1, 2$ by constructing low-dimensional representations of the data that exploit properties of the stable law.

Cormode et al. (2003) expand the algorithm to $\alpha \in (0, 2]$ for the general case of stream data. For the problem of cardinality estimation in the turnstile case, their estimator is based on the median of weighted sums of symmetric, strictly stable random variables with index $\alpha = 0.02$. Data sketching methods require space logarithmic in the cardinality of the stream and work under the underlying assumption that this cardinality is extremely large, else it would be possible to store the vector of observed frequencies in main computer memory.

Furthermore, data sketching methods that store linear combinations of α-stable random variables also work in the presence of deletions, i.e., when a negative number of elements of a particular type is observed; other synopsis construction methods, e.g., sampling, fail in such instances.

Ganguly (2007) combines the methods of hashing and binning for estimating cardinality in the non-negative, turnstile case. His algorithms, which use d-wise independent random hash functions, improve upon the time and space complexity of existing algorithms for this problem. Ganguly (2004) presents an algorithm for estimating the αth frequency moments for $\alpha > 2$ in the non-negative, turnstile case, via sketching based on linear combinations of randomly chosen αth roots of unity. He reduces the space complexity obtained by Alon et al. (1999) to $O\big(2m^{1-1/(\alpha-1)} \log m\big)$ memory bits, which is further improved upon by Indyk and Woodruff (2005) who reduce the leading factor to $m^{1-2/\alpha}$, for $\alpha > 2$.

3.3 Cardinality estimation via hashing

In this section we propose maximum likelihood estimators of the cardinality based on hashing to continous or discrete random variables, and storing order statistics. We discuss the property of recursive computability, and the requirement of attaining the tight lower bound on storage complexity. Moreover, we show that the estimators of Giroire (2005) are asymptotically efficient.

3.3.1 Hashing to continuous random variables

Suppose that data arrives in simple stream format: $(i_t, 1)$; the data type i_t is mapped to a random variable from the uniform distribution on (0,1) via m independent hash functions, concurrently applied to the data stream. The result is m random samples of size c, where c denotes the cardinality of the data stream, from the uniform distribution on (0,1); we now

apply conventional maximum likelihood estimation to the problem of approximating c.

Let $Y_i^{(k)}$ denote the kth order statistic from the ith sample, $i = 1, \ldots, m$. Assuming that the first k order statistics $Y_i^{(1)} = y_1, \ldots, Y_i^{(k)} = y_k$ are available for each sample i, $i = 1, \ldots, m$, the joint likelihoood of sample i is

$$L(c; y_1, \ldots, y_k) = \frac{c!}{(c-k)!}(1-y_k)^{c-k}\mathbb{I}[y_1 < y_2 < \ldots < y_k].$$

By the Neyman Factorisation Criterion (Lehmann, 1983), $Y_i^{(k)}$ is sufficient for c. Therefore we derive our maximum likelihood estimator of c, denoted by $\hat{c}_C$, from $Y_1^{(k)} = y_1, \ldots, Y_m^{(k)} = y_m$, a random sample of size m of kth order statistics. The likelihood function is

$$L(c; y_1, \ldots, y_m) = \prod_{i=1}^{m} \left\{ \frac{c!}{(c-k)!(k-1)!} y_i^{k-1}(1-y_i)^{c-k} \right\},$$

and $\hat{c}_C$ satisfies

$$\left.\frac{\partial l}{\partial c}\right|_{c=\hat{c}_C} = \left.\frac{\partial \log(L)}{\partial c}\right|_{c=\hat{c}_C} = \sum_{i=1}^{m} \log(1-y_i) + \frac{m}{\hat{c}_C} + \frac{m}{\hat{c}_C - 1} + \ldots + \frac{m}{\hat{c}_C - k + 1} = 0. \quad (3.1)$$

From the definition of Euler's constant $\gamma := \lim_{n\to\infty} \left\{ \sum_{i=1}^{n} 1/i - \log(n) \right\}$, we have the approximation $\sum_{i=1}^{n} 1/i = \log(n) + \gamma_n$, where $\gamma_n \to \gamma$ as $n \to \infty$. So for large c and $k \neq 1, c$, equation (3.1) becomes $m^{-1}\sum_{i=1}^{m} \log(1-y_i) + \log(\hat{c}_C) - \log(\hat{c}_C - k) \approx 0$, and the MLE is approximately given by

$$\hat{c}_C = \frac{k}{1 - \prod_{i=1}^{m}(1-y_i)^{1/m}}. \quad (3.2)$$

Moreover,

$$-\frac{\partial^2 l}{\partial c^2} = \frac{m}{c^2} + \frac{m}{(c-1)^2} + \ldots + \frac{m}{(c-k+1)^2} \approx \frac{mk}{c(c-k)},$$

so the Fisher information (Cramér, 1946) about c contained in the sample is approximately mk/c^2 for large c; hence, $\hat{c}_C \sim \text{Normal}(c, c^2/(mk))$ approximately for large c as $m \to \infty$.

Let F denote an arbitrary continuous distribution, and suppose that the hash functions map to random variables from F. Then the log likelihood function of the kth order statistics depends on the data only through $F(y_i)$ that are independent of the parameter of interest

c, and hence the maximum likelihood estimator of c is distribution-free, i.e., the expression for $\hat{c}_C$ is independent of the underlying distribution of the hashed values.

Therefore, using m copies of the kth order statistic results in a reduction in s.d. by a factor of $\sqrt{m}$ that equals the degree of reduction obtained by stochastic averaging with m buckets. Furthermore, the variance of the estimator depends on m and k only through the factor mk, so using mk copies of the first order statistic or m copies of the kth results in the same approximate variance for large c.

We compare the estimators of Giroire (2005) to the MLE $\hat{c}_C$ in terms of asymptotic relative efficiency (ARE), defined as the ratio of the asymptotic variance of the best possible estimator, the MLE, to that of the estimator under study (Lehmann, 1983), and show that Giroire's estimators are asymptotically efficient as $m \to \infty$ and $k \to \infty$.

$$\begin{aligned}
\mathrm{ARE}(\hat{c}_C, \xi_1) &\approx \frac{mk}{m(k-2)} \approx 1 \quad \text{for large } k, \\
\mathrm{ARE}(\hat{c}_C, \xi_2) &\approx 4k\left\{\frac{1}{k-1}\left(\frac{\Gamma(k)}{\Gamma(k-0.5)}\right)^2 - 1\right\} \quad \text{for } k \text{ large.}
\end{aligned}$$

Now, $\Gamma(k)/\Gamma(k-0.5) = (k-1)\Gamma(k-1)/\Gamma(k-1+0.5)$, and we use the following approximation

$$\begin{aligned}
\frac{\Gamma(k-1+0.5)}{\Gamma(k-1)} &= \sqrt{k-1}\left(1 - \frac{1}{8(k-1)} + \frac{1}{128(k-1)^2} + \frac{5}{1024(k-1)^3} \pm \dots\right) \\
&= \sqrt{k-1} - \frac{1}{8\sqrt{k-1}} + O\left(\frac{1}{(k-1)^{3/2}}\right) \\
&= \frac{8k-9}{8\sqrt{k-1}} + O\left(\frac{1}{(k-1)^{3/2}}\right),
\end{aligned}$$

thus obtaining, for k large,

$$\mathrm{ARE}(\hat{c}_C, \xi_2) \approx 4k\left\{(k-1)\left(\frac{8\sqrt{k-1}}{8k-9}\right)^2 - 1\right\} = \frac{64k^2 - 68k}{64k^2 - 144k + 81} \approx 1.$$

Lastly, for k large,

$$\begin{aligned}
\mathrm{ARE}(\hat{c}_C, \xi_3) &\approx k\Psi'(k) \\
&= k\sum_{i=0}^{\infty} \frac{1}{(k+i)^2}, \quad \text{where } \Psi'(k) = \frac{d}{dk}\ln\Gamma(k) \\
&= \lim_{n\to\infty} k\sum_{i=0}^{n} \frac{1}{(k+i)^2} \\
&= \lim_{n\to\infty} k\left\{\frac{1}{k^2} + \ldots + \frac{1}{(k+n)^2}\right\} \\
&\approx \lim_{n\to\infty} \frac{nk}{(k+n)(k-1)} \\
&= \frac{k}{k-1} \approx 1.
\end{aligned}$$

It should be noted that when estimating an integer valued parameter, such as the cardinality, the derivatives in, for example, equation (3.1) cannot be calculated. Nevertheless, equivalent results can be derived in terms of finite differences and since the standard deviation of the estimators we consider is of the order of c, with c large, the use of derivatives can be justified. For an early discussion of these issues, see Hammersley (1950).

Finally, we show that the estimator $\hat{c}_C$ in (3.2) is recursively computable. The property of recursive computability is particularly important when dealing with massive data sets due to constraints on available storage.

Definition 3.3.1. *A sequence of statistics $T_n(x_1, \ldots, x_n)$ is said to be recursively computable if $\forall n \in \mathbb{N}$, $T_n(x_1, \ldots, x_n) = T_n(z_1, \ldots, z_n) \Rightarrow T_{n+1}(x_1, \ldots, x_n, w) = T_{n+1}(z_1, \ldots, z_n, w)$.*

Lauritzen (1988) proves for independent variables $X_1, \ldots, X_n$ that if $T_n(x_1, \ldots, x_n)$ is minimal sufficient, then the sequence T_n, $n \geq 1$, is recursively computable; this characteristic of sufficient statistics was first remarked by Fisher (1925). We use a theorem of Lehmann and Scheffé (1950) to show that the statistic in (3.2) is minimal sufficient for c. The ratio

$$\frac{L(c; y_1, \ldots, y_m)}{L(c; z_1, \ldots, z_m)} = \frac{\left\{\prod_{i=1}^m y_i\right\}^{k-1}\left\{\prod_{i=1}^m (1-y_i)\right\}^{c-k}}{\left\{\prod_{i=1}^m z_i\right\}^{k-1}\left\{\prod_{i=1}^m (1-z_i)\right\}^{c-k}}$$

is constant as a function of c if and only if $\prod_{i=1}^{m}(1-y_i) = \prod_{i=1}^{m}(1-z_i)$. So, the statistic $T_m(y_1,\ldots,y_m) = \prod_{i=1}^{m}(1-y_i)$ is minimal sufficient for c; since the estimator $\hat{c}_C$ is in one-to-one correspondence with $T_m(y_1,\ldots,y_m)$, it follows that it is minimal sufficient for c and hence recursively computable.

Alternatively, we can show directly that $\hat{c}_C$ is recursively computable. Suppose the goal is to estimate the number of distinct stocks traided during a single day based on monitoring during peak hours in the morning and afternoon. The two estimates, $\hat{c}_{C,m_1}$ and $\hat{c}_{C,m_2}$, based on random samples of sizes m_1 and m_2, respectively, can be combined to give an estimate based on a random sample of size m_1+m_2:

$$\hat{c}_{C,m_1+m_2} = \frac{k}{1-\left[(1-k/\hat{c}_{C,m_1})^{m_1}\times(1-k/\hat{c}_{C,m_2})^{m_2}\right]^{1/(m_1+m_2)}}.$$

For the case $k=c$, i.e., storing the maximum order statistic, the likelihood function is

$$l(c;y_1,\ldots,y_m) = m\log c + \sum_{i=1}^{m}(c-1)\log y_i, \quad 0<y_1,\ldots,y_m<1.$$

The MLE of c is $\hat{c}_C = -m/\sum_{i=1}^{m}\log y_i$, and, as expected, is asymptotically Normal with mean c and variance c^2/m, having the property of recursive computability in m.

Similarly, for the case $k=1$, the MLE of c is given by $\hat{c}_C = -m/\sum_{i=1}^{m}\log(1-y_i)$; it is also asymptotically Normal with mean c and variance c^2/m, and recursively computable.

3.3.2 Hashing to discrete random variables

Hashing to integer values requires less storage as the stream is processed. Let the hash function map to discrete random variables, taking non-negative integer values, with cumulative distribution function F. Consider the special case that $F(x) = 1-\rho^{-x}$, for $\rho>1$, i.e., $X \sim \text{Geometric}(1-\rho^{-1})$, noting that $\rho=2$ is the case analysed by Flajolet (2004).

Let $M_c^{(d)}$ denote the dth largest value in a random sample of Geometric$(1-\rho^{-1})$ variables of size c. Kirschenhofer and Prodinger (1993) show that as $c\to\infty$,

$$\mathbb{E}\big(M_c^{(d)}\big) \approx \frac{\log(c)}{\log(\rho)} + \frac{\gamma}{\log(\rho)} + \frac{1}{2} - \frac{1}{\log(\rho)}H_{d-1} + P_1\big(\log_\rho(c)\big), \tag{3.3}$$

where γ is Euler's constant, $P_1(x)$ is a continuous, periodic function of period 1, mean zero and small amplitude, and H_k is the kth harmonic number defined as $H_k = \sum_{i=1}^{k} i^{-1}$, $H_0 = 0$. Furthermore, the variance of $M_c^{(d)}$ can be made arbitrarily small by choosing d large enough, in the limit as $c \to \infty$. However, Kirschenhofer and Prodinger (1993) point out that estimating c from $M_c^{(d)}$ for $d \geq 2$ "does not work well, because the d-maximum is sensitive (though not much) to multiple appearances of the same element". Indeed, the maximum order statistic ($d = 1$) is typically stored by algorithms hashing to Geometric($1-\rho^{-1}$) random variables because it lies far out in the tail of the distribution where the hash function has low probability of collision; see Figure 3.1.

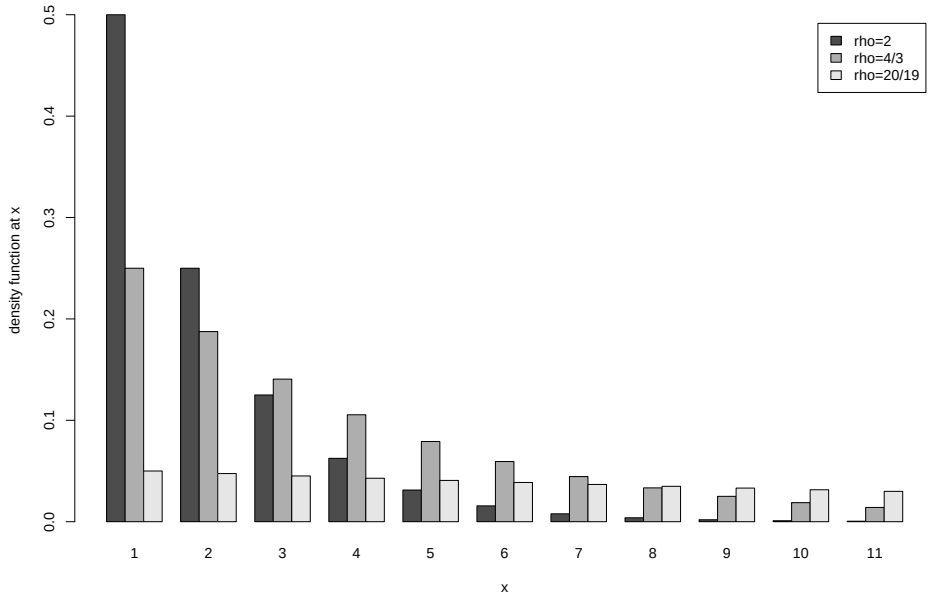


Figure 3.1: Comparison of probability mass functions of Geometric($1 - \rho^{-1}$) distribution.

Figure 3.1 also shows that as $\rho \to 1$, the probability of repetitions appearing in the sample tends to 0; this justifies first estimating the cardinality c from a sample $\left(Y_1^{(1)} = y_1, \ldots, Y_m^{(1)} = y_m\right)$ of minimum order statistics with ρ close to 1. The probability mass function (p.m.f.) of the minimum order statistic from a random sample of size c of Geometric($1 - \rho^{-1}$) variables

is given by

$$f(y;c) = \frac{1}{\rho^{(y-1)c}}\left(1 - \frac{1}{\rho^c}\right), \quad y = 1, 2, \ldots,$$

so it has a Geometric$(1-\rho^{-c})$ distribution. The log likelihood is

$$l(y_1, \ldots, y_m; c) = \sum_{i=1}^{m} \big\{ -y_i c \log(\rho) + \log(\rho^c - 1)\big\},$$

so

$$\frac{\partial l}{\partial c} = -\log(\rho)\sum_{i=1}^{m} y_i + \frac{m\rho^c}{\rho^c - 1}\log(\rho), \quad \frac{\partial^2 l}{\partial c^2} = -\frac{m[\log(\rho)]^2\rho^c}{(\rho^c-1)^2} \le 0,$$

and the MLE, denoted by $\hat{c}_D$, has the advantage of being available in closed form:

$$\hat{c}_D = \frac{\log\big(\sum_{i=1}^m y_i\big) - \log\big(\sum_{i=1}^m y_i - m\big)}{\log\rho};$$

furthermore, it has the property of recursive computability.

Fisher's information is, for $m = 1$,

$$I_1(c;\rho) = \sum_{x=1}^{\infty}\Big(-x\log\rho + \frac{\rho^c}{\rho^c-1}\log\rho\Big)^2\frac{\rho^c - 1}{\rho^{xc}} = \frac{\rho^c(\log\rho)^2}{(\rho^c-1)^2}.$$

For fixed c, $I_1(c;\rho)$ is a decreasing function of ρ for $\rho > 1$ with $\lim_{\rho\to 1} I_1(c;\rho) = c^{-2}$.

We compare the efficiency of $\hat{c}_D$ to that of $\hat{c}_C$ with $k = 1$. Let $\rho = 1 + \lambda/c$, for some fixed, positive constant λ. For large c, the ARE of $\hat{c}_D$ compared to $\hat{c}_C$ is

$$\lim_{c\to\infty} c^2 I_1(c;\rho) = \lim_{c\to\infty}\frac{(1+\lambda/c)^c\big[c\log(1+\lambda/c)\big]^2}{\big[(1+\lambda/c)^c - 1\big]^2} = \frac{\lambda^2 e^\lambda}{(e^\lambda - 1)^2},$$

a decreasing function of λ tending to 1 as $\lambda \to 0$. So an asymptotic relative efficiency arbitrarily close to 1 can be obtained by letting $\rho = 1 + \lambda/c$ with $\lambda \approx 0$. For example, $\lambda = 0.01$ results in a relative efficiency of 0.9999 compared with estimation based on the first order statistic from a continuous distribution. Hence, based on m replicates with $\rho \approx 1$, $\hat{c}_D \sim$ Normal$(c, c^2/m)$ approximately for large m. We therefore conclude that estimating the cardinality c by the MLE based on a random sample of size m of first order statistics from an arbitrary continuous distribution, for large m, is equivalent to estimating via the MLE

based on a random sample of size m of first order statistics from the Geometric$(1-\rho^{-1})$ distribution, for $\rho \approx 1$, $\rho > 1$.

Similar results are obtained for the MLE based on a random sample of maximum order statistics, $\big(Y_1^{(c)} = y_1, \ldots, Y_m^{(c)} = y_m\big)$, from the same distribution. The joint log likelihood is

$$l(c; y_1, \ldots, y_m) = \sum_{i=1}^{m} \log\Big\{\big(1-\rho^{-y_i}\big)^c - \big(1-\rho^{-(y_i-1)}\big)^c\Big\},$$

and the MLE $\hat{c}_D$ satisfies

$$\left.\frac{\partial l}{\partial c}\right|_{c=\hat{c}_D} = \sum_{i=1}^{m} \frac{\log\big(1-\rho^{-y_i}\big)\big(1-\rho^{-y_i}\big)^{\hat{c}_D} - \log\big(1-\rho^{-(y_i-1)}\big)\big(1-\rho^{-(y_i-1)}\big)^{\hat{c}_D}}{\big(1-\rho^{-y_i}\big)^{\hat{c}_D} - \big(1-\rho^{-(y_i-1)}\big)^{\hat{c}_D}} = 0. \quad (3.4)$$

Since $\hat{c}_D$ is not tractable in closed form, it is not recursively computable, and, more importantly, must be approximated by an iterative numerical technique like Newton-Raphson:

$$\hat{c}^{(r+1)} = \hat{c}^{(r)} - \left\{\frac{\partial^2 l}{\partial c^2}\right\}^{-1}\Bigg|_{c=\hat{c}^{(r)}} \cdot \left.\frac{\partial l}{\partial c}\right|_{c=\hat{c}^{(r)}}, \quad (3.5)$$

where

$$\frac{\partial^2 l}{\partial c^2} = \sum_{i=1}^{m} \frac{\big(1-\rho^{-y_i}\big)^c\big(1-\rho^{-(y_i-1)}\big)^c\big[\log\big(1-\rho^{-y_i}\big) - \log\big(1-\rho^{-(y_i-1)}\big)\big]^2}{\big[\big(1-\rho^{-y_i}\big)^c - \big(1-\rho^{-(y_i-1)}\big)^c\big]^2}.$$

It is known that if $\hat{c}^{(1)}$ is a consistent estimate of c, then $\hat{c}^{(2)}$ is an asymptotically efficient estimate and the procedure need not be iterated further (Rao, 1973).

So we proceed to find a consistent estimate of c. Define $Z = \sum_{i=1}^{m} \mathbb{I}[Y_i \le \Psi]$, for some large constant $\Psi \ge 1$. Then $Z \sim \text{Binomial}\big(m, (1-\rho^{-\Psi})^c\big)$, and let $\hat{Z} = \big|\{y_i; y_i \le \Psi\}\big|$. By the method of moments, if $\hat{Z} \neq 0$, equate $\mathbb{E}(Z) = m(1-\rho^{-\Psi})^c$ with $\hat{Z}$ and solve for c. The proposed estimator is

$$\hat{c} = \begin{cases} \frac{\log(\hat{Z}/m)}{\log(1-\rho^{-\Psi})} & \text{if } \hat{Z} \neq 0 \\ T & \text{otherwise,} \end{cases}$$

where T is the length of the observed data stream. Expanding the function $\log(Z/m)$ into a Taylor series about $\mathbb{E}Z$, we obtain that $\log(Z/m) = c\log(1-\rho^{-\Psi}) + m^{-1}(1-\rho^{-\Psi})^{-c}(Z - \mathbb{E}Z) + O(1/m)$, so $\mathbb{E}\log(Z/m) \approx c\log(1-\rho^{-\Psi})$.

Now, $\mathbb{E}\hat{c} \approx c\big\{1-(1-[1-\rho^{-\Psi}]^c)^m\big\} + T(1-[1-\rho^{-\Psi}]^c)^m$, and the bias of $\hat{c}$ tends to 0 as $m \to \infty$, since $\rho > 1$. Similarly, $\text{var}\left(\log(Z/m)/\log(1-\rho^{-\Psi})\right) \approx \rho^{2\Psi} m^{-1}\big[\exp(c\rho^{-\Psi}) - 1\big]$, using the approximation $(1-\rho^{-\Psi})^c \approx \exp(-c\rho^{-\Psi})$ for large c, and it is easy to show that $\text{var}(\hat{c}) \to 0$ as $m \to \infty$. Hence $\hat{c}$ is consistent for c. Setting $\hat{c}^{(1)} = \hat{c}$ in equation (3.5) results in an asymptotically efficient estimate $\hat{c}_D = \hat{c}^{(2)}$ of c.

For $m = 1$, Fisher's information is

$$I_1(c;\rho) = \sum_{x=1}^{\infty} \frac{[\log(1-\rho^{-x})(1-\rho^{-x})^c - \log(1-\rho\rho^{-x})(1-\rho\rho^{-x})^c]^2}{(1-\rho^{-x})^c - (1-\rho\rho^{-x})^c};$$

by letting $c = \rho^r$ and $x = r + k$, where r is an integer, it can be shown that

$$\lim_{c\to\infty} c^2 I_1(c;\rho) = \sum_{k=-\infty}^{\infty} \frac{1}{\rho^{2k}} \frac{(\rho-1)^2}{[\exp(\rho\rho^{-k}) - \exp(\rho^{-k})]},$$

a decreasing function of ρ as $\rho \to \infty$. In particular, for $\rho = 1.1$, the ARE of the MLE based on a sample of maximum order statistics from the Geometric distribution as compared to the MLE based on a random sample of minimum order statistics from an arbitrary continuous distribution is 0.9985. For the algorithm of Flajolet (2004) with $\rho = 2$, the ARE is 0.9304.

In the extreme case of discretisation to Bernoulli random variables with probability of success p, the MLE of c based on the random sample $\big(Y_1^{(c)} = y_1, \ldots, Y_m^{(c)} = y_m\big) \sim \text{Bernoulli}\big(1-(1-p)^c\big)$ is given by

$$\hat{c} = \frac{\log(1-\sum_{i=1}^m y_i)}{\log(1-p)}.$$

The Fisher information, for $m = 1$,

$$I(c;p) = \frac{(1-p)^c\big[\log(1-p)\big]^2}{1-(1-p)^c}$$

is maximised by $p_{max} = 1 - \exp\big(-\lambda_0/c\big) \approx \lambda_0/c$, for large c, where $\lambda_0 \approx 1.593623$. Letting $p = \lambda/c$, for fixed $0 < \lambda < c$, results in an asymptotic relative efficiency of

$$\lim_{c\to\infty} c^2 I(c;p) = \lim_{c\to\infty} \frac{(1-\lambda/c)^c}{1-(1-\lambda/c)^c}\Big[\log\Big(1-\frac{\lambda}{c}\Big)^c\Big]^2 = \frac{\lambda^2}{e^\lambda - 1},$$

that equals approximately 0.65 when $\lambda = \lambda_0$. We observe that the ARE is above 30% on the range $\lambda \in (0.4, 4)$, and conclude that if c is known in advance to lie in the range $(0.4c_0, 4c_0)$, for some fixed c_0, then $p = c_0^{-1}$ guarantees at least 30% efficiency.

3.3.3 Complexity results and tail bounds

Let $\hat{c}_C$, $\hat{c}_D$ denote the maximum likelihood estimator based on a random sample of minimum order statistics from an arbitrary continuous distribution (see Subsection 3.3.1), and the MLE based on a random sample of maximum order statistics from the Geometric$(1-\rho^{-1})$, $\rho = 1.1$, distribution (see Subsection 3.3.2), respectively.

In hashing to discrete random variables, the maximum order statistic is preferred to the minimum because the former requires less space to store and lies far out into the tail of the distribution where the hash function has low collision probability. If the algorithm is based on the minimum order statistic, then a succession of decreasing values is stored, and the size of the largest of these values determines the storage requirement. The largest value is the first one observed, a Geometric$(1 - \rho^{-1})$ random variable with expectation $\rho/(\rho - 1)$, where $\rho = 1 + \lambda/c$. In terms of c, the expected value is $1 + c/\lambda$, so it is of order $O(c)$. Hence, a conservative estimate of the space required to store one sample of the minimum order statistic is $O(\log c)$. On the other hand, a very optimistic estimate of the space required to store one sample of the maximum order statistic is obtained by looking at the expected value of the maximum, given by (3.3) with $d = 1$, which is of order $O(\log c)$; so the algorithm requires $O(\log \log c)$ space.

We proceed to obtain tail bounds for the distribution of $\hat{c}_C$, where $\hat{c}_C = -m/\sum_{i=1}^{m} \log(1 - F(y_i))$, and $(y_1, \ldots, y_m)$ is a random sample of minimum order statistics, each coming from an independent sample of size c from a continous distribution F. It follows that $F(y_i)$ is the minimum order statistic of a random sample of size c from the Unif$(0, 1)$ distribution (David and Nagaraja, 2003), and $-\log(1 - F(y_i))$ follows the Exponential distribution with

mean c^{-1}. Hence, $-\sum_{i=1}^{m}\log(1-F(y_i))$ is Gamma distributed with shape parameter m and scale parameter c^{-1}. We use Chernoff's approach (Chernoff, 1952) to obtain tail bounds. For $0<\epsilon<1$,

$$\begin{aligned}
\mathbb{P}\Big(\hat{c}_C \geq (1+\epsilon)c\Big) &= \mathbb{P}\Big(-\frac{m}{\sum_{i=1}^{m}\log(1-F(y_i))} \geq (1+\epsilon)c\Big) \\
&= \mathbb{P}\Big(\lambda\sum_{i=1}^{m}\log(1-F(y_i)) \geq -\frac{\lambda m}{(1+\epsilon)c}\Big), \text{ for } \lambda>0 \\
&= \mathbb{P}\Big(\exp\big(\lambda\sum_{i=1}^{m}\log(1-F(y_i))\big) \geq \exp\big(-\lambda m/[c(1+\epsilon)]\big)\Big) \\
&\leq \mathbb{E}\Big(\exp\big(\lambda\sum_{i=1}^{m}\log(1-F(Y_i))\big)\Big)\exp\big(\lambda m/[c(1+\epsilon)]\big) \\
&\leq \inf_{\lambda>0}\Big\{\exp\big(\lambda m/[c(1+\epsilon)]\big)\times\big(1+\lambda/c\big)^{-m}\Big\} \\
&= \exp\big(m\epsilon/(1+\epsilon)\big)\times(1+\epsilon)^{-m} \\
&= \exp\big(-m\epsilon^2/G_1\big),
\end{aligned}$$

is exponentially decreasing with constant

$$G_1 = \frac{\epsilon^2(1+\epsilon)}{-\epsilon+(1+\epsilon)\log(1+\epsilon)}.$$

Similarly,

$$\mathbb{P}\Big(\hat{c}_C \leq (1-\epsilon)c\Big) \leq \exp\big(-m\epsilon/(1-\epsilon)\big)\times(1-\epsilon)^{-m} = \exp\big(-m\epsilon^2/G_2\big),$$

with constant

$$G_2 = \frac{\epsilon^2(1-\epsilon)}{\epsilon+(1-\epsilon)\log(1-\epsilon)}.$$

On the other hand, the asymptotic tail bounds are

$$\mathbb{P}\big(|\hat{c}_C - c| \geq \epsilon c\big) \leq 2\exp\big(-m\epsilon^2/2\big),$$

and it can easily be shown that $\lim_{\epsilon\to 0^+} G_1 = \lim_{\epsilon\to 0^+} G_2 = 2$ agree with what is expected in the limit as $m\to\infty$. Letting $G=\max\{G_1,G_2\}$, for approximation parameter $0<\epsilon<1$ and finite m, we have

$$\mathbb{P}\big(|\hat{c}_C - c| \geq \epsilon c\big) \leq 2\exp\big(-m\epsilon^2/G\big).$$

Given confidence parameter $1-\delta \in (0,1)$, $\hat{c}_C$ is an (ϵ, δ)-approximation to c provided

$$2\exp\big(-m\epsilon^2/G\big) \leq \delta \iff m \geq G\epsilon^{-2}\log(\delta/2),$$

i.e., the sketch size m must be of $O(\epsilon^{-2})$.

For $\rho = 1.1$, the ARE of $\hat{c}_D$ compared to $\hat{c}_C$ is 0.9985; however, the estimator does not exist in closed form, so we cannot attempt to obtain small sample tail bounds. Instead, we approximate the Geometric$(1-\rho^{-1})$ distribution with c.d.f. $F_g(x) = 1-\rho^{-x}$ by the Exponential distribution with mean λ^{-1} satisfying $\lambda = \log\rho$ and c.d.f. $F_e(x) = 1-e^{-\lambda x}$. So the log-likelihood is approximately given by

$$l(c; y_1, \ldots, y_m) = m\log(\lambda c) - \lambda\sum_{i=1}^{m} y_i + (c-1)\sum_{i=1}^{m}\log\big(1-e^{-\lambda y_i}\big),$$

and the maximum likelihood estimator is $\hat{c}_D^{\star} = -m/\sum_{i=1}^{m}\log(1-e^{-\lambda y_i})$. Now, $\prod_{i=1}^{m}(1-e^{-\lambda y_i}) = \prod_{i=1}^{m}(1-\rho^{-y_i})$ is sufficient for c, so, to this degree of approximation, the MLE is recursively computable. Furthermore, it can be shown that $-\log\big(1-e^{-\lambda Y_i}\big) \sim \text{Exp}(c)$, so $m/\hat{c}_D^{\star} \sim \Gamma(m, c^{-1})$ follows the same distribution as $m/\hat{c}_C$, based on hashing to a continuous distribution and storing the minimum order statistics. Hence the tail bounds are as detailed above, and, to obtain an (ϵ, δ)-approximation of c, the sketch size m must be of order $O(\epsilon^{-2})$. Therefore the cardinality estimation algorithm that hashes to Geometric$(1-\rho^{-1})$ random variables (with ρ nearly 1) and stores the maximum order statistics requires $O(\epsilon^{-2}\times\log\log c)$ space, thus attaining the tight lower bound of Indyk and Woodruff (2003).

We recommend estimating the cardinality c by maximum likelihood based on a random sample of maximum order statistics obtained by hashing to Geometric$(1-\rho{-}1)$ random variables with $\rho = 1.1$ via m independent hash functions; the algorithm appears below, and it applies to streams in the cash register case.

Step 1:

for j in $1:m$ set $y_j = 0$

Step 2:

for every pair (i_t, d_t) with $d_t > 0$ {

use i_t to seed the random number generator

for j in $1 : m$ {

generate $z \sim \text{Geometric}(1 - \rho^{-1})$

set $y_j = \max\{y_j, z\}$

}

}

Step 3:

return c the root of

$$\sum_{i=1}^{m} \frac{\log\left(1-\rho^{-y_i}\right)\left(1-\rho^{-y_i}\right)^c - \log\left(1-\rho^{-(y_i-1)}\right)\left(1-\rho^{-(y_i-1)}\right)^c}{\left(1-\rho^{-y_i}\right)^c - \left(1-\rho^{-(y_i-1)}\right)^c} = 0$$

3.4 Cardinality estimation via stable law sketching

This section presents the method of data sketching of Cormode et al. (2003); based on the idea of sketching to positive, strictly stable random variables, we derive a maximum likelihood estimator for the cardinality that is twice as efficient as the estimator in Cormode et al. (2003). Furthermore, we show a previously unsuspected link between the method of hashing to continuous random variables and storing order statistics, and that of data sketching.

Cormode et al. (2003) introduce stable law sketching for estimation of the cardinality of a general data stream ($a_i \in \mathbb{R}$). Let h be a hash function mapping $i \in \mathcal{D}$ to $h(i)$, an independent copy of a random variable $X \sim F(x; \alpha, 0, 1, 0)$. Rather than storing and updating the minimum or maximum of these variables, the sum $V = \sum_{t=1}^{T} d_t h(i_t)$ is calculated as the stream is processed. Noticing that $\sum_{t=1}^{T} d_t h(i_t) = \sum_{i \in \mathcal{D}} a_i h(i)$ and since $h(i) \overset{\mathcal{D}}{=} X$, $\forall i \in \mathcal{D}$, it follows from Theorem 2.1.1 that $V \overset{\mathcal{D}}{=} \left(\sum_{i \in \mathcal{D}} |a_i|^\alpha\right)^{1/\alpha} X$, i.e., V is a scaled version of X, where the scaling factor is precisely the α-norm of $\mathbf{a}$. This is the basis of the ideas developed

in Indyk (2006) for $\alpha = 1, 2$ and extended in Cormode et al. (2003) to $\alpha \in (0, 2]$.

This procedure is repeated independently m times to produce a sample $V_1, \ldots, V_m$, called a *data sketch*, where $V_i \overset{\mathcal{D}}{=} l_\alpha(\mathbf{a})X$, for $i = 1, \ldots, m$. Since the density of X is not known in tractable form for most values of $\alpha \in (0, 2]$, Cormode et al. (2003) propose estimating $l_\alpha(\mathbf{a})$ by $\tilde{V}/\tilde{\mu}_\alpha$, where $\tilde{V}$ is the median of $|V_1|, \ldots, |V_m|$, and $\tilde{\mu}_\alpha$ is the theoretical median of $|X|$. Given $\epsilon, \delta > 0$, letting m be of order $O\big(\epsilon^{-2}\log(\delta^{-1})\big)$ ensures that, with probability $(1 - \delta)$,

$$(1 - \epsilon)\frac{\tilde{V}}{\tilde{\mu}_\alpha} \leq l_\alpha(\mathbf{a}) \leq (1 + \epsilon)\frac{\tilde{V}}{\tilde{\mu}_\alpha}.$$

Furthermore, Cormode et al. (2003) show that if $|a_i| \leq \beta \ \forall i$, then

$$|\mathbf{a}|_H \leq \big(l_\alpha(\mathbf{a})\big)^\alpha \leq (1 + \epsilon)|\mathbf{a}|_H,$$

provided $0 < \alpha \leq \epsilon/\log(\beta)$. Therefore, $\big(\tilde{V}/\tilde{\mu}_\alpha\big)^\alpha$ estimates the Hamming norm $|\mathbf{a}|_H$ (and hence the cardinality of a simple data stream) up to multiplicative factor $1 \pm \epsilon$ with probability $1 - \delta$ requiring $O\big(\epsilon^{-2}\log(\delta^{-1})\big)$ storage. In the implementation of the algorithm, the authors set the parameter α equal to 0.02.

3.4.1 Data sketching and maximum likelihood estimation

First, consider the problem of cardinality estimation for non-negative data streams ($a_i \geq 0 \ \forall i$). Let h_j, $j = 1, \ldots, m$, be independent hash functions mapping from $\mathcal{D}$ to positive, strictly stable random variables of index α. For $j = 1, \ldots, m$, define

$$Y_j := V_j^\alpha = \Big(\sum_{i \in \mathcal{D}} a_i h_j(i)\Big)^\alpha \overset{\mathcal{D}}{=} \sum_{i \in \mathcal{D}} a_i^\alpha X^\alpha,$$

where $X \sim S(x; \alpha)$. Letting $\alpha \to 0$, $\sum_{i \in \mathcal{D}} a_i^\alpha \to c$, and $X^\alpha \to 1/L$ by Lemma 2.3.1, where $L \sim \text{Exp}(1)$. So, the likelihood of the random sample $\big(Y_1 = y_1, \ldots, Y_m = y_m\big)$ tends to

$$L(c; y_1, \ldots, y_m) = \prod_{i=1}^m \exp\big(-c/y_i\big)\frac{c}{y_i^2} = c^m \exp\big(-c\sum_{i=1}^m y_i^{-1}\big)\prod_{i=1}^m y_i^{-2},$$

and the MLE of c is given by

$$\hat{c}_{DS} = m/\sum_{i=1}^{m} y_i^{-1}. \tag{3.6}$$

Fisher's information is m/c^2, so $\hat{c}_{DS} \sim \text{Normal}\left(c, c^2/m\right)$ approximately for large m. Moreover, $mc/\hat{c}_{DS}$, with $\Gamma(m,1)$ distribution, is a pivot for confidence interval purposes; this follows from the fact that as $\alpha \to 0$, $Y_i^{-1} \sim \text{Exp}(c)$ (mean c^{-1}), so $\sum_{i=1}^{m} Y_i^{-1} \sim \Gamma(m, c^{-1})$. It follows that the tail bounds are identical to those of $\hat{c}_C$, the MLE based on hashing to continuous random variables and storing minimum order statistics; see Section 3.3.3 for details. Moreover, as in Section 3.3.1, it can easily be shown that $\sum_{i=1}^{m} y_i^{-1}$ is minimal sufficient for c, so it follows that $\hat{c}_{DS}$ is minimal sufficient for c and hence recursively computable. The algorithm is presented below, and it applies to non-negative data streams. In practice, we set $\alpha = 0.02$.

Step 1:

for j in $1:m$ set $V_j = 0$

Step 2:

for every pair (i_t, d_t) {

use i_t to seed the random number generator

for j in $1:m$ {

generate $L \sim \text{Exp}(1)$, $U \sim \text{Unif}(0,\pi)$

set $a(U) = \left[\sin(\alpha U)/\sin(U)\right]^{(1-\alpha)^{-1}} \times \left[\sin\left((1-\alpha)U\right)/\sin(\alpha U)\right]$

set $V_j = V_j + d_t\left[a(U)/L\right]^{(1-\alpha)/\alpha}$

}

}

Step 3:

return $m/\sum_{j=1}^{m} V_j^{-\alpha}$.

Next, we compare the efficiency of the maximum likelihood estimator, $\hat{c}_{DS}$, to that of the estimator of Cormode et al. (2003), denoted by $\tilde{c} = \left(\tilde{V}/\tilde{\mu}_\alpha\right)^\alpha$, where $\tilde{V}^\alpha$ is the median of

$\left(V_1^\alpha, \ldots, V_m^\alpha\right)$. Let $\tilde{V}^0 = \lim_{\alpha \to 0} \tilde{V}^\alpha$. As $\alpha \to 0$, $V_j^\alpha \to c/L$, where $L \sim \text{Exp}(1)$, so $\tilde{V}^0$ is the median of a random sample of variables distributed as c/L. From Cramér (1946),

$$\tilde{V}^0 \sim \text{Normal}\left(\frac{-c}{\log(0.5)}, \frac{c^2}{m[\log(0.5)]^4}\right) \quad \text{as } m \to \infty,$$

and $\left(\tilde{\mu}_\alpha\right)^\alpha \to -1/\log(0.5)$ as $\alpha \to 0$, so, the variance of $\tilde{c}$ is approximately $c^2/[m(\log 0.5)^2]$. Then the ARE of $\hat{c}_{DS,1}$ with respect to $\tilde{c}$ is

$$\text{ARE}(\hat{c}_{DS}, \tilde{c}) = \lim_{m \to \infty} \left\{ \frac{c^2}{m[\log(0.5)]^2} \times \frac{m}{c^2} \right\} = (\log 0.5)^{-2} \approx 2.08,$$

i.e., the MLE obtained from projections, $\hat{c}_{DS}$, is twice as efficient as the estimator based on the median $\tilde{c}$ with $\alpha \approx 0$.

Finally, consider the problem of cardinality estimation for general data streams. Let h_j, $j = 1, \ldots, m$, map from $\mathcal{D}$ to independent copies of $X \sim F(x; \alpha, 0, 1, 0)$, i.e., a symmetric, strictly stable random variable. Then, as the stream is processed, we store

$$Y_j := |V_j|^\alpha = \left| \sum_{i \in \mathcal{D}} a_i h_j(i) \right|^\alpha \overset{\mathcal{D}}{=} \sum_{i \in \mathcal{D}} |a_i|^\alpha |X|^\alpha \to \frac{c}{L},$$

as $\alpha \to 0$ by Lemma 2.3.2. Hence the maximum likelihood estimator of c is identical to $\hat{c}_{DS}$ in the limit as $\alpha \to 0$, having the same asymptotic distribution.

3.4.2 Connection between data sketching and hashing to continuous random variables

In this subsection we present a previously unsuspected link between the methods pioneered by Flajolet (2004), that hash to continuous random variables and store order statistics, and Cormode et al. (2003), that store sketches of linear combinations of α-stable random variables for the problem of cardinality estimation of non-negative data streams.

In Subsection 3.3.1 we showed that the efficiency of the MLE based on kth order statistics from a continuous distribution does not depend on the particular distribution simulated by

the hash function. The simplest continuous distribution is obviously Unif$(0,1)$, but it is interesting to see what happens when sampling from $S(x;\alpha)$, the positive, strictly stable distribution. The first thing to observe is that these distributions have extremely heavy tails when α is small, so that storing the maximum of c such variables, for c large, requires high precision floating point numbers.

Theorem 3.4.1. *Consider a simple data stream over $\mathcal{D}$ observed up to time T. Let* $\mathbf{a}$ *denote the accumulation vector, and* c *the cardinality. Let $\{h_j\}$ be independent hash functions mapping from $\mathcal{D}$ to random copies of $X \sim S(x;\alpha)$, for fixed $\alpha \in (0,1)$. Define $h_j(i_t)$ to be the hashed value returned by function h_j after i_t is used to seed the pseudo-random number generator.*

In the limit as $\alpha \to 0$, the method of hashing and storing the maximum order statistic $\big(X_{(c)}\big)^\alpha = \max\big\{[h_j(i_t)]^\alpha, t = 1,\ldots,T\big\}$, for $j = 1,\ldots,m$, is identical to the method of data sketching that stores $\sum a_{\pi(i)} X_{(i)}$, where $X_{(1)},\ldots,X_{(c)}$ are the order statistics of a sample of size c from $S(x;\alpha)$ and π is a random permutation of $\{i_t, t = 1,\ldots,T\}$. In particular, if $\alpha \log c \to 0$ (condition A), and $\alpha \log(\mathbf{a}_{\max}) \to 0$ (condition B) hold as $\alpha \to 0$, where $\mathbf{a}_{\max} = \max\{a_i, i = i_t, t = 1,\ldots,T\}$, then the pivotal quantities for the two methods are identical.

Proof. First, consider the method of hashing, and storing, for $j = 1,\ldots,m$,

$$Y_j = \big[\max\{h_j(i_t), t = 1,\ldots,T\}\big]^\alpha = \big(X_{(c)}\big)^\alpha.$$

From subsection 3.3.1, the MLE of c based on $(Y_1,\ldots,Y_m)$ is $\hat{c} = -m/\sum_{j=1}^m \log F(Y_j)$, where $F(y) = \mathbb{P}\left(X^\alpha \le y\right)$, $X \sim S(x;\alpha)$. Since $X^\alpha \to 1/L$ where $L \sim \text{Exp}(1)$ as $\alpha \to 0$, it follows that

$$\mathbb{P}\left(X^\alpha \le y\right) \to \mathbb{P}\left(1/L \le y\right) = \mathbb{P}\left(L \ge 1/y\right) = e^{-1/y}, \quad 0 < y < 1,$$

so $\hat{c}$ can be approximated arbitrarily closely by $m/\sum_{j=1}^m Y_j^{-1}$ with α small. In other words, under this condition, the estimator has the same form as (3.6), the MLE of c obtained from

projections. Furthermore, just as in the case of a projection sketch, the ratio $mc/\hat{c}$ provides a pivot with $\Gamma(m,1)$ distribution as $\alpha \to 0$. This follows from the fact that as $\alpha \to 0$,

$$\mathbb{P}\left(1/Y_i \leq y\right) = 1 - \mathbb{P}\left(Y_i < 1/y\right) \to 1 - \left[\mathbb{P}\left(L \geq y\right)\right]^c = 1 - e^{-cy},$$

so $1/Y_i \xrightarrow{\mathcal{D}} L$, where $L \sim \text{Exp}(c)$. By Claim 2.3.1, the pivot distribution does not converge uniformly to $\Gamma(m,1)$ for all c, i.e., as $T \to \infty$ and $|\{i_t, t = 1, \ldots, T\}|$ increases, but it is sufficient that $\alpha \to 0$ such that $\alpha \log c \to 0$ (condition A). We note that this condition does not involve the size of the elements of $\mathbf{a}$.

Similarly, to show that the pivot distribution for the projection sketch converges uniformly, we need to show that as $\alpha \to 0$,

$$\frac{(\sum a_i X_i)^\alpha}{c} \overset{\mathcal{D}}{=} X^\alpha \frac{\sum a_i^\alpha}{c} \xrightarrow{\mathcal{D}} \frac{1}{L}.$$

Since we have already noted that $X^\alpha \xrightarrow{\mathcal{D}} 1/L$, this only requires that $\sum a_i^\alpha / c \to 1$. Now,

$$1 \leq \frac{\sum a_i^\alpha}{c} \leq \mathbf{a}_{\max}^\alpha,$$

so a sufficient condition is $\alpha \log(\mathbf{a}_{\max}) \to 0$ as $\alpha \to 0$ (condition B). We note that this condition does not involve the cardinality of the stream.

We are now in a position to relate the two methods. Since $a_i \geq 1$, for $i = i_t, t = 1, \ldots, T$, the ratio of the pivotal quantities satisfies

$$\left(1/\sum a_i\right)^\alpha \leq \left(X_{(c)}/\sum a_{\pi(i)} X_{(i)}\right)^\alpha \leq 1,$$

and when both conditions A and B apply, the ratio converges to 1 since the left hand side converges to 1. We conclude that, to this level of approximation, the pivotal quantities will be identical for the two methods of sketching, thus revealing a previously unsuspected link between these two methods of cardinality estimation. □

3.5 Summary

This chapter introduces two methods for cardinality estimation in streaming data: hashing and storing order statistics, and data sketching via projections to stable random variables. In both cases, we derive recursively computable, maximum likelihood estimators of the cardinality. In the former, we recommend hashing to Geometric$(1-\rho^{-1})$ random variables with $\rho = 1.1$, and storing maximum order statistics; in this case, the algorithm attains the tight lower bound on space complexity. In the latter, we recommend projecting to the $S(x;\alpha)$ distribution with $\alpha = 0.02$, and show that the resulting estimator is twice as efficient as that of Cormode et al. (2003).

Finally, the main result of this chapter is Theorem 3.4.1 that proves an unsuspected link between these two methods: under certain conditions, the pivotal quanties from hashing to $S(x;\alpha)$ random variables and storing maximum order statistics, and data sketching, are identical. Of course, since there is no gain in efficiency for the method of hashing and storing order statistics in using the $S(x;\alpha)$ distribution, rather than the simpler Unif$(0,1)$ distribution, the latter is to be preferred. Moreover, since we have shown in Subsection 3.3.2 that discrete hash functions are capable of comparable efficiency but with reduced storage requirements, hashing to discrete random variables must be the overall method of choice.

Chapter 4

Distance estimation via random projections

Efficient estimation of distances between high-dimensional data vectors is an increasingly important objective in modern statistical analysis with many applications, e.g., clustering and classification. This chapter analyses the problem of distance-preserving dimension reduction with the aim of preserving l_α distances (quasi-distances) for $\alpha \in (0, 2]$. We introduce the problem and current literature in the context of data in high dimensions, and relate this to streaming data in Section 4.2, where a data stream, represented by its accumulation vector, can be viewed as a high-dimensional point.

Let V be a collection of n points in $\mathbb{R}^m$, where m is very large. The data is arranged into a matrix $\mathbf{V}$ with n rows and m columns, i.e., one row for each of the n data points. We are interested in projecting the m-dimensional points into a lower k-dimensional space via α-stable random projections (Indyk, 2006) such that l_α distances (quasi-distances) between original points can be recoved with full statistical efficiency, for fixed $\alpha \in (0, 2]$. The cases $\alpha = 1$ and $\alpha = 2$, corresponding to l_1 and l_2 distances respectively, are of special interest, as is the limiting case $\alpha \to 0$ which yields the Hamming distance.

In this chapter we propose asymptotically efficient estimators of l_α distances (quasi-

distances), for $\alpha \in (0, 2]$, derived from projections to symmetric, or maximally skewed strictly stable random variables of index α. We analyse the small sample performance of these estimators, and propose improvements via trimming and winsorising.

4.1 Dimension reduction in l_1 and l_2

The first result on projecting onto a lower dimensional space is the Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss, 1984) for dimension reduction in l_2. It states that for any $0 < \epsilon < 1$ and any positive integer k satisfying $k \geq k_0 = O(\epsilon^{-2} \log n)$, there exists a map $f_V : \mathbb{R}^m \to \mathbb{R}^k$, which depends on V, such that

$$(1-\epsilon)\|u-v\|_2^2 \leq \|f_V(u) - f_V(v)\|_2^2 \leq (1+\epsilon)\|u-v\|_2^2 \quad \forall u, v \in V, \tag{4.1}$$

i.e., l_2 distances between original points in V are well approximated by l_2 distances between corresponding projected points. The result proves the existence of a map f_V, but does not provide a deterministic formulation for it. Indeed, the latter would be an unrealistic objective as it would involve computing all pairwise distances between points in V to check that requirement (4.1) holds. Frankl and Maehara (1988) prove that the Johnson-Lindenstrauss lemma holds for any $0 < \epsilon < 0.5$, provided that $k \geq k_0 = \lceil 9(\epsilon^2 - 2\epsilon^3/3)^{-1} \log n \rceil + 1$, and V is sufficiently large; again, the existence of the map is proved, but f_V is not found explicitly. Dasgupta and Gupta (2003) tighten further the lower bound k_0 by obtaining $k_0 = 4(\epsilon^2/2 - \epsilon^3/3)^{-1} \log n$, for any $0 < \epsilon < 1$ and any set V of n points in $\mathbb{R}^m$.

The proof of Dasgupta and Gupta (2003) proceeds by constructing an explicit function f_V such that (4.1) holds with high probability. This approach suggests pursuing a more realistic objective: find a lower bound k_0 that depends on ϵ, δ, and n, such that if $k \geq k_0$, then there exists a mapping $f : \mathbb{R}^m \mapsto \mathbb{R}^k$, independent of V, satisfying

$$\mathbb{P}\left((1-\epsilon)\|u-v\|_2^2 \leq \|f(u) - f(v)\|_2^2 \leq (1+\epsilon)\|u-v\|_2^2, \quad \forall u, v \in V\right) > 1 - \delta, \tag{4.2}$$

for $0 < \delta < 1$, $0 < \epsilon < 1$ small. This objective is a relaxed version of (4.1), but has the advantage of describing the projection map f deterministically. The random projection method (Vempala, 2004) for dimension reduction in l_2 pursues this objective.

Definition 4.1.1. *Let* $\mathbf{P} \in \mathbb{R}^{m \times k}$ *be a matrix whose entries are independent random variables.* $\mathbf{P}$ *is called a random projection matrix mapping from* $\mathbb{R}^m$ *to* $\mathbb{R}^k$*, where* $k \ll m$*, via the linear map* $\mathbf{V} \mapsto \mathbf{B} = \mathbf{VP}$*.*

The random projection method maps points $u = (u_1, \ldots, u_m)$ and $v = (v_1, \ldots, v_m)$ in V to linear combinations of random variables from which the distances (quasi-distances) $\left[d_\alpha(u,v)\right]^\alpha = \sum_{i=1}^m |u_i - v_i|^\alpha$ can be recovered with high degree of accuracy. We remark that the Hamming distance is obtained as $\lim_{\alpha \to 0} \left[d_\alpha(u,v)\right]^\alpha$.

Dimension reduction via random projections is similar in aim to multidimensional scaling (MDS) (Torgerson, 1958). See Cox and Cox (2001) for an introduction. MDS searches for a low dimensional, graphical representation of high dimensional data that approximately preserves dissimilarities between original points. There are many flavours of MDS, with the most widely used being least squares scaling, in which a k-dimensional representation is seeked (with k fixed, typically, $k = 2$) such that Euclidean distances between projected points closely match some continuous, monotonic function of original dissimilarities. The projected points are found by minimising a loss function via least squares, where the choice of loss function is to a large extent arbitrary. The main difficulties in implementing MDS are avoiding local minima of the loss function, and choosing a dimension k that results in an informative low dimensional configuration.

Definition 4.1.2. *The matrix* $\mathbf{P} \in \mathbb{R}^{m \times k}$ *with independent* Normal$(0,1)$ *entries, scaled by* $\sqrt{k}$*, is called a conventional random projection matrix.*

For dimension reduction in l_2 with conventional random projections, the distance between u and v, $d_2(u,v)$, is estimated via maximum likelihood. The MLE is linear and equals

$\sum_{z=1}^{k}(a_z - b_z)^2$ (Achlioptas, 2003; Vempala, 2004), where $a = (a_1, \ldots, a_k)$ and $b = (b_1, \ldots, b_k)$ are rows of the matrix of projected points $\mathbf{B}$, corresponding to u and v, respectively. Given $\epsilon, \beta > 0$, if $k \geq k_0 = (4 + 2\beta)\big(\epsilon^2/2 - \epsilon^3/3\big)^{-1} \log n$, then the requirement (4.2) is satisfied with $\delta = n^{-\beta}$ (Achlioptas, 2003). Li et al. (2006a) improve the lower bound k_0 when the marginal norms $\|u\|_2^2$ are known. Furthermore, they show that sign random projections, that store only the sign of the projected data, are as efficient as conventional random projections, while significantly reducing the storage cost per projection.

The computational cost is further reduced by using a sparse random projection matrix $\mathbf{P}$, and the lower bound k_0 is unchanged (Achlioptas, 2003). In fact, it is sufficient that the entries of $\mathbf{P}$ be independent, identically distributed with zero mean and bounded moments (Arriaga and Vempala, 2006). Li et al. (2006b) generalise this result by considering the following probability distribution on the entries p_{ij} of $\mathbf{P}$:

$$\mathbb{P}(p_{ij} = \sqrt{s}) = \mathbb{P}(p_{ij} = -\sqrt{s}) = (2s)^{-1}, \ \mathbb{P}(p_{ij} = 0) = 1 - s^{-1},$$

where $s = 1, 3$ correspond to the distributions in Achlioptas (2003). Li et al. (2006b) suggest taking $s = \sqrt{m}$ or $m/\log(m)$ when the data points are approximately normal; the corresponding projections are called very sparse random projections. Under the assumptions that all fourth moments exist, very sparse random projections with $s = o(m)$ have asymptotically the same distribution as conventional random projections (Vempala, 2004).

For dimension reduction in l_1, the entries of $\mathbf{P}$ are independent Cauchy$(0, 1)$ random variables (Indyk, 2006). Li et al. (2007) propose an analog of the result of Achlioptas (2003) for conventional random projections; they define a bias-corrected maximum likelihood estimator of the distance $d_1(u, v)$ based on the random sample $\{a_z - b_z; z = 1, \ldots, k\}$ of Cauchy$(0, d_1(u, v))$ variables, and show that the dimension of the projection space, k, must be of order $O(\log n/\epsilon^2)$, for given $0 < \epsilon < 1$. The MLE is nonlinear and can be found by a few iterations of the Newton-Raphson algorithm. Brinkman and Charikar (2005) prove that recovering l_1 distances with constant distortion, in the sense of the Johnson-Lindenstrauss

lemma, by corresponding l_1 distances between projected points requires a number of dimensions k polynomial in n, so an analog of the Johnson-Lindenstrauss lemma does not exist.

4.2 Dimension reduction in l_α, $0 < \alpha \leq 2$

Let the entries of $\mathbf{P}$ be independent, strictly stable random variables of index α, skewness β, scale $\gamma = 1$, and location $\delta = 0$, i.e., $p_{ij} \sim F(x; \alpha, \beta, 1, 0)$. Consider $u = (u_1, \ldots, u_m)$ and $v = (v_1, \ldots, v_m) \in V$, with corresponding rows $a = (a_1, \ldots, a_k)$ and $b = (b_1, \ldots, b_k)$ in $\mathbf{B} = \mathbf{VP}$. Then, for $i = 1, \ldots, k$, we have

$$x_j := a_j - b_j = \sum_{i=1}^{m} (u_i - v_i) p_{ij} \sim F\big(x; \alpha, \beta, \gamma = d_\alpha(u, v), 0\big),$$

independently by Theorem 2.1.1, provided the difference $u_i - v_i > 0$ for all i. The aim is to recover the parameter $\gamma = d_\alpha(u, v)$, or equivalently $\theta := \gamma^\alpha = \big[d_\alpha(u, v)\big]^\alpha$, from (a, b). Since $(x_1, \ldots, x_k)$ is a random sample from a distribution with unknown parameter $\gamma = d_\alpha(u, v)$, we are in a position to apply the usual repertoire of statistical estimation techniques to obtain estimators with specified accuracy.

Dimension reduction via random projections can also be applied in the context of streaming data. Consider two data streams with accumulation vectors $\mathbf{a}$ and $\mathbf{b}$ defined over $\mathcal{D}$. Let $h_1, \ldots, h_k$ be hash functions mapping from $\mathcal{D}$ to independent copies of random variables having distribution $F(x; \alpha, \beta, 1, 0)$. Moreover, let $\tilde{\mathbf{a}}$ and $\tilde{\mathbf{b}}$ be k-dimensional sketch vectors representing $\mathbf{a}$ and $\mathbf{b}$, respectively; at $t = 0$, the entries in these vectors are set to 0.

At time $t > 0$, we observe pair (i_{1t}, d_{1t}) in the first stream, and (i_{2t}, d_{2t}) in the second, and update the sketch vectors as follows, for $j = 1, \ldots, k$,

$$\tilde{a}_j = \tilde{a}_j + d_{1t} h_j(i_{1t}), \quad \tilde{b}_j = \tilde{b}_j + d_{2t} h_j(i_{2t}),$$

where $h_j(i_{1t})$ is a random variable from distribution $F(x; \alpha, \beta, 1, 0)$, generated after using i_{1t}

to seed the pseudo-random number generator; $h_j(i_{2t})$ is defined similarly. At time $T > 0$,

$$\tilde{a}_j = \sum_{t=1}^{T} d_{1t} h_j(i_{1t}) = \sum_{i \in \mathcal{D}} a_i h_j(i) \sim F\Big(x; \alpha, \beta, \gamma = \Big[\sum_{i \in \mathcal{D}} |a_i|^{\alpha}\Big]^{1/\alpha}, 0\Big),$$

for $j = 1, \ldots, k$. This is equivalent to multiplying the row vector $\mathbf{a}$ of length $m = |\mathcal{D}|$ (m denotes the size of the universe of data types) by the random projection matrix $\mathbf{P} \in \mathbb{R}^{m \times k}$ with independent entries $p_{ij} \sim F(x; \alpha, \beta, 1, 0)$. It then follows that $\big\{\tilde{a}_j - \tilde{b}_j : z = 1, \ldots, k\big\}$ is a random sample from the distribution $F\big(x; \alpha, \beta, \gamma = d_\alpha(\mathbf{a}, \mathbf{b}), 0\big)$, and the problem reduces to that of estimating the parameter γ, or equivalently, $\theta = \gamma^\alpha$. We point out that $\big[d_\alpha(\mathbf{a}, \mathbf{b})\big]^\alpha$, for $\alpha \leq 1$, is a meaningful measure of distance between data streams. In the extreme case $\alpha \to 0$, $\big[d_\alpha(\mathbf{a}, \mathbf{b})\big]^\alpha$ tends to the number of mismatches between the two sequences.

For the space complexity of estimating the $F_\alpha(\mathbf{a}) = \big(l_\alpha(\mathbf{a})\big)^\alpha$ frequency moments in the non-negative turnstile case, i.e., $a_i \geq 0\ \forall i$, by an (ϵ, δ)-approximation via a one-pass algorithm, Woodruff (2004) proves that the space must be of order $\Omega\big(\epsilon^{-2}\big)$ for any real $\alpha \neq 1$ and any $\epsilon = \Omega\big(m^{-1/2}\big)$, where m is the size of the universe of data types.

Ping Li argues that the space complexity can be reduced to $O\big(\epsilon^{-1}\big)$ when $\alpha = 1 \pm \Delta$, as $\Delta \to 0$, by projecting to linear combinations of skewed stable random variables with $\beta = 1$ (Li, 2008c, 2009), thus significantly reducing the lower bound on the number of projections, k, required for estimating the l_α frequency moment when $\alpha \sim 1$. This approach is discussed further in Section 4.5.

4.3 The method of L-estimation

Chernoff et al. (1967) consider the problem of asymptotically efficient linear estimation of location and scale parameters by weighted linear combinations of ordered statistics. Let $(y_{(1)}, \ldots, y_{(k)})$ be the ordered statistics of a random sample of k observations from a family with c.d.f. and p.d.f. given by $F(y; \mu, \theta) = F_0\big((y - \mu)/\theta\big)$, and $f(y; \mu, \theta) = \theta^{-1} f_0\big((y - \mu)/\theta\big)$, respectively, where μ and θ are location and scale parameters. Consider estimators of the

form

$$T_k = \sum_{i=1}^{k} w_{ik} y_{(i)} = \frac{1}{k} \sum_{i=1}^{k} J\Big(\frac{i}{k+1}\Big) y_{(i)},$$

where $J(u)$ is a function defining the weights w_{ik}; these are called L-estimators. Define

$$L_1(y) = -\frac{f_0'(y)}{f_0(y)}, \quad L_2(y) = -1 - y\frac{f_0'(y)}{f_0(y)}.$$

Under regularity conditions for the validity of the Cramér-Rao lower bound, conditions allowing integration by parts, and conditions of Corollaries 3 and 4 in Chernoff et al. (1967), the Fisher information matrix is of the form $\theta^{-2} I$, where

$$I = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix} = \begin{pmatrix} \int_{-\infty}^{\infty} L_1'(y) f_0(y) dy & \int_{-\infty}^{\infty} L_2'(y) f_0(y) dy \\ \int_{-\infty}^{\infty} y L_1'(y) f_0(y) dy & \int_{-\infty}^{\infty} y L_2'(y) f_0(y) dy \end{pmatrix},$$

$I_{12} = I_{21}$. Moreover, the estimators of location and scale presented in Chernoff et al. (1967) for particular J functions, defined below, are asymptotically efficient.

Most of these conditions are readily verified in our case. Let $H(u)$ denote the inverse function of $F(y; \mu, \theta)$, i.e., $H(u) = \mu + \theta F_0^{-1}(u)$, and $H'(u)du = \theta dy$, $du = f_0(y)dy$, where $y = F_0^{-1}(u)$. Corollaries 3 and 4 in Chernoff et al. (1967) hold under assumptions on the behaviour of $H(\cdot)$ (i.e., $H(\cdot)$ is continuous and satisfies a first-order Lipschitz condition, and $H'(\cdot)$ exists and is continous - these are known as Assumption A*), a tail smoothness assumption (Assumption E) that is satisfied since the ratio $H'(u_1)/H'(u_2)$ is constant, and Assumption B** on the absolute convergence of certain Riemann integrals (satisfied since the integrands are continuous functions, and the intervals of integration are closed and bounded).

It remains to show that $f_0(y)$ exists and is finite, and that conditions allowing integration by parts are satisfied, namely the existence of $f_0''(y)$ and $y^2 f_0'(y) \to 0$ as $y \to \pm\infty$.

For location estimation, assuming θ to be known, the estimator T_k is defined with weights

$$J(u) = \frac{L_1'(F_0^{-1}(u))}{I_{11}} = -\frac{\ell''(F_0^{-1}(u))}{I_{11}}, \tag{4.3}$$

where $\ell(y) = \log f_0(y)$, and, as $k \to \infty$,

$$\sqrt{k}\big(T_k - I_{11}^{-1} I_{12}\theta\big) \xrightarrow{\mathcal{D}} \text{Normal}(\mu, \theta^2/I_{11}).$$

This result also appears in Huber (2004), and can be derived by linearising the MLE as follows. For simplicity, let $\theta = 1$. The log-likelihood is

$$\ell(\mu; y_1, \ldots, y_k) = \log \prod_{i=1}^{k} f_0(y_i - \mu) = \sum_{i=1}^{k} \log f_0(y_i - \mu),$$

and the likelihood equation

$$\frac{\partial}{\partial \mu} \sum_{i=1}^{k} \log f_0(y_i - \mu) = 0.$$

The inverse empirical distribution function of the sample is defined as

$$H_Y(u) = y_{(i)}, \text{ if } (i-1)/k < u \leq i/k, \ i = 1, \ldots, k.$$

Then, we notice that

$$\int_0^1 \log f_0\big(H_Y(u) - \mu\big) du = \frac{1}{k} \sum_{i=1}^{k} \log f_0(y_i - \mu),$$

and from the likelihood equation, we have that

$$\frac{\partial}{\partial \mu} \int_0^1 \log f_0\big(H_Y(u) - \mu\big) du = 0.$$

By a first-order Taylor expansion around the population quantile $F^{-1}(u)$, we obtain

$$\begin{aligned}
0 &= \int_0^1 \ell'\big(H_Y(u) - \mu\big) du \\
&\approx \int_0^1 \ell'\big(F^{-1}(u) - \mu\big) du + \int_0^1 \big[H_Y(u) - F^{-1}(u)\big] \ell''\big(F^{-1}(u) - \mu\big) du \\
&= \int_0^1 \ell'\big(F_0^{-1}(u)\big) du + \int_0^1 \big[H_Y(u) - F^{-1}(u)\big] \ell''\big(F_0^{-1}(u)\big) du \\
&= \int_{-\infty}^{\infty} \ell'(y) f_0(y) dy + \int_0^1 H_Y(u) \ell''\big(F_0^{-1}(u)\big) du - \int_{-\infty}^{\infty} (y + \mu) \ell''(y) f_0(y) dy \\
&= \sum_{i=1}^{k} y_{(i)} \int_{(i-1)/k}^{i/k} \ell''\big(F_0^{-1}(u)\big) du - \int_{-\infty}^{\infty} y \ell''(y) f_0(y) dy - \mu \int_{-\infty}^{\infty} \ell''(y) f_0(y) dy \\
&= \sum_{i=1}^{k} y_{(i)} \int_{F_0^{-1}((i-1)/k)}^{F_0^{-1}(i/k)} \ell''(y) f_0(y) dy - \int_{-\infty}^{\infty} y \ell''(y) f_0(y) dy + \mu I_{11}.
\end{aligned}$$

So, the MLE of μ can be approximated by

$$\begin{aligned}\hat{\mu} &\approx I_{11}^{-1}\Big\{\int_{-\infty}^{\infty} y\ell''(y)f_0(y)dy - \sum_{i=1}^{k} y_{(i)}\int_{F_0^{-1}((i-1)/k)}^{F_0^{-1}(i/k)} \ell''(y)f_0(y)dy\Big\} \\ &\approx -\frac{1}{k}\sum_{i=1}^{k} I_{11}^{-1}\ell''\Big(F_0^{-1}\Big(\frac{i}{k+1}\Big)\Big)y_{(i)} - I_{11}^{-1}I_{12},\end{aligned}$$

giving the L-estimator T_k with weights (4.3).

For scale estimation, assuming μ to be known, T_k is defined with weights

$$J(u) = \frac{L_2'(F_0^{-1}(u))}{I_{22}} = -\frac{\ell'(F_0^{-1}(u)) + F_0^{-1}(u)\ell''(F_0^{-1}(u))}{I_{22}}, \tag{4.4}$$

and, as $k \to \infty$,

$$\sqrt{k}\big(T_k - I_{22}^{-1}I_{12}\mu\big) \xrightarrow{\mathcal{D}} \text{Normal}(\theta, \theta^2/I_{22}).$$

4.4 Symmetric stable random projections

4.4.1 Introduction

Let the entries of $\mathbf{P}$ be independent, symmetric, strictly stable random variables of index α, skewness $\beta = 0$, scale $\gamma = 1$, and location $\delta = 0$, i.e., having c.f. $e^{-|t|^\alpha}$, for $-\infty < t < \infty$. The requirement in Theorem 2.1.1 that the constants in the linear combination be positive is not needed in the symmetric case. We have a random sample $(x_1, \ldots, x_k) \sim F(x; \alpha, 0, \gamma, 0)$.

Li (2008b) and Li and Hastie (2008) propose various estimators of $\theta = \gamma^\alpha$, derived from the following statistical result: if $X \sim F(x; \alpha, 0, \gamma, 0)$, then for $-1 < \lambda < \alpha$, $\mathbb{E}|X|^\lambda = \theta^{\lambda/\alpha} 2/\pi\Gamma(1 - \lambda/\alpha)\Gamma(\lambda)\sin(\pi\lambda/2)$. Li (2008b) proposes the geometric mean estimator for $0.344 < \alpha < 2$, and the harmonic mean estimator for $\alpha \leq 0.344$. Combined, these estimators have asymptotic relative efficiency exceeding 70% and increasing to 100% as $\alpha \to 0$. Moreover, the geometric mean estimator has exponentially decreasing tail bounds, as does the harmonic mean estimator in the limit as $\alpha \to 0$.

Li and Hastie (2008) define the fractional power estimator (proposed previously by Nikias and Shao (1995)):

$$\begin{aligned}\hat{\theta} &= \left(\frac{k^{-1}\sum_{z=1}^{k}|x_z|^{\lambda\alpha}}{2/\pi\Gamma(1-\lambda)\Gamma(\lambda\alpha)\sin(\pi\lambda\alpha/2)}\right)^{1/\lambda} \times \\ &\left(1-\frac{1}{2k\lambda}\left(\frac{1}{\lambda}-1\right)\left(\frac{2/\pi\Gamma(1-2\lambda)\Gamma(2\lambda\alpha)\sin(\pi\lambda\alpha)}{[2/\pi\Gamma(1-\lambda)\Gamma(\lambda\alpha)\sin(\pi/2\lambda\alpha)]^2}-1\right)\right), \qquad (4.5)\end{aligned}$$

where λ is chosen to minimise the asymptotic variance which is a convex function; this estimator is unbiased up to terms of order $O(1/k^2)$. Moreover, it outperforms both the geometric mean and the harmonic mean estimators in terms of ARE, which exceeds 75% for the entire range of α, increasing to 100% as $\alpha \to 0$; its small sample performance is good for values of k as small as 10, unless α is close to 2. For fixed α, evaluating the estimator in (4.5) requires finding λ, and pre-computing the constant term, so the computation only involves the sum $\left(\sum_{z=1}^{k}|x_z|^{\lambda\alpha}\right)^{1/\lambda}$. The main drawback of the fractional power estimator is the fact that it does not have exponentially decreasing tail bounds, so large mean square errors can be obtained in small samples, particularly for α close to 2.

Finally, Li (2008a) propose an asymptotically unbiased estimator called the optimal quantile estimator, defined as follows:

$$\hat{\theta} = \frac{1}{B_{\alpha,k}} \times \left(\frac{q-Quantile\{|x_j|, j=1,\ldots,k\}}{q-Quantile\{|F(x;\alpha,0,1,0)|\}}\right)^{\alpha}, \qquad (4.6)$$

where q is chosen to minimise the asymptotic variance, and $B_{\alpha,k}$ is the small sample bias correction factor. The latter is estimated via Monte Carlo simulations as it is not possible to find an approximate expression for it by Taylor expansions. When $\alpha < 1$, it performs as well as the geometric mean estimator in terms of ARE, and slightly better than the fractional power estimator when $1 < \alpha \leq 1.8$. The estimator in (4.6) has two advantages over previous estimators; first, it has exponentially decreasing tail bounds that are exact, not approximate. Li (2008a) expresses tail probabilities involving the empirical quantile function as tail probabililites involving the empirical c.d.f., which follows the Binomial distribution, and uses the binomial Chernoff bounds to derive the tail bounds. Second, the optimal quantile estimator

is more efficient computationally than the geometric mean, harmonic mean, and fractional power estimators, in terms of computing time (up to orders of one magnitude), as it involves a single fractional power computation (which dominates the computing time).

Next, we apply the theory of L-estimation to the problem of approximating the parameter θ, or equivalently, the scale parameter $\gamma = \theta^{1/\alpha}$.

4.4.2 Location parameter estimation

Estimation Consider the random sample $x_1, \ldots, x_k \sim f(x; \alpha, 0, \gamma, 0)$. Define

$$y_i := \log |x_i| \stackrel{\mathcal{D}}{=} \log \gamma + z_i := \mu + z_i, \quad i = 1, \ldots, k,$$

where z_i is distributed as the logarithm of the absolute value of a symmetric, strictly stable random variable of index α and scale 1, and $\mu = \log \gamma$. So, $y_i \sim f_\mu(y) = f_0(y - \mu)$ independently, where the probability density function of z_i is

$$f_0(z) = 2e^z f(e^z; \alpha, 0, 1, 0), \quad -\infty < z < \infty. \tag{4.7}$$

Proposition 4.4.1. *The density function $f_0(z)$ defined in* (4.7) *satisfies the conditions for L-estimation.*

Proof. In the tail as $z \to \infty$, we know from Nolan (2007), page 14, that $f(e^z; \alpha, 0, 1, 0) \sim e^{-z(\alpha+1)}$, so $f_0(z) \sim e^{-\alpha z} \to 0$. Moreover, using the same approximation, $z^2 f_0'(z) \sim z^2 e^{-\alpha z} \to 0$ as $z \to \infty$. It is more involved to show that $z^2 f_0'(z) \to 0$ and $f_0(z) \to 0$ as $z \to -\infty$.

As $z \to -\infty$, $e^z \to 0$, and we employ the asymptotic representation from Theorem 2.5.1 in Zolotarev (1986), page 94, for $\alpha < 1$ and $\beta \neq 1$. We obtain that

$$f(e^z; \alpha, 0, 1, 0) \sim \frac{1}{\pi\alpha}\big[\Gamma(\alpha^{-1}) - 0.5\Gamma(3/\alpha)e^{2z}\big],$$

and

$$z^2 f_0'(z) \sim 2e^z z^2 \frac{1}{\pi\alpha}\big[\Gamma(\alpha^{-1}) - 0.5\Gamma(3/\alpha)e^{2z}\big] \sim z^2 e^z \to 0 \quad \text{as } z \to -\infty.$$

In the case $\alpha > 1$, we use the reflection property in equation (2.5.5), page 94, to show that as $e^z \to 0$, $f(e^z; \alpha, 0, 1, 0)$ is constant, so $z^2 f_0'(z) \sim z^2 e^z \to 0$ as $z \to -\infty$. It follows that as $z \to -\infty$, $f(e^z; \alpha, 0, 1, 0)$ is constant, so $f_0(z) \sim e^z \to 0$. □

Having verified the conditions, the L-estimator of the location parameter μ (scale parameter 1) with weights given by formula (4.3) is

$$\hat{\mu} = \sum_{i=1}^{k} w_{ik} y_{(i)} = -\frac{1}{kI_{11}} \sum_{i=1}^{k} \ell''\Big(F_0^{-1}\Big(\frac{i}{k+1}\Big)\Big) y_{(i)},$$

where $I_{11} = I_\mu(F_\mu)$ is the Fisher information about μ contained in $(y_1, \ldots, y_k)$.

Proposition 4.4.2. *The L-estimator $\sum_{i=1}^{k} w_{ik} Y_{(i)}$ has finite variance.*

Proof. The L-estimator satisfies

$$\Big|\sum_{i=1}^{k} w_{ik} Y_{(i)}\Big| \le \sum_{i=1}^{k} |w_{ik} Y_{(i)}| \le k \max\{|w_{ik}|, i = 1, \ldots, k\} \max\{|Y_{(i)}|, i = 1, \ldots, k\},$$

so we need to show that $Y_{(i)}$ has finite variance. Hence we're interested in the behaviour of the ith order statistic in the tails as $y \to \pm\infty$. From Proposition 4.4.1, the p.d.f. is

$$f_\mu(y) \sim \begin{cases} \exp(-y\alpha) & \text{as } y \to \infty \\ \exp(y) & \text{as } y \to -\infty. \end{cases}$$

For the ith order statistic,

$$f_{(i)}(y) = \frac{k!}{(i-1)!(k-i)!} \big[F_\mu(y)\big]^{i-1} \times \big[1 - F_\mu(y)\big]^{k-i} f_\mu(y).$$

As $y \to \infty$, $F_\mu(y) \sim 1 - \alpha^{-1} e^{-\alpha y} \to 1$, so $f_{(i)}(y) \sim e^{-\alpha y(k-i+1)}$, and $\int_0^\infty y^2 e^{-\alpha y(k-i+1)} dy = 2/[\alpha(k-i+1)]^{-3} < \infty$. As $y \to -\infty$, $F_\mu(y) \sim e^y \to 0$, so $f_{(i)}(y) \sim e^{iy}$, and $\int_{-\infty}^0 y^2 e^{iy} dy < \infty$. Therefore, $Y_{(i)}$ has finite variance, and so does the L-estimator. □

Now,

$$I_\mu(F_\mu) = -\int_0^1 \frac{\partial^2}{\partial \mu^2} \log f_0(F_0^{-1}(t)) dt = -\int_0^1 \ell''(F_0^{-1}(t)) dt,$$

so $-k^{-1}\sum_{i=1}^{k}\ell''(F_0^{-1}(i/(k+1))) \to I_{11}$ as $k \to \infty$, and, we can assume that the weights sum to 1 in large samples.

Moreover, computing w_{ik} requires solving $F_0(z) = i/(k+1)$ for z, i.e., $z = F_0^{-1}(i/(k+1)) := \inf\{s; F_0^{-1}(s) \geq i/(k+1)\}$. Let $W \sim f(w;\alpha,0,1,0)$. Now, $F_0(z) = P(Z \leq z) = i/(k+1)$ if and only if $P(W \leq e^z) = 0.5 + i/[2(k+1)]$, i.e., e^z is the $\big(0.5 + i/[2(k+1)]\big)$ quantile of the symmetric, strictly stable distribution of index α, $\gamma = 1$, $\delta = 0$.

In finite samples, we approximate the weights

$$w_{ik} = \frac{\ell''(F_0^{-1}(i/(k+1)))}{\sum_{j=1}^{k}\ell''(F_0^{-1}(j/(k+1)))}, \tag{4.8}$$

and the estimator of μ becomes

$$\hat{\mu} = \sum_{i=1}^{k}\frac{\ell''(F_0^{-1}(i/(k+1)))}{\sum_{j=1}^{k}\ell''(F_0^{-1}(j/(k+1)))}y_{(i)},$$

with expectation

$$\mathbb{E}(\hat{\mu}) = \mu + \sum_{i=1}^{k}\frac{\ell''(F_0^{-1}(i/(k+1)))}{\sum_{j=1}^{k}\ell''(F_0^{-1}(j/(k+1)))}\mathbb{E}Z_{(i)} \approx \mu + \sum_{i=1}^{k}\frac{F_0^{-1}(i/(k+1))\ell''(F_0^{-1}(i/(k+1)))}{\sum_{j=1}^{k}\ell''(F_0^{-1}(j/(k+1)))},$$

where $\mathbb{E}Z_{(i)} \approx F_0^{-1}(i/(k+1))$ in large samples by a first order Taylor approximation. In the limit as $k \to \infty$, the expected value converges as follows

$$\mathbb{E}(\hat{\mu}) \to \mu + \frac{\int_0^1 F_0^{-1}(t)\ell''(F_0^{-1}(t))dt}{\int_0^1 \ell''(F_0^{-1}(t))dt} = \mu - \frac{1}{I_\mu(F_\mu)}\int_{-\infty}^{\infty} z\ell''(z)f_0(z)dz = \mu + I_{11}^{-1}I_{12},$$

as expected. We define the asymptotic bias-correction term

$$BC = \mathbb{E}(\hat{\mu}) - \mu = -\frac{1}{I_\mu(F_\mu)}\int_{-\infty}^{\infty} z\ell''(z)f_0(z)dz,$$

and propose the limiting bias-corrected estimator of μ

$$\hat{\mu}_{BC} = \sum_{i=1}^{k} w_{ik}y_{(i)} - BC, \tag{4.9}$$

that satisfies the following asymptotic result $\sqrt{k}\big(\hat{\mu}_{BC} - \mu\big) \xrightarrow{\mathcal{D}} \text{Normal}\big(0, 1/I_\mu(F_\mu)\big)$ as $k \to \infty$. In finite samples, the bias-corrected estimator becomes

$$\hat{\mu}_{BC} = \sum_{i=1}^{k} w_{ik}\Big(y_{(i)} - F_0^{-1}\Big(\frac{i}{k+1}\Big)\Big), \tag{4.10}$$

with weights w_{ik} given by (4.8).

We have two special cases to consider: $\alpha = 2$ and $\alpha = 1$. First, assume $x_1, \ldots, x_k \sim f(x; 2, 0, \gamma, 0)$, for $\gamma > 0$, i.e., $x_i \sim \text{Normal}(0, 2\theta)$ i.i.d. Then,

$$f_0(y) = \exp(y)/\sqrt{\pi}\exp\big(-\exp(2y)/4\big), \quad -\infty < y < \infty,$$

and $\ell'(y) = 1 - e^{2y}/2$, $\ell''(y) = -e^{2y}$. Hence the Fisher information equals

$$I_{11} = \int_{-\infty}^{\infty} \exp(3y)/\sqrt{\pi}\exp\big(-\exp(2y)/4\big)dy = 2,$$

and the bias is

$$BC = \frac{1}{I_{11}}\int_{-\infty}^{\infty} \frac{y\exp(3y)}{\sqrt{\pi}}\exp\big(-\exp(2y)/4\big)dy = 1 - \frac{\xi}{2} \approx 0.7114,$$

by formulae (3.361) and (3.482) in Gradshteyn and Ryzhik (1980), where ξ is Euler's constant. Moreover, $F_0^{-1}(u) = \log\big(F^{-1}((1+u)/2)\big) = \log\big\{\sqrt{2}\Phi^{-1}((1+u)/2)\big\}$, where F is the c.d.f. of the Normal$(0, 2)$ distribution, so the weight function is given by

$$w_{ik} = \frac{1}{k}\Big[\Phi^{-1}\Big(\frac{1}{2} + \frac{i}{2(k+1)}\Big)\Big]^2.$$

Next, assume that $x_1, \ldots, x_k \sim f(x; 1, 0, \gamma, 0)$, i.e., $x_i \sim \text{Cauchy}(0, \gamma)$ i.i.d. Then,

$$f_0(y) = \frac{2\exp(y)}{\pi(1+\exp(2y))}, \quad -\infty < y < \infty,$$

and, $\ell'(y) = 1 - 2e^{2y}/[1+e^{2y}]$, $\ell''(y) = -4e^{2y}/(1+e^{2y})^2$. So the Fisher information equals

$$I_{11} = \frac{8}{\pi}\int_{-\infty}^{\infty} \frac{\exp(3y)}{[1+\exp(2y)]^3}dy = \frac{1}{2},$$

and the bias is

$$BC = \frac{16}{\pi}\int_{-\infty}^{\infty} \frac{y\exp(3y)}{[1+\exp(2y)]^2}dy = \frac{2}{\pi}\bigg\{\int_{-\infty}^{\infty} \frac{\exp(y)}{[1+\exp(2y)]^2}dy + \int_{-\infty}^{\infty} \frac{y\exp(y)}{[1+\exp(2y)]^2}dy\bigg\} = 0,$$

using formulae (3.241), (4.231) and (8.366) in Gradshteyn and Ryzhik (1980). Moreover, $F_0^{-1}(u) = \log\big(F^{-1}((1+u)/2)\big) = \log\big(\tan(\pi u/2)\big)$, where F is the c.d.f. of the Cauchy$(0, 1)$ distribution. Therefore, the weight function is of the form

$$w_{ik} = \frac{8}{k} \times \frac{\big[\tan(\pi i/[2(k+1)])\big]^2}{\big[1 + \big(\tan(\pi i/[2(k+1)])\big)^2\big]^2}.$$

In the case $\alpha > 1$, we are interested in estimating $\gamma = e^{\mu}$ by $\hat{\gamma} = \exp\big(\hat{\mu}_{BC}\big)$. It follows that $\sqrt{k}\big(\hat{\gamma} - \gamma\big) \xrightarrow{\mathcal{D}} \text{Normal}\big(0, 1/I_{\gamma}(F_{\gamma})\big)$ as $k \to \infty$ by the Delta Method, where $I_{\gamma}(F_{\gamma})$ is the Fisher information about γ contained in $(x_1, \ldots, x_k)$.

To relate the Fisher information about μ contained in y_1, $I_{\mu}(F_{\mu})$, to the Fisher information about γ contained in x_1, $I_{\gamma}(F_{\gamma})$, where $F_{\gamma}(x) = F(x; \alpha, 0, \gamma, 0)$, we remark that

$$\begin{aligned}
\big(\log f_{\mu}(y)\big)'' &= \frac{\partial^2}{\partial \mu^2} \log f\big(e^y; \alpha, 0, e^{\mu}, 0\big) \\
&= \frac{\partial^2}{\partial \mu \partial \gamma} \log f\big(e^y; \alpha, 0, \gamma, 0\big) \times \frac{d\gamma}{d\mu} \\
&= \frac{\partial}{\partial \mu} \left\{ e^{\mu} \frac{d}{d\gamma} \log f\big(e^y; \alpha, 0, \gamma, 0\big) \right\} \\
&= \gamma^2 \frac{\partial^2}{\partial \gamma^2} \log f\big(e^y; \alpha, 0, \gamma, 0\big) + \gamma \frac{\partial}{\partial \gamma} \log f\big(e^y; \alpha, 0, \gamma, 0\big). \qquad (4.11)
\end{aligned}$$

Taking expectations on both sides with respect to $Y \sim f_{\mu}(y)$, we obtain

$$\begin{aligned}
I_{\mu}(F_{\mu}) &= \gamma^2 \mathbb{E}_Y \Big(- \frac{\partial^2}{\partial \gamma^2} \log f\big(e^y; \alpha, 0, \gamma, 0\big) \Big) \\
&= -\gamma^2 \int_{-\infty}^{\infty} \frac{\partial^2}{\partial \gamma^2} \log f\big(e^y; \alpha, 0, \gamma, 0\big) 2e^y f\big(e^y; \alpha, 0, \gamma, 0\big) dy = \gamma^2 I_{\gamma}(F_{\gamma}).
\end{aligned}$$

Another way of showing that $I_{\mu}(F_{\mu}) = \gamma^2 I_{\gamma}(F_{\gamma})$ proceeds as follows. Let $\hat{\mu}_{MLE}$ denote the MLE of μ, which, under regularity conditions, is asymptotically normally distributed with mean μ and variance $1/[kI_{\mu}(F_{\mu})]$. It follows by the Delta Method that the MLE of γ, $\exp(\hat{\mu}_{MLE})$, satisfies $\sqrt{k}\big\{ \exp(\hat{\mu}_{MLE}) - \exp(\mu) \big\} \xrightarrow{\mathcal{D}} \text{Normal}\big(0, \gamma^2/I_{\mu}(F_{\mu})\big)$ as $k \to \infty$. So, $I_{\mu}(F_{\mu}) = \gamma^2 I_{\gamma}(F_{\gamma})$.

By a Taylor expansion around μ,

$$\hat{\gamma} = \exp\big(\hat{\mu}_{BC}\big) = e^{\mu} + e^{\mu}\big(\hat{\mu}_{BC} - \mu\big) + \frac{e^{\mu}}{2}\big(\hat{\mu}_{BC} - \mu\big)^2 + \frac{e^{\mu}}{6}\big(\hat{\mu}_{BC} - \mu\big)^3 + \sum_{i=4}^{\infty} \frac{e^{\mu}}{i!}\big(\hat{\mu}_{BC} - \mu\big)^i.$$

Taking expectations, and using the fact that the i*th* central moment of a normal random variable is 0, if i is odd, and equals a constant times the i*th* power of the standard deviation,

if i is even, we show that the bias incurred by exponentiating is

$$\begin{aligned}\mathbb{E}(\hat{\gamma}) - \gamma &= \frac{\gamma}{2}\mathbb{E}\big(\hat{\mu}_{BC} - \mu\big)^2 + \frac{\gamma}{8}\mathbb{E}\big(\hat{\mu}_{BC} - \mu\big)^4 + \sum_{i=3}^{\infty} \frac{\gamma}{(2i)!}\mathbb{E}\big(\hat{\mu}_{BC} - \mu\big)^{2i} \\ &\approx \gamma\bigg(\frac{1}{2kI_\mu(F_\mu)} + \frac{1}{8k^2[I_\mu(F_\mu)]^2} + O\Big(\frac{1}{k^4}\Big)\bigg).\end{aligned}$$

In finite samples, the bias-corrected estimator of γ

$$\hat{\gamma}_{BC} = \hat{\gamma}\bigg(1 - \frac{1}{2kI_\mu(F_\mu)}\bigg) = \exp\bigg\{\sum_{i=1}^{k} w_{ik}\Big(y_{(i)} - F_0^{-1}\Big(\frac{i}{k+1}\Big)\Big)\bigg\}\bigg[1 - \frac{1}{2kI_\mu(F_\mu)}\bigg] \quad (4.12)$$

is unbiased up to terms of order $O(1/k^2)$, where $\hat{\gamma} = \exp(\hat{\mu}_{BC})$ using formulation (4.10).

Similar calculations provide an asymptotically efficient estimator for $\theta(p) = \gamma^p$, $0 < p < 1$, a more relevant parameter to estimate when $\alpha < 1$,

$$\hat{\theta}_{BC}(p) = \exp\big(p\hat{\mu}_{BC}\big)\bigg(1 - \frac{p^2}{2kI_\mu(F_\mu)}\bigg);$$

in finite samples, the approximate estimator becomes

$$\hat{\theta}_{BC}(p) = \exp\bigg\{p\sum_{i=1}^{k} w_{ik}\Big(y_{(i)} - F_0^{-1}\Big(\frac{i}{k+1}\Big)\Big)\bigg\}\bigg[1 - \frac{p^2}{2kI_\mu(F_\mu)}\bigg]. \quad (4.13)$$

Next, we improve the unbiased estimators $\hat{\theta}_{BC}$ and $\hat{\gamma}_{BC}$ by replacing the asymptotic bias term in expression (4.9) by a second order Taylor approximation as follows.

$$\hat{\mu} = \sum_{i=1}^{k} w_{ik} y_{(i)} = \mu\sum_{i=1}^{k} w_{ik} + \sum_{i=1}^{k} w_{ik} z_{(i)} = \mu + \sum_{i=1}^{k} w_{ik} z_{(i)},$$

since the weights are normalised to add to 1, and the bias term equals

$$BC = \mathbb{E}\hat{\mu} - \mu = \sum_{i=1}^{k} w_{ik}\mathbb{E}Z_{(i)} = \sum_{i=1}^{k} w_{ik}\mathbb{E}F_0^{-1}(U_{(i)}),$$

where $U_i \sim \text{Unif}(0,1)$ i.i.d. for $i = 1, \ldots, k$. By a second order Taylor expansion of F_0^{-1} around $i/(k+1)$, we have

$$F_0^{-1}(U_{(i)}) = F_0^{-1}\Big(\frac{i}{k+1}\Big) + \Big(U_{(i)} - \frac{i}{k+1}\Big)\frac{dF_0^{-1}(y)}{dy}\Big|_{y=\frac{i}{k+1}} + \frac{1}{2}\Big(U_{(i)} - \frac{i}{k+1}\Big)^2\frac{d^2F_0^{-1}(y)}{dy^2}\Big|_{y=\frac{i}{k+1}},$$

and

$$\mathbb{E}F_0^{-1}(U_{(i)}) = F_0^{-1}\Big(\frac{i}{k+1}\Big) + \frac{i(k-i+1)}{2(k+1)^2(k+2)} \times \frac{d^2}{dy^2}F_0^{-1}(y)\Big|_{y=\frac{i}{k+1}}. \tag{4.14}$$

Letting $u = F_0^{-1}(y)$ such that $y = F_0(u)$, it can be shown easily that

$$\frac{du}{dy} = \frac{1}{f_0(F_0^{-1}(y))}, \quad \frac{d^2u}{dy^2} = -\frac{f_0'(u)}{[f_0(u)]^3}.$$

Hence

$$\frac{d^2}{dy^2}F_0^{-1}(y)\Big|_{y=\frac{i}{k+1}} = -\frac{\ell'\big(F_0^{-1}(i/(k+1))\big)}{\big[f_0\big(F_0^{-1}(i/(k+1))\big)\big]^2},$$

and the bias correction term becomes

$$BC = \sum_{i=1}^{k} w_{ik}\Bigg\{F_0^{-1}\Big(\frac{i}{k+1}\Big) - \frac{i(k-i+1)}{2(k+1)^2(k+2)} \times \frac{\ell'\big(F_0^{-1}(i/(k+1))\big)}{\big[f_0\big(F_0^{-1}(i/(k+1))\big)\big]^2}\Bigg\}.$$

The improved bias-corrected estimator of μ is

$$\hat{\mu}_{BC} = \sum_{i=1}^{k} w_{ik}\Bigg\{y_{(i)} - F_0^{-1}\Big(\frac{i}{k+1}\Big) + \frac{i(k-i+1)}{2(k+1)^2(k+2)} \times \frac{\ell'\big(F_0^{-1}(i/(k+1))\big)}{\big[f_0\big(F_0^{-1}(i/(k+1))\big)\big]^2}\Bigg\}, \tag{4.15}$$

and the corresponding estimators of γ and θ follow. Let the estimators of γ and θ in (4.12) and (4.13), respectively, be called version 1, if computed with fBasics commands, and version 2, if computed with our numerically stable commands. Let the improved bias-corrected estimators of γ and θ, derived from expression (4.15), computed with our commands be called version 3 estimators. We proceed to compare the performance of these three estimators to that of the fractional power estimator in terms of mean square error.

Computation In computing $\hat{\gamma}_{BC}$ and $\hat{\theta}_{BC}$, the evaluation of $\ell''(x) = \big(\log f_\mu(x)\big)''\big|_{\mu=0}$ and $F_0^{-1}(x)$ is required. The latter is related to the quantile function of the $f(x;\alpha,0,1,0)$ density as follows: $F_0(z) = x$ if and only if $P(W \le e^z) = (1+x)/2$, where $W \sim f(w;\alpha,0,1,0)$. So,

$$F_0^{-1}(x) = \log\Big(F^{-1}\Big(\frac{1+x}{2};\alpha,0,1,0\Big)\Big),$$

and the quantile can be obtained approximately using the *qstable* function in fBasics. We estimate $\ell''(x)$ by a second order finite difference scheme by

$$\begin{aligned} \ell''(x) &\approx \frac{1}{h}\left[\frac{\ell(x+h)-\ell(x)}{h} - \frac{\ell(x)-\ell(x-h)}{h}\right] \\ &= h^{-2}\big[\ell(x+h) - 2\ell(x) + \ell(x-h)\big], \end{aligned} \tag{4.16}$$

with grid width $h = 0.01$, where $\ell(x) = \log\big(2e^x f(e^x; \alpha, 0, 1, 0)\big)$.

Table 4.1 gives the Fisher information, $I_\mu(F_\mu)$, for various α, computed with the fBasics functions *dstable* and *qstable*; $I_\mu(F_\mu)$ for $\alpha = 1, 2$ is exact. Given $\delta, n > 0$,

$$\begin{aligned} I_\mu(F_\mu) &= -\int_0^1 \ell''\big(F_0^{-1}(t)\big)dt \\ &\approx -\frac{h}{2}\sum_{i=1}^{n}\Big[\ell''\big(F_0^{-1}(\delta+(i-1)h)\big) + \ell''\big(F_0^{-1}(\delta+ih)\big)\Big], \end{aligned} \tag{4.17}$$

where the trapezoid rule is employed to approximate the integral on $[\delta, 1-\delta]$ with n subintervals of width $h = (1-2\delta)/n$. Figure 4.1 displays the result for various α.

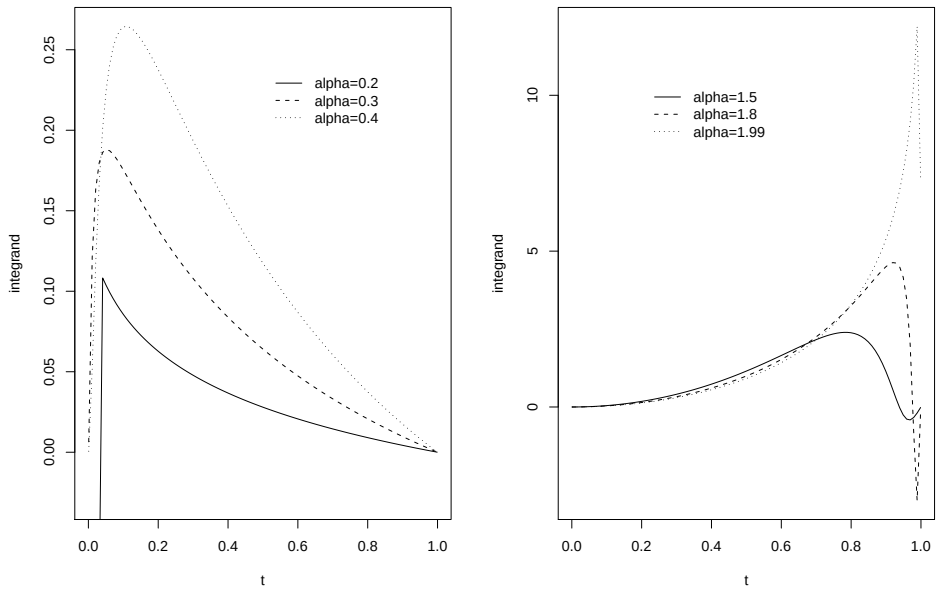


Figure 4.1: Approximations to the integrand in expression (4.17) ($\delta = 0.001$, $n = 100$).

α	$I_\mu(F_\mu)$	$I_\mu(F_\mu)^\star$	MT	Nolan	α	$I_\mu(F_\mu)$	$I_\mu(F_\mu)^\star$	MT	Nolan
0.2	0.0363	0.0367	0.0367		1.15	0.6182	0.6181		
0.25	0.0547	0.0549			1.2	0.6604	0.6603	0.6603	0.6142
0.3	0.0755	0.0756	0.0756		1.25	0.7042	0.7042		
0.35	0.982	0.0983			1.3	0.7499	0.7498	0.7498	0.6875
0.4	0.1226	0.1226	0.1226		1.35	0.7976	0.7975		
0.45	0.1483	0.1483			1.4	0.8476	0.8475	0.8475	0.7668
0.5	0.1753	0.1753	0.1753	0.1240	1.45	0.9002	0.9000		
0.55	0.2034	0.2034			1.5	0.9558	0.9556	0.9556	0.8542
0.6	0.2325	0.2325	0.2325	0.1919	1.55	1.0148	1.0145		
0.65	0.2626	0.2626			1.6	1.0780	1.0775	1.0775	0.9537
0.7	0.2937	0.2937	0.2937	0.2660	1.65	1.1459	1.1453		
0.75	0.3256	0.3257			1.7	1.2198	1.2189	1.2189	1.0738
0.8	0.3585	0.3586	0.3586	0.3392	1.75	1.3011	1.2998		
0.85	0.3924	0.3924			1.8	1.3920	1.3898	1.3898	1.2237
0.9	0.4272	0.4272	0.4272	0.4093	1.85	1.4968	1.4922		
0.95	0.4631	0.4631	0.4631		1.9	1.6270	1.6127	1.6127	1.4345
1.0	0.5	0.5	0.5	0.4698	1.95	1.7882	1.7631	1.7631	1.5982
1.05	0.5379	0.5381	0.5381		1.99	1.8861	1.9322	1.9321	1.8361
1.1	0.5774	0.5774	0.5774	0.5446	2.0	2.0	2.0	2.0	2.0006

Table 4.1: Approximate Fisher information about μ tabulated for $\alpha \in [0.2, 2]$. The values in columns $I_\mu(F_\mu)$ and $I_\mu(F_\mu)^\star$ are obtained via the finite difference scheme using fBasics functions ($n = 1000$, $\delta = 0.005$), and our improved versions ($n = 1000$, $\delta = 0.001$), respectively; those in MT and Nolan appear in Matsui and Takemura (2006), and Nolan (2001).

The values in column $I_\mu(F_\mu)$ of Table 4.1 agree with those presented by Matsui and Takemura (2006) (column MT) to within 3-4 significant digits for $\alpha \in (0.3, 1.8)$, but appear to differ slightly more for α outside this range; for example, for $\alpha = 1.8$, our estimate is 1.3920, whereas that of Matsui and Takemura (2006) is 1.3898. However, the Fisher information values presented by Nolan (2001) (column Nolan) are strikingly different. They were obtained by numerically computing the integral defining the Fisher information, and the partial derivatives of the density; the paper offers no further details. We try to reproduce the values in column MT by implementing the approximation formulas in Matsui and Takemura (2006), and using *qstable* for quantile estimation. We omit these results, but point out that they were different than expected for $\alpha \approx 1, 2$, possibly due to the instability of the fBasics functions identified in Chapter 2.

Column $I_\mu(F_\mu)^\star$ of Table 4.1 presents the estimates of Fisher information obtained by our improved R functions for density, distribution, and quantile estimation. These estimates agree with those of Matsui and Takemura (2006) to four decimal places for all α, with the exception of $\alpha = 1.99$, where we report a value of 1.9322, compared to 1.9321. The values reported by Nolan (2001) are significantly different and we dismiss them as incorrect.

In recomputing the Fisher information to obtain the values in column $I_\mu(F_\mu)^\star$, we analyse the shape of the integrand function in Figure 4.1. We suspect that the integration procedure misses the sharp drop at t close to 0 and $\alpha < 0.4$, and again at t close to 1 and $\alpha \geq 1.8$, underestimating the integral. For $\alpha < 0.4$, we transform from t to $\sqrt{t}$; Figure 4.2 displays the tranformed integrand. For $\alpha \geq 1.8$, we use a smaller width $h = (1 - 2\delta)/(10n)$ in the quadrature rule integration algorithm over the region $t > 0.8$.

Moreover, we extrapolate the value of Fisher information for $\alpha < 0.2$ based on the estimates in column $I_\mu(F_\mu)^\star$ of Table 4.1. Table 4.2 summarises our results. We model the Fisher information as a function of α via four different no-intercept models using least squares, and choose the quadratic model with no linear term. We justify fitting no-intercept

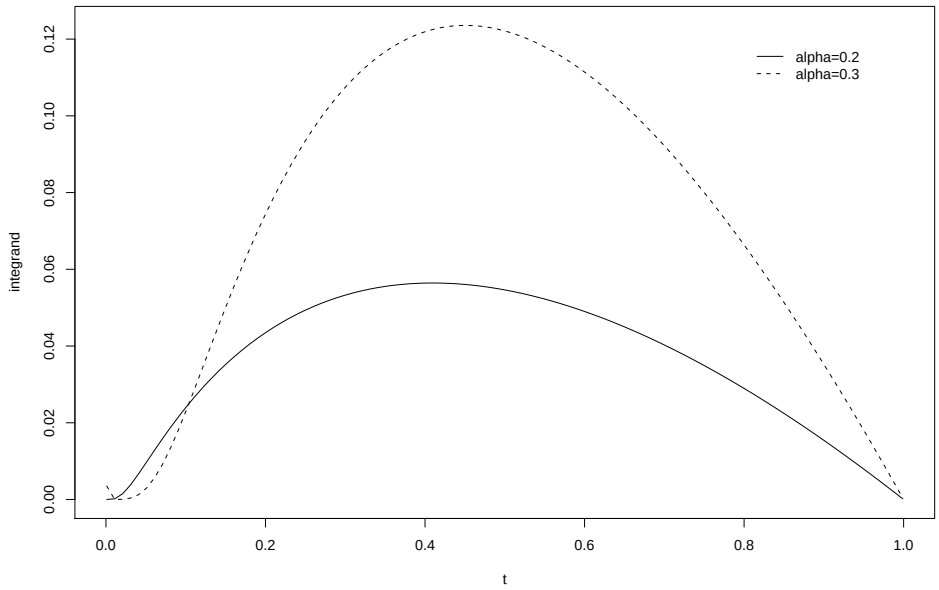


Figure 4.2: Approximations to the transformed integrand in the expression giving the Fisher information (4.17) with $\delta = 0.001$ and $n = 100$.

models in Section 4.4.3, showing that $I_\mu(F_\mu) \to 0$ as $\alpha \to 0$. Figure 4.3 displays the data and the fitted curves of the models.

α	$I_\mu(F_\mu)$	prediction interval	α	$I_\mu(F_\mu)$	prediction interval
0.01	4.523×10^{-5}	(-0.1372, 0.1373)	0.1	4.523×10^{-3}	(-0.1328, 0.1418)
0.05	1.131×10^{-3}	(-0.1362, 0.1384)	0.15	1.018×10^{-2}	(-0.1271, 0.1475)

Table 4.2: Predicted values of Fisher information for $\alpha = 0.01, 0.05, 0.1, 0.15$ from no-intercept model $I_\mu(F_\mu) = 0.4523 \times \alpha^2$; the model is fitted by least squares.

The linear model $I_\mu(F_\mu) \sim \alpha$ explains about 94.4% of the variability in the data; adding the quadratic term α^2 increases the R^2 to 0.995, but the linear term α is no longer significant at 0.95 level. We fit the quadratic model with linear term removed; the R^2 remains unchanged. We also fit the cubic model with all lower order terms, except the intercept; all

variables are significant and the R^2 rises to 0.998. Our final model is the quadratic model with no linear term, as we believe that the cubic one overfits the data. The predicted values and prediction intervals in Table 4.2 are obtained from the final model.

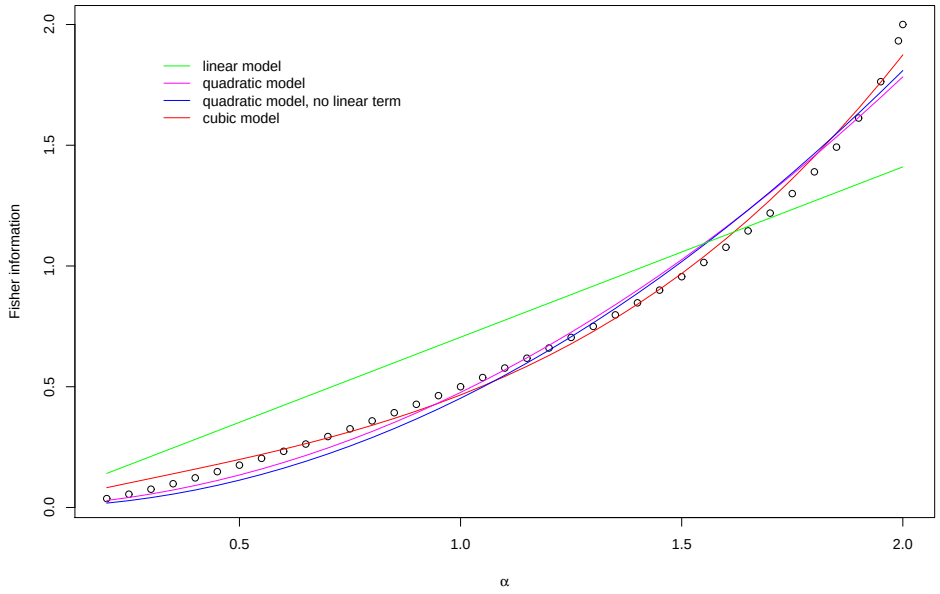


Figure 4.3: Fitted models of Fisher information as a function of α; the black circles are the data from column $I_\mu(F_\mu)^\star$ of Table 4.1.

Similarly, the asymptotic bias is estimated by

$$\begin{aligned} BC &= -\frac{1}{I_\mu(F_\mu)} \int_0^1 F_0^{-1}(t)\ell''\big(F_0^{-1}(t)\big)dt \\ &\approx \frac{-h}{2\hat{I}_\mu(F_\mu)} \sum_{i=1}^{n} \Big[F_0^{-1}(\delta+ih-h)\ell''\big(F_0^{-1}(\delta+ih-h)\big) + F_0^{-1}(\delta+ih)\ell''\big(F_0^{-1}(\delta+ih)\big)\Big], \end{aligned}$$

where $\hat{I}_\mu(F_\mu)$ is the Fisher information estimate. For $\alpha \geq 1.5$ and $t > 0.8$, we use width $h = (1-2\delta)/(10n)$ with $n = 1000$, $\delta = 0.001$; we perform no transformations on the integrand function. Table 4.3 gives approximate values of the bias for various $\alpha \in [0.2, 2.0]$ computed using the improved version of the fBasics functions; the bias for $\alpha = 1, 2$ is exact.

Figure 4.4 displays the weight function in (4.8) for various α computed using the second order finite difference scheme ($h = 0.01$) and the improved version of the fBasics commands.

α	BC	α	BC	α	BC	α	BC
0.2	-2.3005	0.7	-0.3035	1.2	0.1405	1.7	0.4383
0.25	-1.7918	0.75	-0.2399	1.25	0.1719	1.75	0.4702
0.3	-1.4335	0.8	-0.1829	1.3	0.2024	1.8	0.5042
0.35	-1.1671	0.85	-0.1312	1.35	0.2321	1.85	0.5411
0.4	-0.9617	0.9	-0.0840	1.4	0.2614	1.9	0.5824
0.45	-0.7979	0.95	-0.0405	1.45	0.2904	1.95	0.6318
0.5	-0.6640	1.0	0	1.5	0.3190	1.99	0.6870
0.55	-0.5523	1.05	0.0380	1.55	0.3483	2.0	0.7114
0.6	-0.4573	1.1	0.0738	1.6	0.3751		
0.65	-0.3753	1.15	0.1078	1.65	0.4073		

Table 4.3: Approximate asymptotic bias tabulated for $\alpha \in [0.2, 2]$, obtained via the finite difference scheme using our improved versions of functions in fBasics ($n = 1000$, $\delta = 0.001$).

For α small, the weighted sum in the formulations of $\hat{\gamma}_{BC}$ and $\hat{\theta}_{BC}$ places significant weight on small order statistics, and negligible weight on large ones, gradually shifting the weight balance towards large order statistics as $\alpha \to 2$. For $\alpha = 1.2, 1.5, 1.8$, the value of w_{ik}, for i very close to k, oscillates below zero, with the larger the α, the sharper the oscillation. For $\alpha = 2$, the weight w_{ik} displays a strictly increasing behaviour as $i \to k$.

Numerical results The L-estimator is easily computable as the weights depend only on α and k, and can be tabulated once-and-or-all for any required value of α. This calculation depends on accurate approximations to the quantiles and the density of the symmetric, strictly stable distribution. Whereas it is possible to obtain a good approximation to the MLE via an iterative procedure with a suitably large table of pre-calculated derivatives for fixed α, the L-estimation procedure has the advantage of achieving the same asymptotic performance

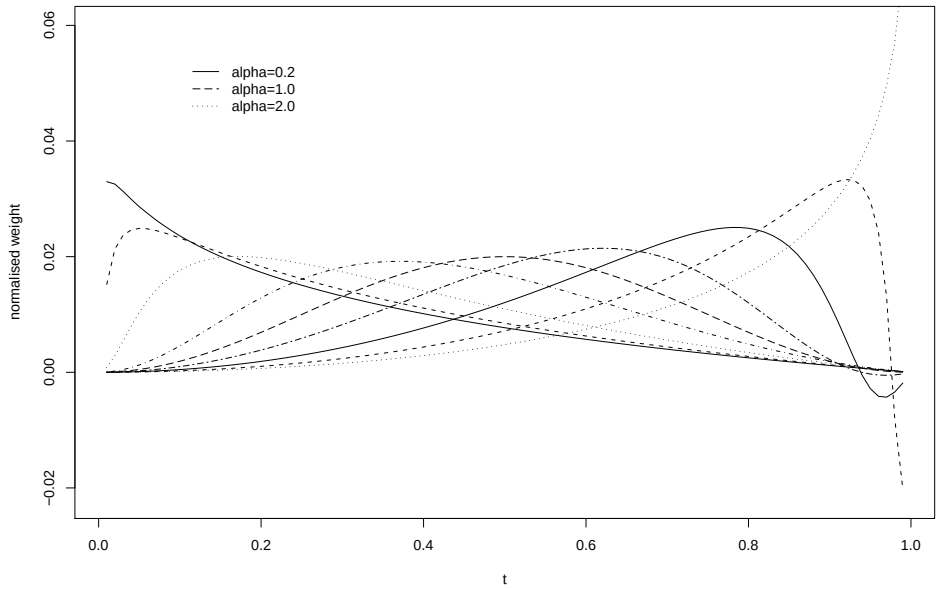


Figure 4.4: This plot displays approximate weights w_{ik} for $t := i/(k+1)$, $i = 1, \ldots, k = 99$, and, starting from the left, following the peaks, $\alpha = 0.2, 0.3, 0.5, 0.8, 1.0, 1.2, 1.5, 1.8, 2.0$.

without iteration. The L-estimator has modest computing requirements: $O(k \log k)$ running time and $O(k)$ storage given a table of pre-calculated weights for given α.

To confirm the superior performance of our L-estimator $\hat{\theta}_{BC}$, we simulate its relative mean square error (m.s.e.), defined by the ratio of $\mathbb{E}(\hat{\theta}_{BC} - \theta)^2$ to θ^2, for various k and α. Since we want to compare the performance of various estimators of θ, generically denoted by $\hat{\theta}$, in terms of deviations from θ, the natural measure to consider is the m.s.e., $\mathrm{MSE}_{\hat{\theta}} = \mathbb{E}(\hat{\theta}_{BC} - \theta)^2$. The best possible m.s.e. is given by the Cramér-Rao lower bound, and equals $[kI_\theta(F_\theta)]^{-1}$. So, to obtain a dimensionless graph, we plot $\mathrm{MSE}_{\hat{\theta}}\, kI_\theta(F_\theta)$, i.e., we compare the simulated m.s.e. for particular estimator $\hat{\theta}$ to the best possible m.s.e. Since $kI_\theta(F_\theta)$ is in units $1/\theta^2$, this is equivalent to comparing the relative m.s.e., $\mathrm{MSE}_{\hat{\theta}}\, /\theta^2$.

Now, when switching between parameterisations in terms of $\theta = \gamma^\alpha$, $\mu = \log \gamma$, and γ, we have to relate the Fisher information and the m.s.e. correctly for the purposes of comparison. We have already shown that $I_\mu(F_\mu) = \gamma^2 I_\gamma(F_\gamma)$; similarly, it can be shown that

$I_\mu(F_\mu) = \alpha^2\theta^2 I_\theta(F_\theta)$. So, the Cramér-Rao lower bound for estimating θ relative to θ^2 equals $1/[k\theta^2 I_\theta(F_\theta)] = \alpha^2/[kI_\mu(F_\mu)]$ in terms of the Fisher information of μ.

Moreover, suppose that one of the methods is estimating γ. Then the m.s.e. for estimating γ, $\mathrm{MSE}_{\hat{\gamma}}\, kI_\gamma(F_\gamma) = \mathrm{MSE}_{\hat{\gamma}}\, kI_\mu(F_\mu)/\gamma^2$, is comparable to that for estimating θ, $\mathrm{MSE}_{\hat{\theta}}\, kI_\theta(F_\theta) = \mathrm{MSE}_{\hat{\theta}}\, kI_\mu(F_\mu)/(\theta^2\alpha^2)$, provided it is multiplied by a factor of α^2. In other words, we plot on the same graph $\mathrm{MSE}_{\hat{\theta}}\,/\theta^2$ and $\alpha^2\,\mathrm{MSE}_{\hat{\gamma}}\,/\gamma^2$ for comparison.

Figure 4.5 displays the relative m.s.e. of the L-estimators of θ, alongside the relative m.s.e. of the fractional power estimator of Li and Hastie (2008) estimating γ, scaled by a factor of α^2, and the relative Cramér-Rao lower bound. The L-estimators have a consistently better performance in terms of m.s.e. than the fractional power estimator. The perturbations in the m.s.e. of the L-estimator at $\alpha = 1.9$ are caused by an oscillation of the weight function; the latter becomes negative when $i/(k+1)$ is close to 1 (see Figure 4.4). The effect can be minimised by trimming or winsorising the L-estimator, explored in Section 4.4.4. The improved bias-corrected estimator (version 3) improves upon the previous two versions by reducing the perturbation in m.s.e. for large values of α (this is particularly noticeable with a sample size of 50); however, versions 2 and 3 of the L-estimator introduce a small perturbation at $\alpha = 0.2$ for $k = 50$.

We investigate the performance of the improved bias-corrected L-estimator for $\alpha > 1.5$; Figure 4.6 displays the weight function for various α, and $k = 50, 100, 150, 200$. For large α and t, the sharp decrease in the value of the weight function is missed in small samples; this is particularly evident for $\alpha = 1.9$. Hence the large order statistics are given positive rather than negative weights, increasing the value of the resulting estimates, and, in turn, inflating the m.s.e. Hence we have reason to believe that trimming or winsorising will significantly improve the performance, particularly in small samples.

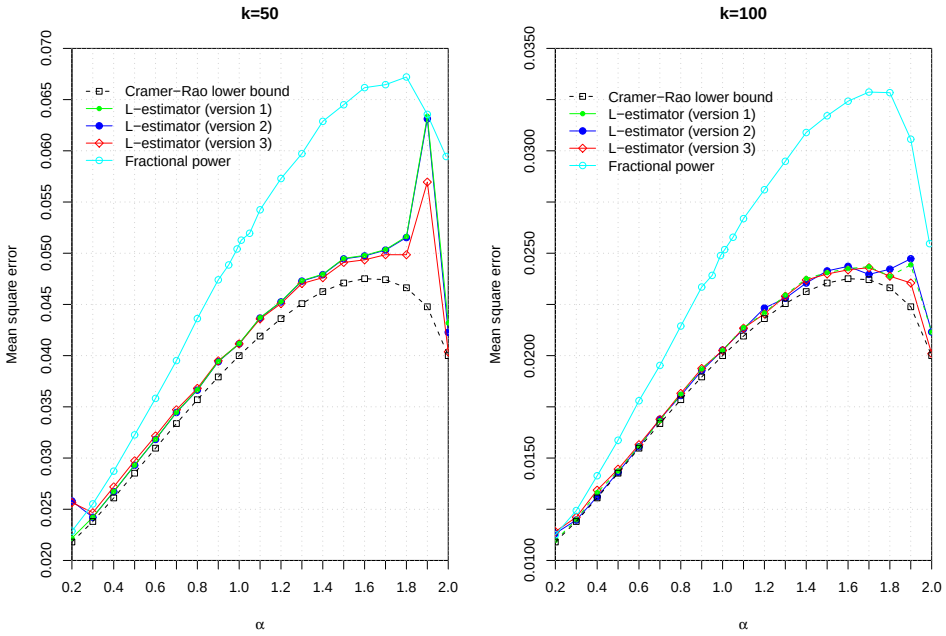


Figure 4.5: Comparison in terms of m.s.e. of the L-estimators of θ with the fractional power estimator of Li and Hastie (2008) (10^5 replicates). The Cramér-Rao lower bound is plotted for comparison. The equivalent plot for estimators of $\gamma = \theta^{1/\alpha}$ shows a similar pattern.

4.4.3 Scale parameter estimation

Estimation Consider the random sample $x_1, \ldots, x_k \sim f(x; \alpha, 0, \gamma, 0)$ with scale parameter $\gamma = \theta^{1/\alpha} > 0$. Define

$$y_i := |x_i|^\alpha := \theta w_i, \quad i = 1, \ldots, k,$$

where w_i is distributed as the αth power of the absolute value of a symmetric, strictly stable random variable of index α and scale parameter 1, with density function

$$f_0(w) = \frac{2}{\alpha} w^{1/\alpha - 1} f(w^{1/\alpha}; \alpha, 0, 1, 0), \quad w > 0,$$

displayed in Figure 4.7. Therefore $(y_1, \ldots, y_k)$ is a random sample from a scale family with p.d.f. $f_\theta(y) = \theta^{-1} f_0\big(y/\theta\big)$, and the problem reduces to that of estimating the parameter θ.

As seen in Figure 4.7, and easy to show analytically, the density function $f_0(w)$ tends to infinity as w tends to 0, for $\alpha > 1$, so an L-estimator of θ cannot be defined from the

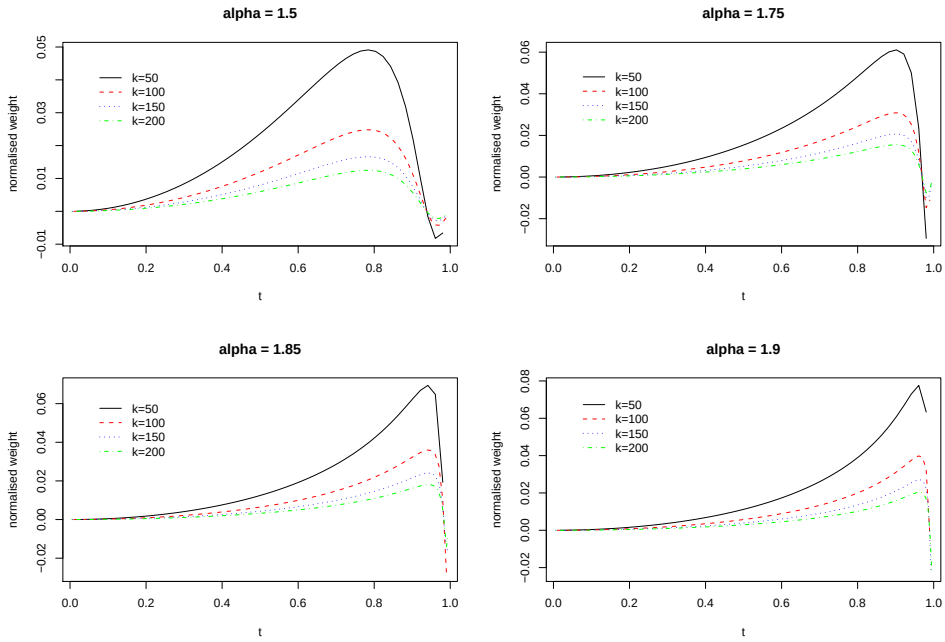


Figure 4.6: Approximate weights w_{ik} for $k = 50, 100, 150, 200$ and $\alpha = 1.5, 1.75, 1.85, 1.9$.

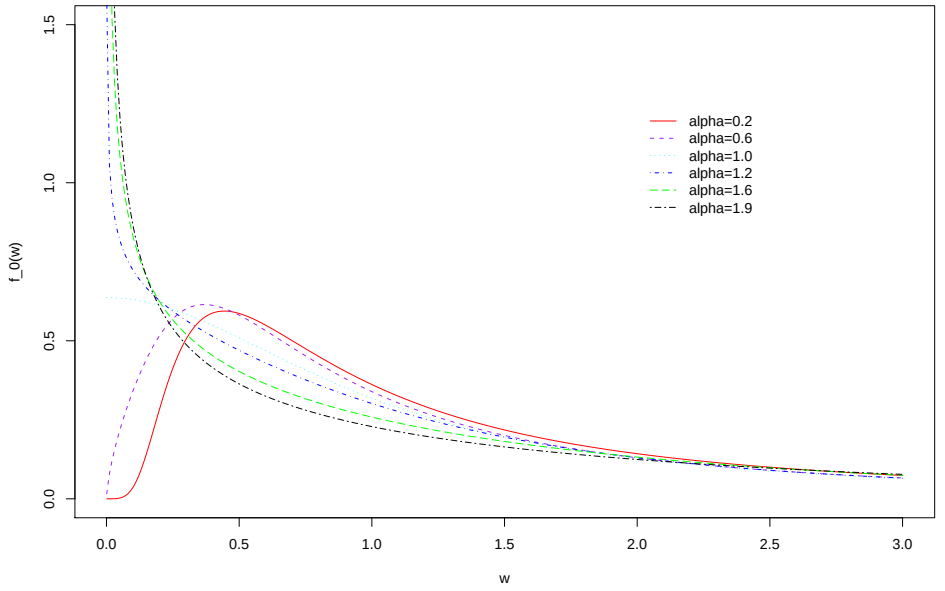


Figure 4.7: Density function $f_0(w)$ computed at $w \in [0.001, 5]$, $\alpha = 0.2, 0.6, 1.0, 1.2, 1.6, 1.9$.

random sample $(y_1, \ldots, y_k)$ for the entire range of α values. However, the expression giving the Fisher information is valid, and we proceed to compute it. Let $I_\theta(F_\theta) = I_{22}/\theta^2$ denote the Fisher information about θ contained in the random sample $(y_1, \ldots, y_k)$, where

$$\begin{aligned} I_{22} = I_1 + I_2 &= -\int_0^\infty z\ell'(z)f_0(z)dz - \int_0^\infty z^2\ell''(z)f_0(z)dz \\ &= -\int_0^1 F_0^{-1}(t)\ell'(F_0^{-1}(t))dt - \int_0^1 \big(F_0^{-1}(t)\big)^2\ell''(F_0^{-1}(t))dt; \end{aligned} \tag{4.18}$$

the second line is obtained by making the transformation $t = F_0(z)$. We approximate I_1 and I_2 by the trapezoid rule with n subintervals of width $h = (1-2\delta)/n$, for given $\delta, n > 0$:

$$I_1 \approx -\frac{h}{2}\sum_{i=1}^{n}\Big[F_0^{-1}(\delta+(i-1)h)\ell'(F_0^{-1}(\delta+(i-1)h)) + F_0^{-1}(\delta+ih)\ell'(F_0^{-1}(\delta+ih))\Big],$$

and

$$I_2 \approx -\frac{h}{2}\sum_{i=1}^{n}\Big[\big(F_0^{-1}(\delta+(i-1)h)\big)^2\ell''(F_0^{-1}(\delta+(i-1)h)) + \big(F_0^{-1}(\delta+ih)\big)^2\ell''(F_0^{-1}(\delta+ih))\Big].$$

This involves solving the equation $F_0(z) = u$ for $z > 0$, where $u \in (0,1)$ fixed. Let $W \sim f(w;\alpha,0,1,0)$. Now, $F_0(z) = 2P(W \le z^{1/\alpha}) - 1$, so $F_0(z) = u$ if and only if $P(W \le z^{1/\alpha}) = (1+u)/2$, i.e., $z^{1/\alpha}$ is the $(1+u)/2$ quantile of the $F(x;\alpha,0,1,0)$ distribution.

To avoid the problem of infinite density, we consider the following transformation for fixed $0 < \beta < 1$:

$$y_i(\beta) = |x_i|^\beta = \theta^{\beta/\alpha}w_i := \eta w_i, \quad i = 1, \ldots, k,$$

giving a random sample from the scale family $f_\eta(y) = \eta^{-1}f_0\big(y/\eta\big)$ with

$$f_0(w) = \frac{2}{\beta}w^{1/\beta-1}f(w^{1/\beta};\alpha,0,1,0), \quad w > 0. \tag{4.19}$$

Proposition 4.4.3. *The density function f_0 defined in* (4.19) *satisfies the conditions for L-estimation with $0 < \beta < 1$.*

Proof. In the limit as $w \to \infty$, $f_0(w) \sim w^{1/\beta-1}\big(w^{1/\beta}\big)^{-(\alpha+1)} = w^{-(\alpha/\beta+1)} \to 0$, and $w^2 f_0'(w) \sim w^{-\alpha/\beta} \to 0$. As $w \to 0$, $f(w^{1/w};\alpha,0,1,0)$ is constant, so $f_0(w) \sim w^{1/\beta-1} \to 0$ since $0 < \beta < 1$, and similarly, $w^2 f_0'(w) \sim w^{1/\beta} \to 0$. □

The L-estimator of η is of the form

$$\hat{\eta} = \sum_{i=1}^{k} w_{ik} y_{(i)}(\beta)$$

with weights given by (4.4) (location parameter 0)

$$w_{ik} = -\frac{\ell'\big(F_0^{-1}(i/(k+1))\big) + F_0^{-1}(i/(k+1))\ell''\big(F_0^{-1}(i/(k+1))\big)}{kI_{22}(\eta)}, \tag{4.20}$$

where $I_\eta(F_\eta) = I_{22}(\eta)/\eta^2$ is the Fisher information about η contained in one sample $y_1(\beta)$. This L-estimator satisfies $\sqrt{k}\big(\hat{\eta} - \eta\big) \xrightarrow{\mathcal{D}} \text{Normal}\big(0, 1/I_\eta(F_\eta)\big)$, where $I_\eta(F_\eta)$ is related to $I_\theta(F_\theta)$ as follows:

$$I_\eta(F_\eta) = I_\theta(F_\theta)\Big(\frac{\partial\theta}{\partial\eta}\Big)^2 = \frac{1}{\eta^2}\frac{\alpha^2}{\beta^2} I_\theta(F_\theta)\theta^2, \tag{4.21}$$

so $I_{22}(\eta) = I_{22}\alpha^2/\beta^2$.

Proposition 4.4.4. *The L-estimator $\sum_{i=1}^{k} w_{ik} Y_{(i)}(\beta)$ has finite variance if $0 < \beta < \alpha/2$.*

Proof. The L-estimator satisfies

$$\Big|\sum_{i=1}^{k} w_{ik} Y_{(i)}(\beta)\Big| \le \sum_{i=1}^{k} |w_{ik}| Y_{(i)}(\beta) \le k \max\{|w_{ik}|, i = 1, \ldots, k\} Y_{(k)}(\beta),$$

so we're interested in the behaviour of $Y_{(k)}(\beta)$ in the tail as $y \to \infty$. From Proposition 4.4.3, $f_\eta(y) \sim y^{-(\alpha/\beta+1)}$ as $y \to \infty$, so $f_{(k)}(y) \sim y^{-(\alpha/\beta+1)}$. Now, $\int_0^\infty y^2 y^{-(\alpha/\beta+1)} dy = \int_0^\infty y^{-(\alpha/\beta-1)} dy < \infty$, provided $\alpha/\beta - 1 > 1$, i.e., $\beta < \alpha/2$. □

The sum of the weights $\sum_{i=1}^{k} w_{ik}$ tends to $I_{12}(\eta)/I_{22}(\eta)$ as $k \to \infty$, where $I_{12}(\eta) = \int_0^\infty L_2'(y) f_0(y) dy = I_{12}\alpha^2/\beta^2$, and I_{12} is computed from the sample $(y_1, \ldots, y_k)$. So the normalised weights are:

$$w_{ik} = \frac{I_{12}(\eta)}{I_{22}(\eta)} \times \frac{\ell'\big(F_0^{-1}(i/(k+1))\big) + F_0^{-1}(i/(k+1))\ell''\big(F_0^{-1}(i/(k+1))\big)}{\sum_{j=1}^{k}\Big[\ell'\big(F_0^{-1}(j/(k+1))\big) + F_0^{-1}(j/(k+1))\ell''\big(F_0^{-1}(j/(k+1))\big)\Big]}. \tag{4.22}$$

In two special cases, we have exact expressions for the weights: for $\alpha = 1$,

$$w_{ik} = \frac{8}{k} \times \frac{\Big[\tan\big(\pi/2 \times i/(k+1)\big)\Big]^{2-\beta}}{\Big[1 + \tan^2\big(\pi/2 \times i/(k+1)\big)\Big]^2},$$

and for $\alpha = 2$,

$$w_{ik} = \frac{1}{2k} \times \Big(\Phi^{-1}\Big(\frac{1}{2} + \frac{i}{2(k+1)}\Big)\Big)^{2-\beta}.$$

Now, the estimator for θ

$$\hat{\theta} = \hat{\eta}^{\alpha/\beta} = \Big(\sum_{i=1}^{k} w_{ik} y_{(i)}(\beta)\Big)^{\alpha/\beta} \tag{4.23}$$

is asymptotically efficient; in fact, it can be shown by the Delta method and equality (4.21) that $\sqrt{k}\theta^{-1}\big\{\hat{\theta} - \theta\big\} \to \text{Normal}(0, 1/I_{22})$ as $k \to \infty$. Although it is asymptotically unbiased, it is biased in finite samples; we proceed by estimating the finite sample bias of $\hat{\eta}$ by a first order Taylor expansion.

$$\mathbb{E}\hat{\eta} = \eta \sum_{i=1}^{k} w_{ik} \mathbb{E} w_{(i)} \approx \eta \sum_{i=1}^{k} w_{ik} F_0^{-1}\Big(\frac{i}{k+1}\Big),$$

for k large; we propose the following bias-corrected estimator

$$\hat{\eta}_{BC} = \frac{\sum_{i=1}^{k} w_{ik} y_{(i)}(\beta)}{\sum_{j=1}^{k} w_{ik} F_0^{-1}(j/(k+1))},$$

where the denominator tends to 1 as $k \to \infty$. Again, $\hat{\eta}_{BC}$ is asymptotically efficient. Let $\hat{\theta} = \hat{\eta}_{BC}^{\alpha/\beta}$. By a second-order Taylor expansion around η, we obtain

$$\hat{\theta} \approx \theta + \frac{\alpha}{\beta}\frac{\theta}{\eta}\Big(\hat{\eta}_{BC} - \eta\Big) + \frac{1}{2}\frac{\alpha}{\beta}\Big(\frac{\alpha}{\beta} - 1\Big)\frac{\theta}{\eta^2}\Big(\hat{\eta}_{BC} - \eta\Big)^2.$$

Taking expectations,

$$\mathbb{E}\hat{\theta} - \theta \approx \frac{1}{2}\frac{\alpha}{\beta}\Big(\frac{\alpha}{\beta} - 1\Big)\frac{\theta}{\eta^2} \operatorname{var} \hat{\eta}_{BC} \approx \frac{\theta}{2}\Big(1 - \frac{\beta}{\alpha}\Big)\frac{1}{kI_{22}}.$$

We propose the bias-corrected estimator (called version 1)

$$\hat{\theta}_{BC} = \Big(1 - \frac{1}{2}\Big(1 - \frac{\beta}{\alpha}\Big)\frac{1}{kI_{22}}\Big)\hat{\eta}_{BC}^{\alpha/\beta}, \tag{4.24}$$

which is unbiased up to terms of order $O(1/k^2)$ and asymptotically efficient.

Furthermore, to improve the estimation procedure, we approximate the bias of η by a second order Taylor approximation as in equation (4.14), starting from

$$\mathbb{E}\hat{\eta} = \eta \sum_{i=1}^{k} w_{ik} \mathbb{E} w_{(i)} = \eta \sum_{i=1}^{k} w_{ik} \mathbb{E} F_0^{-1}(U_{(i)}),$$

to obtain the improved bias-corrected estimator

$$\hat{\eta}_{BC} = \frac{\sum_{i=1}^{k} w_{ik} y_{(i)}(\beta)}{\sum_{i=1}^{k} w_{ik} \Big\{ F_0^{-1}\big(i/(k+1)\big) - \frac{i(k-i+1)}{2(k+1)^2(k+2)} \frac{\ell'(F_0^{-1}(i/(k+1)))}{[f_0(F_0^{-1}(i/(k+1)))]^2} \Big\}}.$$

The improved bias-corrected estimators of θ and γ follow; we refer to these as version 2.

Computation Figure 4.8 displays the integrand functions of integrals I_1 and I_2 for various α, computed at n equally spaced points in the interval $[\delta, 1-\delta]$ ($\delta = 0.001$, $n = 100$).

We simplify the expressions for $x\ell'(x)$ and $x^2\ell''(x)$, where $\ell(x) = \log f_0(x)$, instead of estimating the first and second derivatives directly; the latter approach suffers from numerical difficulties in computing $f_0(x)$ for x close to 0 when $\alpha > 1$ because $f_0(x) \to \infty$, as shown in Figure 4.7. We have

$$\ell(x) = \log\left(\frac{2}{\alpha}\right) + \left(\frac{1}{\alpha} - 1\right) \log x + \log f(x^{1/\alpha}; \alpha, 0, 1, 0),$$

so

$$\begin{aligned} x\ell'(x) &= \frac{1}{\alpha} - 1 + x \frac{\partial}{\partial x} \log f(x^{1/\alpha}; \alpha, 0, 1, 0), \\ x\ell''(x) &= 1 - \frac{1}{\alpha} + x^2 \frac{\partial^2}{\partial x^2} \log f(x^{1/\alpha}; \alpha, 0, 1, 0), \end{aligned}$$

where the second order derivative of $\log f(x^{1/\alpha}; \alpha, 0, 1, 0)$ with respect to x is estimated via a second order finite difference scheme as in (4.16), and the first order derivative is estimated similarly:

$$\frac{\partial}{\partial x} \log f(x^{1/\alpha}; \alpha, 0, 1, 0) \approx \frac{1}{2h} \Big[\log f\big((x+h)^{1/\alpha}; \alpha, 0, 1, 0\big) - \log f\big((x-h)^{1/\alpha}; \alpha, 0, 1, 0\big) \Big].$$

The value of h in the finite difference scheme depends both on α and on x; we take $h = 0.01$ for $\alpha < 0.3$, and $h = 0.001$ otherwise, and we increase the width to $h = 0.01$ for x very large, because the density function in the tail varies very slowly. We perform all computations with our improved R functions for density, distribution and quantile estimation.

Next, we estimate $I_{22} = I_1 + I_2$, where $I_{22} = \theta^2 I_\theta(F_\theta)$, by approximating the integrals I_1 and I_2 via the trapezoid rule applied to the interval $[\delta, 1-\delta]$ with n points placed at a distance of $h = (1-2\delta)/n$, where $\delta = 0.001$ and $n = 1000$. As seen in Figure 4.8, the integrand functions change rapidly for $t \geq 0.9$ and $\alpha \geq 1.6$; in this region, we use a finer grid with of $h/20$. The results appear in Figure 4.9 and Table 4.4.

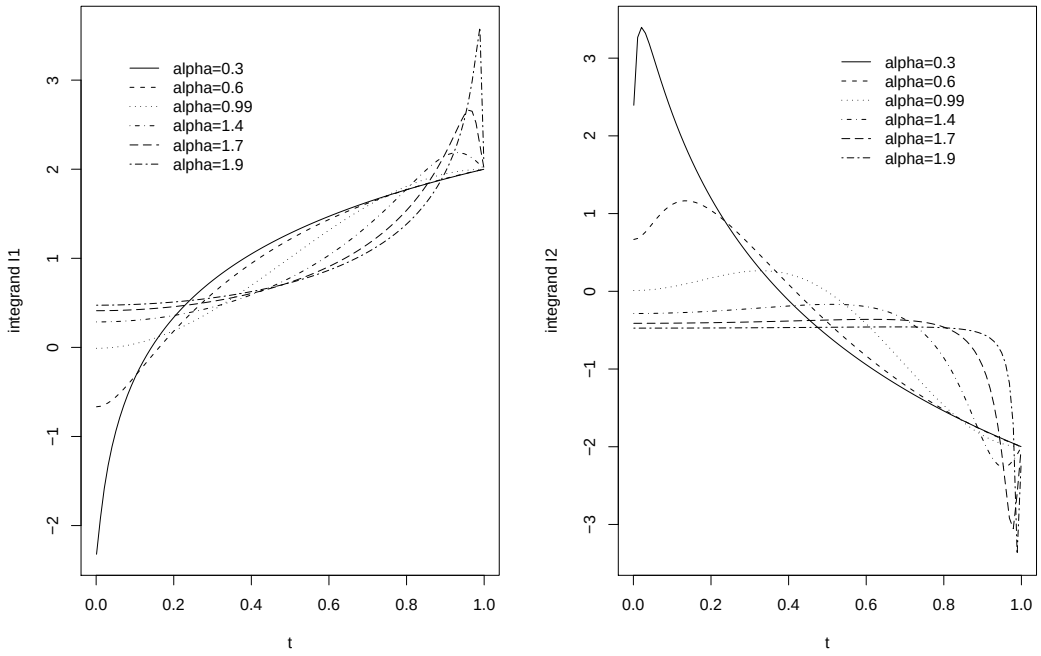


Figure 4.8: Approximations to the integrand functions of I_1 (left) and I_2 (right) in equation (4.18) for $\alpha = 0.3, 0.6, 0.99, 1.4, 1.7, 1.9$, with $\delta = 0.001$, $n = 100$.

The Fisher information for θ can also be computed via a transformation of the Fisher information for μ, given in Table 4.1, as follows: $\mu = (\log\theta)/\alpha$, so $\theta = \exp(\mu\alpha)$, and $I_\mu(F_\mu) = I_\theta(F_\theta) \times (\partial\theta/\partial\mu)^2 = I_{22}\alpha^2/\theta^2$. The third and sixth columns in Table 4.4 give

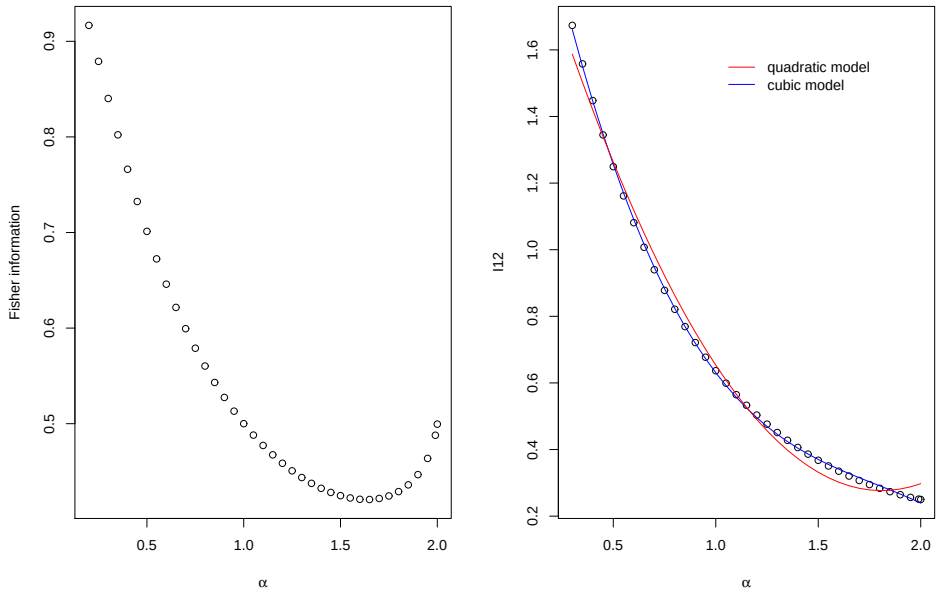


Figure 4.9: Left: Approximate Fisher information I_{22} as a function of α, for $\alpha \in [0.2, 2.0]$. Right: Fitted models of I_{12} as a function of α; black circles indicate estimated values.

the Fisher information for θ computed as $I_\mu(F_\mu)/\alpha^2$. The values I_{22} and $I_\mu(F_\mu)/\alpha^2$ agree to within 4 significant digits for all α, with the exception of $\alpha < 0.3$; we use the latter approximations to the Fisher information (third and sixth columns) in the sequel.

As seen in Figure 4.9, I_{22} tends to 1 as $\alpha \to 0$; we prove this as follows. Recall that $y_i = |x_i|^\alpha = \theta w_i$; based on previously shown results, we know that the distribution of w_i converges to the inverse of the Exp(1) distribution as $\alpha \to 0$. Hence $y_i \xrightarrow{\mathcal{D}} \theta/L_i$, where $L_i \sim \text{Exp}(1)$ as $\alpha \to 0$, and the Fisher information about the scale parameter θ, denoted by I_{22}/θ^2, converges to $1/\theta^2$ as $\alpha \to 0$. It follows that $I_{22} \to 1$ as $\alpha \to 0$. Moreover, we showed in the previous paragraph that $I_\mu(F_\mu) = I_{11}/\theta^2 = \alpha^2 I_{22}/\theta^2$, so $I_{11} = \alpha^2 I_{22}$; hence, if $I_{22} \to 1$ as $\alpha \to 0$, then it follows that $I_{11} \to 0$ as $\alpha \to 0$. This justifies modelling I_{11} by zero-intercept models in Section 4.4.2.

α	I_{22}	$I_\mu(F_\mu)/\alpha^2$	α	I_{22}	$I_\mu(F_\mu)/\alpha^2$
0.2	0.9167	0.9173	1.15	0.4674	0.4674
0.25	0.8790	0.8792	1.2	0.4586	0.4586
0.3	0.8402	0.8402	1.25	0.4507	0.4507
0.35	0.8022	0.8023	1.3	0.4437	0.4437
0.4	0.7661	0.7661	1.35	0.4376	0.4376
0.45	0.7324	0.7324	1.4	0.4324	0.4324
0.5	0.7012	0.7012	1.45	0.4281	0.4281
0.55	0.6724	0.6724	1.5	0.4247	0.4247
0.6	0.6459	0.6459	1.55	0.4223	0.4223
0.65	0.6216	0.6216	1.6	0.4209	0.4209
0.7	0.5994	0.5994	1.65	0.4207	0.4207
0.75	0.5789	0.5789	1.7	0.4218	0.4218
0.8	0.5603	0.5602	1.75	0.4244	0.4244
0.85	0.5431	0.5431	1.8	0.4289	0.4289
0.9	0.5275	0.5274	1.85	0.4360	0.4360
0.95	0.5131	0.5131	1.9	0.4467	0.4467
1.0	0.5	0.5	1.95	0.4637	0.4637
1.05	0.4881	0.4881	1.99	0.4879	0.4879
1.1	0.4772	0.4772	2.0	0.5	0.5

Table 4.4: Approximate Fisher information I_{22}, tabulated for values of $\alpha \in [0.2, 2]$, computed via two different methods: estimating the value of the integrals in expression (4.18), and by a tranformation of the Fisher information for μ given in Table 4.1.

We proceed to compute the weight functions given by (4.20) using the simplified formula

$$\ell'(x) + x\ell''(x) = \frac{\partial}{\partial x} \log f(x^{1/\beta}; \alpha, 0, 1, 0) + x \frac{\partial^2}{\partial x^2} \log f(x^{1/\beta}; \alpha, 0, 1, 0),$$

with $k = 99$, for $\alpha \in [0.2, 2.0]$ and $\beta = 0.025$; the results are shown in Figure 4.10. For $\alpha = 1, 2$ we plot the exact weight functions. We also compute the normalised weights as in expression (4.22); this latter approach removes the bias incurred in estimating the first and second derivatives by finite differences, provided k, the number of weights, is large enough. The two approaches give nearly identical weight functions.

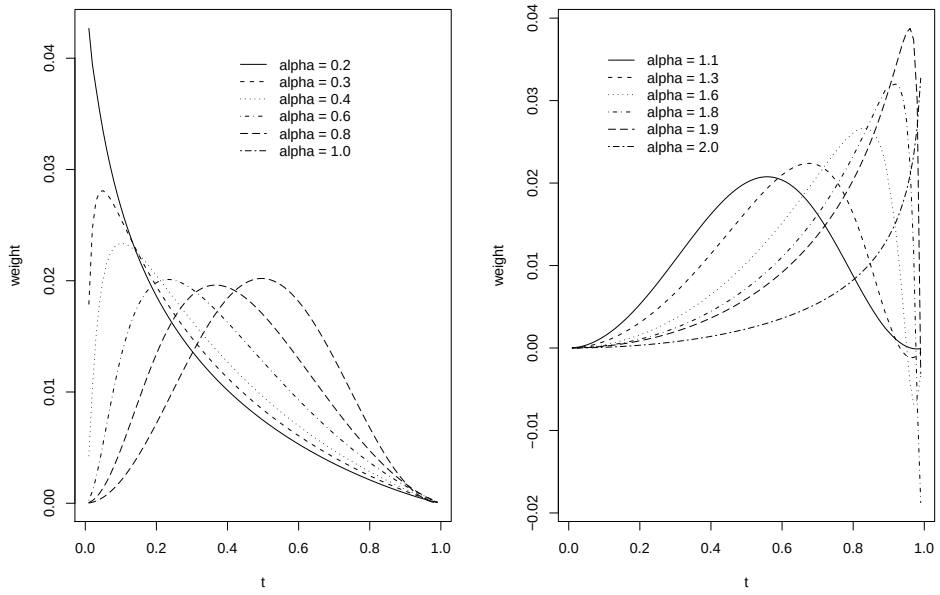


Figure 4.10: Approximate weight functions as given by expression (4.20) for various values of α and $\beta = 0.025$; exact expressions are used for $\alpha = 1, 2$.

Computing the normalised weights involves estimating the integral I_{12} by the trapezoid rule, following the same approach as in estimating I_{22}; for $\alpha = 1$, $I_{12} = 2/\pi$, and for $\alpha = 2$, $I_{12} = 1/4$. Table 4.5 presents the approximations to I_{12} for $\alpha \in [0.3, 2]$; for $\alpha < 0.3$, the approximations appear unreliable, so we estimate the value of I_{12} in this range via extrapolation. Figure 4.9 shows the quadratic and cubic models fitted to the data via the method of least squares; our final model is the cubic model: $I_{12} \sim 2.48170 - 3.20779\alpha + 1.67035\alpha^2 - 0.31353\alpha^3$ that explains 99.97% of the variability of the data. The predicted values are as follows: for $\alpha = 0.2$, $I_{12} = 1.9044$ with prediction interval $(1.8852, 1.9237)$, and

for $\alpha = 0.25$, $I_{12} = 1.7793$ with prediction interval $(1.7615, 1.7970)$.

α	I_{12}	α	I_{12}	α	I_{12}	α	I_{12}
0.3	1.6736	0.75	0.8779	1.2	0.5035	1.65	0.3204
0.35	1.5579	0.8	0.8212	1.25	0.4762	1.7	0.3071
0.4	1.4478	0.85	0.7691	1.3	0.4511	1.75	0.2948
0.45	1.3447	0.9	0.7212	1.35	0.4278	1.8	0.2836
0.5	1.2493	0.95	0.6772	1.4	0.4062	1.85	0.2735
0.55	1.1616	1.0	0.6366	1.45	0.3862	1.9	0.2644
0.6	1.0811	1.05	0.5992	1.5	0.3678	1.95	0.2564
0.65	1.0074	1.1	0.5647	1.55	0.3507	1.99	0.2510
0.7	0.9398	1.15	0.5329	1.6	0.3350	2.0	0.2500

Table 4.5: Approximation to integral I_{12}, tabulated for $\alpha \in [0.3, 2]$, obtained via the finite difference approach ($n = 1000$, $\delta = 0.001$).

Numerical results Figure 4.11 compares the small sample performance of the version 1 estimator of θ, for various β, to that of the fractional power estimator; best performance in terms of m.s.e. is obtained with $\beta = 0.025$. For a dimensionless graph, we plot the empirical m.s.e. divided by θ^2 against the Cramér-Rao lower bound computed as $1/(k \times I_{22})$. The m.s.e. is close to the theoretical lower bound except in the region $\alpha \in (1.8, 2.0)$, where the weight function drops sharply in value for $t = i/(k+1)$ close to 1, as shown in Figure 4.12. When the sample size k is small, the weight function is estimated crudely at a few points, so it is possible to miss the change from large, positive weights to negative ones, and to place a positive weight on the largest order statistics instead of a negative one. This has the effect of inflating the m.s.e. Removing the finite sample bias reduces this inflation, as does decreasing the power β, which tempers the effect of large positive or small negative values drawn from

the stable distribution. See Figure 4.11. In the location estimation case, the effect of these extreme values was decreased by taking the logarithm of the absolute value.

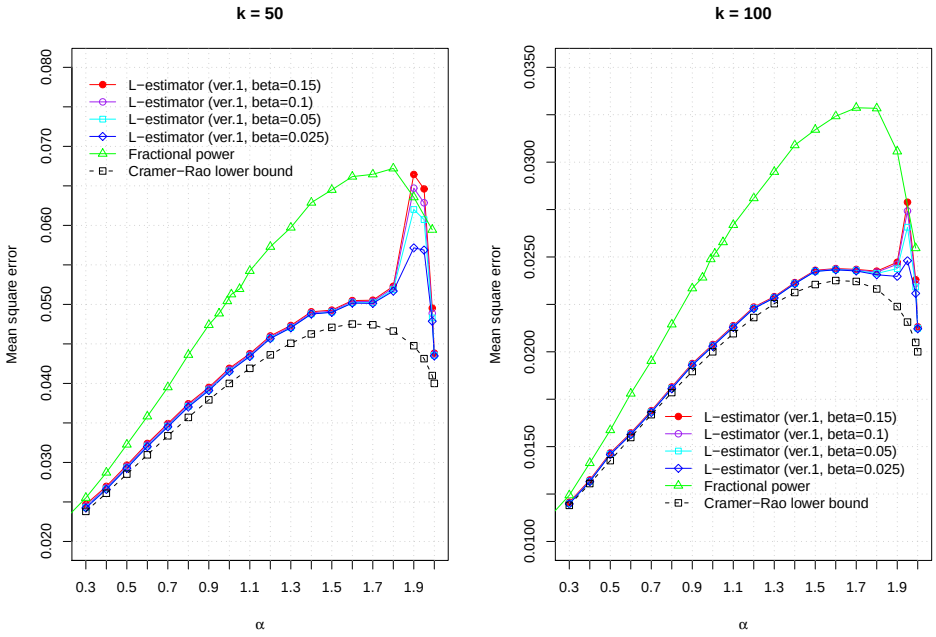


Figure 4.11: Comparison in terms of m.s.e. of the version 1 estimator of θ against the fractional power estimator (10^5 replicates), alongside the Cramér-Rao lower bound.

Figure 4.18 displays the m.s.e. of the improved bias-corrected estimator (version 2) of θ with $\beta = 0.025$; this estimator has better performance than version 1, but also exhibits a small perturbation (particularly for a sample size of 50) in the region of $\alpha = 1.9$. In the following subsection, we attempt to eliminate this perturbation via trimming and winsorising.

4.4.4 Trimmed and winsorised estimation

In this subsection, we improve the small sample performance of the L-estimators by trimming, i.e., setting some of the weights to zero and normalising, and winsorising (Tukey, 1962), i.e., setting the values of extreme order statistics to a specified percentile. These approaches are justified by the behaviour of the weight function in the range corresponding to the large

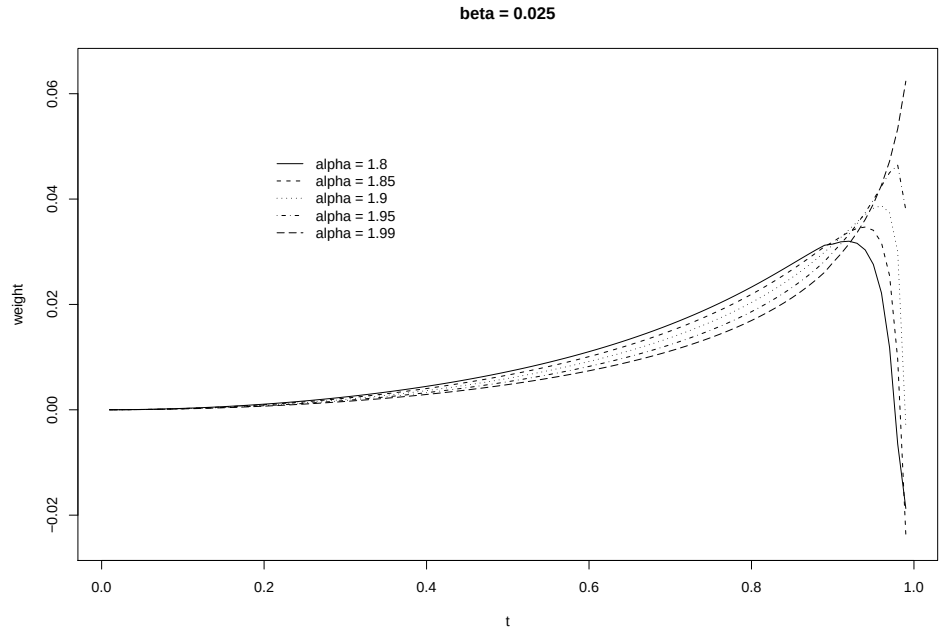


Figure 4.12: Approximate weight function (4.22) for $\alpha = 1.8, 1.9, 1.95, 1.99$, and $\beta = 0.025$.

order statistics in the weighted linear combination. We show that trimming and winsorising improves the performance for large α by reducing the m.s.e. close to the Cramér-Rao lower bound, and decreasing the length of the 95% confidence interval.

For the location estimation problem, Figure 4.5 shows a significant perturbation in the m.s.e. of the L-estimator at $\alpha = 1.9$ and a barely noticeable one at $\alpha = 0.2$. We attribute the latter to numerical instabilities in fBasics commands for α small, discussed in Subsection 2.5. However, we believe the former to be caused by the oscillation in weight function corresponding to large order statistics for large α, shown in Figure 4.6.

Recall that the location parameter μ is estimated by $\hat{\mu} = \sum_{i=1}^{k} w_{ik} y_{(i)}$, where $\sum_{i=1}^{k} w_{ik} = 1$. We trim symmetrically by defining weights $w'_{ik} \propto w_{ik}$, if $i \neq 1, k$, and $w'_{ik} = 0$ otherwise; after normalising, we have

$$w'_{ik} = \begin{cases} \left(\sum_{j=2}^{k-1} w_{jk}\right)^{-1} w_{ik} & \text{if } i \neq 1, k \\ 0 & \text{if } i = 1, k. \end{cases}$$

The computation of the bias remains unchanged.

The winsorised estimator at percentile $p(> 0.5)$ is obtained by setting the bottom $(1-p)/2 \times 100\%$ order statistics to the $(1-p)/2$ empirical quantile, and the top $(1-p)/2 \times 100\%$ order statistics to the $(1+p)/2$ empirical quantile. Let i_1, i_2 denote the index of the $(1-p)/2$ and $(1+p)/2$ empirical quantiles, respectively. For $i < i_1$, set $y_{(i)} = y_{(i_1)}$, and for $i > i_2$, set $y_{(i)} = y_{(i_2)}$. Then the estimator $\hat{\mu}$ becomes $\hat{\mu} = \sum_{i=1}^{i_1-1} w_{ik} y_{(i_1)} + \sum_{i=i_1}^{i_2} w_{ik} y_{(i)} + \sum_{i=i_2+1}^{k} w_{ik} y_{(i_2)}$, and the bias is modified accordingly. The winsorised estimator effectively draws in the $(1-p) \times 100\%$ extreme order statistics. It has been shown that winsorising can result in highly efficient estimators, as in the case of estimating the mean of a normal population (Dixon, 1960).

Figure 4.13 compares the m.s.e. of the L-estimator (version 3, trimmed and winsorised with $p = 0.9$), and of the fractional power estimator, to the theoretical lower bound for sample sizes $k = 50, 100$. For $\alpha \in (1.8, 2.0)$ and $k = 50$, trimming at the two extreme endpoints and winsorising both remove the perturbation, reducing the m.s.e. close to the Cramér-Rao lower bound; the same effect is noticed for $\alpha < 0.3$. Trimming results in a small loss of efficiency for $\alpha \in (1.6, 1.8)$; for all other values, there is no noticeable loss in efficiency. This is explained by the fact that for $\alpha \in [0.3, 1.6]$, the weight function attains values very close to zero at the two extreme endpoints, so trimming has no effect. Overall, the winsorised estimator appears to perform slightly better than the trimmed one. Figure 4.14 compares the length of the 95% confidence interval of the L-estimator of θ to the theoretical lower bound derived from the asymptotic distribution. The latter is computed as $2 \times 1.96 \times \alpha / \sqrt{k \times I_\mu(F_\mu)}$, and the former is estimated as the difference between the 0.975 and 0.025 empirical quantiles based on 10^5 replicates. Both trimming and winsorising decrease the length of the confidence interval in the critical region $\alpha \in [1.8, 2)$. Overall, we expect the L-estimator, which is unbiased up to terms of order $O\big(1/k^2\big)$, to fall within a small interval around the true value θ with high probability.

Figure 4.15 displays the significant jump in m.s.e. at $\alpha = 1.91$ for $k = 50$ and $\alpha \in$

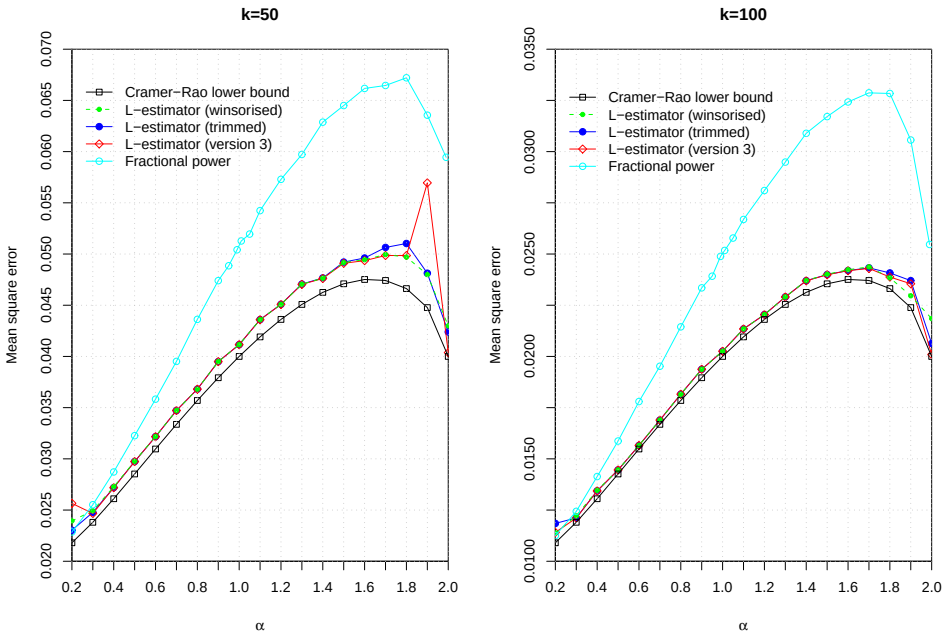


Figure 4.13: Comparison in terms of m.s.e. of the L-estimator of θ (ver. 3, trimmed and winsorised) obtained by location estimation approach (10^5 replicates).

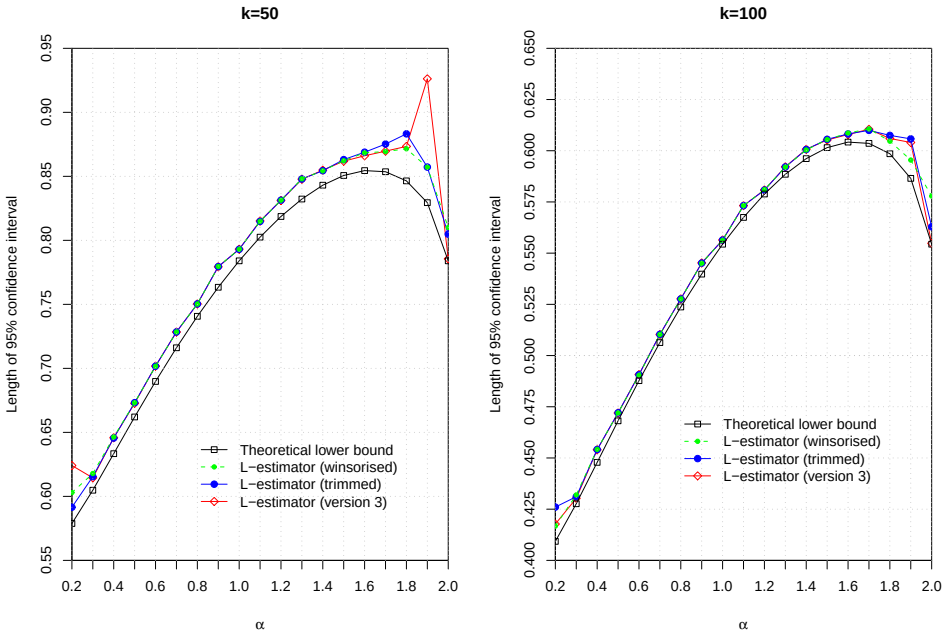


Figure 4.14: Comparison in terms of length of 95% confidence interval for L-estimator of θ (ver. 3, trimmed and winsorised) obtained by location estimation approach (10^5 replicates).

$(1.94, 1.96)$ for $k = 100$, considerably reduced by trimming at the two extreme endpoints, and by winsorising with $p = 0.9$. In fact, the winsorised estimator performs slightly better than the trimmed one for $\alpha \in [1.8, 1.9]$ with $k = 50$, and $\alpha \in [1.8, 1.94]$ with $k = 100$. Both methods avoid the sharp drop in weight function w_{ik} for i close to k and large α (shown in Figure 4.16) by setting the weight to 0 (trimming), or by replacing the extreme order statistic by a less extreme one (winsorising).

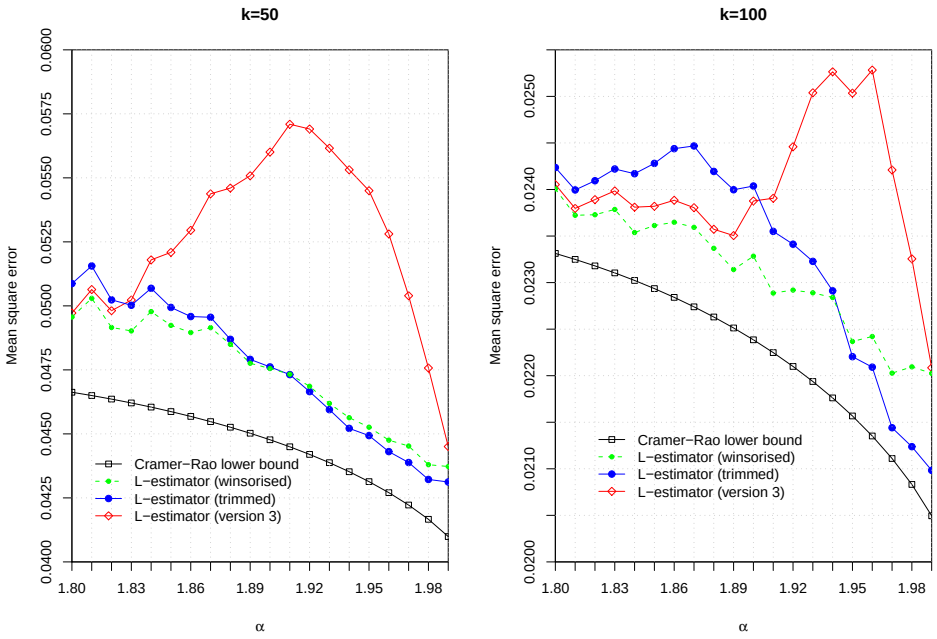


Figure 4.15: Comparison in terms of m.s.e. of the L-estimator of θ (ver. 3, trimmed and winsorised) for large α, obtained by location estimation approach (10^5 replicates).

Figure 4.17 displays the optimal quantile (q) defining the estimator (4.6) of θ based on a single weighted empirical quantile, plotted alongside the asymptotic efficiency relative to the MLE, where, for given α, q is chosen to minimise the asymptotic variance. In the critical region $\alpha \in (1.8, 2.0)$, $q \in (0.78, 0.87)$ with ARE decreasing from 0.76 to 0.65. Although the optimal q does not correspond to the index i of the largest weight w_{ik} (in fact, for $\alpha = 1.95$, w_{ik} is an increasing function of i), it indicates that the top 10% of order statistics are not very important in estimating θ, and that, therefore, removing or replacing them with less

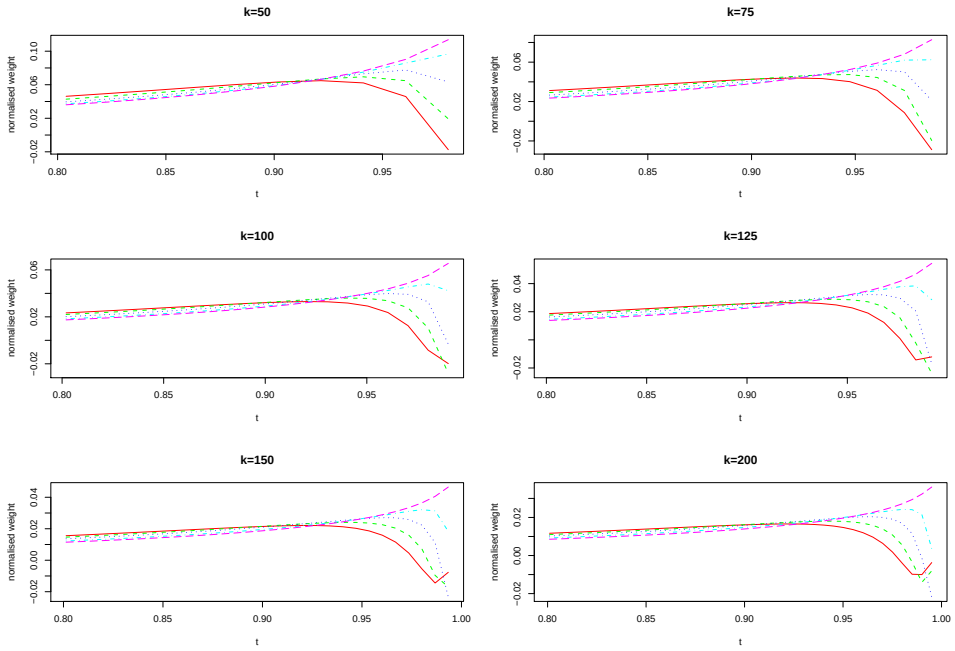


Figure 4.16: Weight function, equation (4.8), with $t = i/(k+1)$, $k = 50, 75, 100, 125, 150, 200$, and $\alpha = 1.8$ (red), 1.85 (green), 1.9 (blue), 1.95 (cyan), 1.99 (magenta).

extreme quantiles will not result in a significant loss in efficiency.

For the scale estimation problem, the estimator $\hat{\eta} = \sum_{i=1}^{k} w_{ik} y_{(i)}(\beta)$ has weights satisfying $\sum_{i=1}^{k} w_{ik} = I_{12}(\eta)/I_{22}(\eta)$. We trim symmetrically at the two endpoints by defining new weights

$$w'_{ik} = \begin{cases} \left(\sum_{j=2}^{k-1} w_{jk}\right)^{-1} w_{ik} I_{12}(\eta)/I_{22}(\eta) & \text{if } i \neq 1, k \\ 0 & \text{if } i = 1, k. \end{cases}$$

The multiplicative bias remains unchanged. For the winsorised estimator at percentile p, we set the top and bottom order statistics to the corresponding empirical quantiles, and modify the bias accordingly.

Figure 4.18 shows the results of trimming and winsorising ($p = 0.9$) for $\alpha \in [0.2, 2.0)$ and $\beta = 0.025$. For $\alpha \in (1.6, 1.83)$ with $k = 50$ and $\alpha \in (1.7, 1.91)$ with $k = 100$, trimming results in small loss of efficiency compared to the untrimmed L-estimator. For $\alpha < 0.4$, although the smallest weight $w_{1k} \in (0.03, 0.04)$ is large, the order statistics to which it corresponds

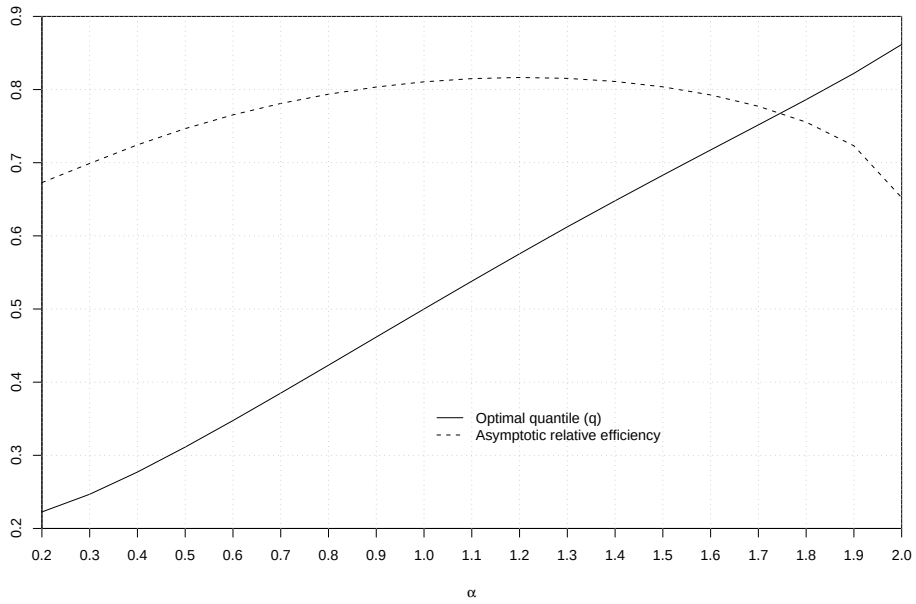


Figure 4.17: Optimal quantile defining the quantile estimator in (4.6), and corresponding asymptotic efficiency relative to the MLE.

are so small in value that only a very small loss in effiency due to trimming is observed for $k = 100$. Overall, again, the winsorised estimator performs slightly better than the trimmed one, with a small deterioration for very large α.

Figure 4.19 compares the length of the 95% empirical confidence interval to the theoretical lower bound, $2 \times 1.96/\sqrt{k \times I_{22}}$, showing that we can expect the L-estimator to fall close to the true value of θ with high probability. For nearly the entire range of α values, the winsorised estimator has confidence interval of shortest length.

In more detail, Figure 4.20 shows that trimming and winsorising ($p = 0.9$) significantly reduce the perturbation in m.s.e. for $\alpha \in [1.83, 1.99]$ with $k = 50$, and for $\alpha \in [1.91, 1.99]$ with $k = 100$. As in the location estimation problem, we believe this perturbation to be caused by the sharp drop in value of the weight function for large α, displayed in Figure 4.12. Again, for $\alpha \in [1.8, 1.93]$ with $k = 50$, and $\alpha \in [1.8, 1.94]$ with $k = 100$, the winsorised estimator outperforms slightly the trimmed estimator.

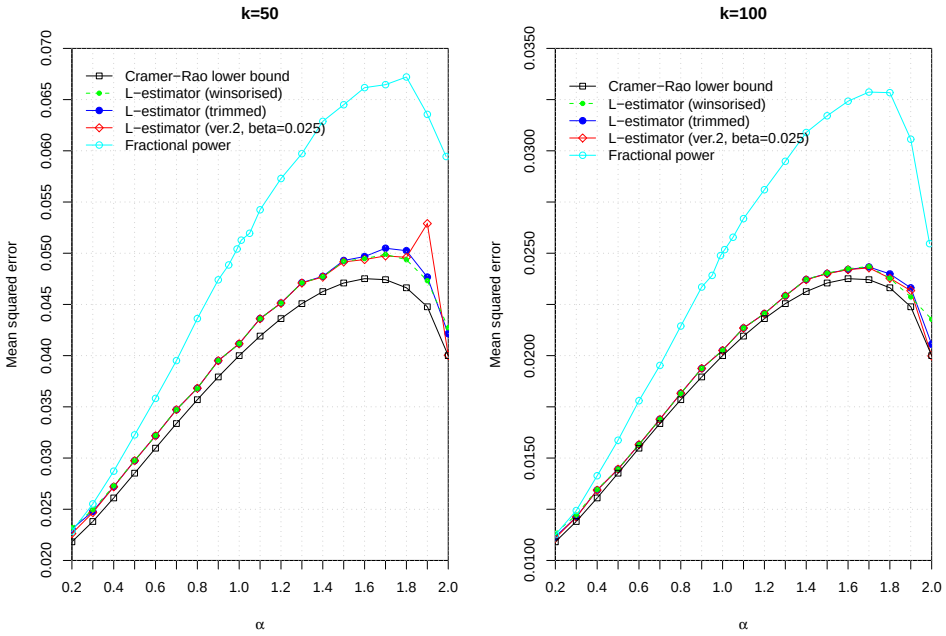


Figure 4.18: Comparison in terms of m.s.e. of the L-estimator of θ (ver. 2, trimmed and winsorised, $\beta = 0.025$) obtained by scale estimation approach (10^5 replicates).

4.5 Maximally skewed stable random projections

4.5.1 Introduction

Li (2008c, 2009) introduces the method of projecting with maximally skewed strictly stable random variables ($\beta = 1$), which he calls *compressed counting*, for approximating frequency moments over data streams, or, equivalently, estimating l_α distances (quasi-distances) between streams, for fixed $0 < \alpha \leq 2$. We note that when projecting to maximally skewed strictly stable random variables, the constants in the linear combination in (2.3) must be positive.

Ping Li argues that compressed counting reduces the space complexity of estimating the αth frequency moment to within factor $1 \pm \epsilon$ by decreasing the lower bound on the dimension of the projection space, k, from $O(\epsilon^{-2})$ to $O(\epsilon^{-1})$ when $\alpha = 1 \pm \Delta$, as $\Delta \to 0$. He proposes geometric mean, harmonic mean, fractional power, and optimal quantile estimators

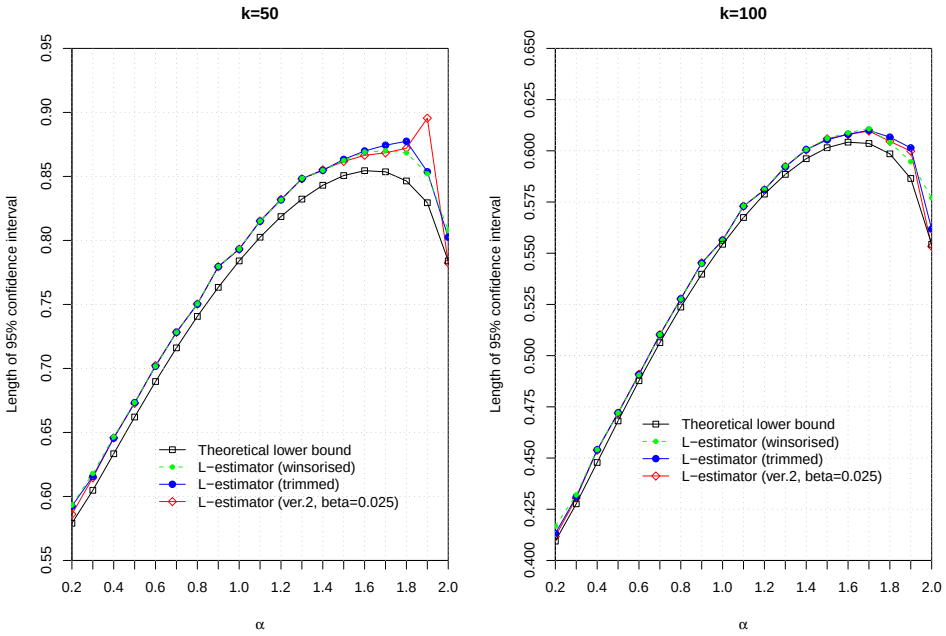


Figure 4.19: Comparison of length of 95% confidence interval for L-estimator of θ (ver.2, trimmed and winsorised, $\beta = 0.025$) obtained by scale estimation approach (10^5 replicates).

for skewed random projections, and shows that for these estimators, the asymptotic variance factor of order k^{-1} tends to 0 as $\alpha \to 1$, achieving infinite improvement over the geometric mean estimator for symmetric stable projections. He computes sub-optimal tails bounds for the geometric mean estimator for skewed stable projections and estimates the constants involved in those bounds in the limit as $\Delta \to 0$, showing graphically that for small Δ, the estimated values are close to the exact ones (Li, 2009).

Moreover, for $\alpha < 1$, the fractional power estimator for skewed projections has absolute moments of all orders, and thus exponential tail bounds exist, unlike in the case of the fractional power estimator for symmetric projections (Li, 2008c). Finally, the optimal quantile estimator for skewed projections is more efficient asymptotically for $\alpha > 1$ than the geometric mean estimator, having a significantly faster running time, but exhibits poor small sample performance (Li, 2008d). Although estimators are compared in terms of variance, there is no comparison to the Cramér-Rao lower bound.

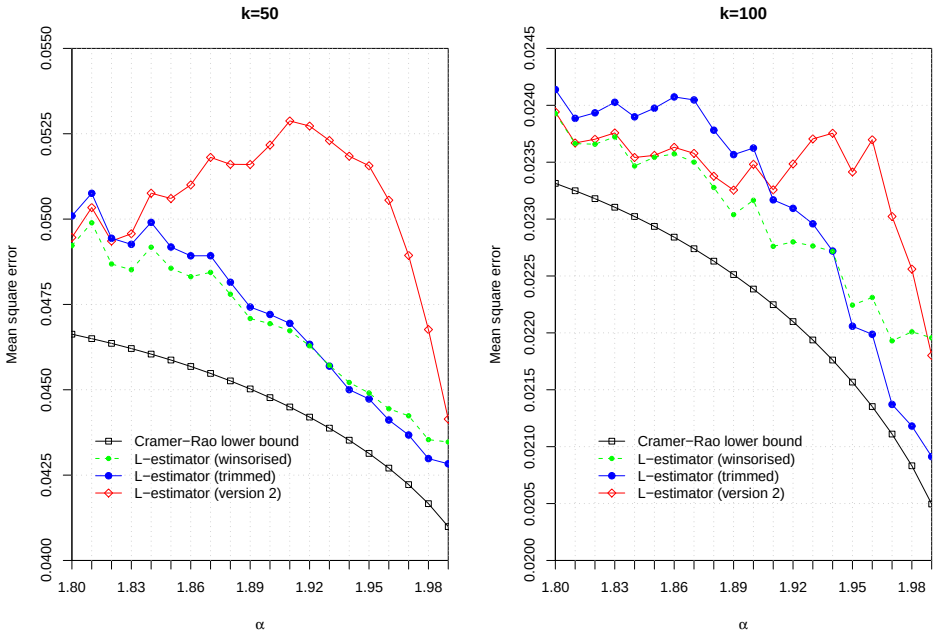


Figure 4.20: Comparison in terms of m.s.e. of the L-estimator of θ (ver. 2, trimmed and winsorised, $\beta = 0.025$, α large) obtained by scale estimation approach (10^5 replicates).

We compare the optimal fractional power estimator (Li, 2008c) in (4.5), and the optimal quantile estimator in (4.6), based on maximally skewed random projections, in Figure 4.21. We plot the asymptotic variance factor term of order k^{-1}, i.e., V satisfying $\text{var}(\hat{\theta}) = V\theta^2/k + O(k^{-2})$, alongside the corresponding term from the Cramér-Rao lower bound, i.e., the inverse of the Fisher information for a sample of size 1. We notice that for $\alpha < 1$, the asymptotic variance of the optimal fractional power estimator attains the theoretical lower bound, up to terms of order k^{-2}, whereas for $\alpha > 1$, the optimal quantile is more efficient, in particular, with small asymptotic variance for $\alpha \in (1.0, 1.6)$. As $\alpha \to 1$, the asymptotic variance factors of both estimators tend to 0, which agrees with the theoretical lower bound, and with intuition, because the l_1 norm of a vector with positive entries can be computed exactly using a counter.

Ping Li uses the fBasics package in R for evaluating density and quantile functions of the skewed stable distribution, pointing out the numerical instability of those procedures, but

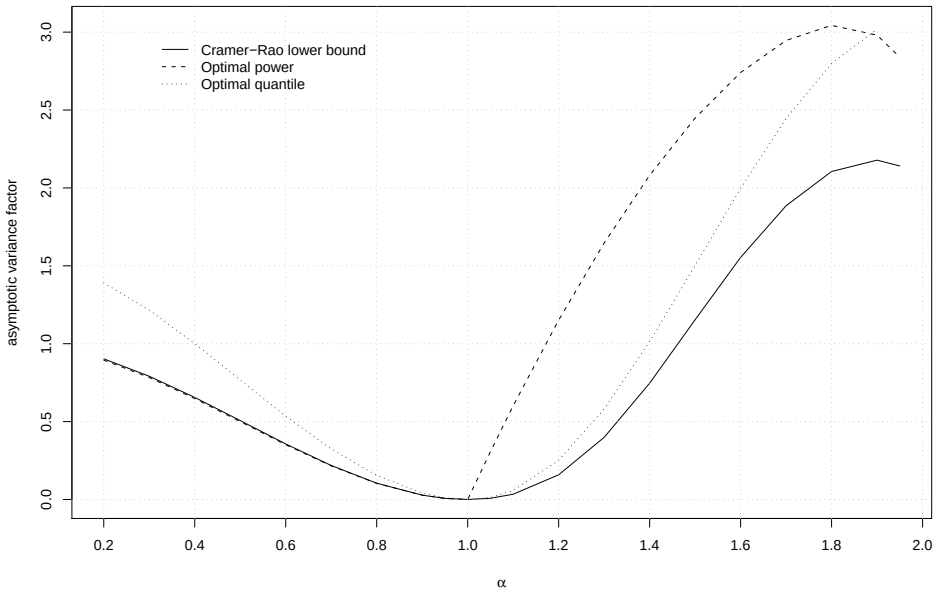


Figure 4.21: Comparison of the asymptotic variance factor of order k^{-1} for the optimal fractional power and the optimal quantile estimators against the Cramér-Rao lower bound.

does not provide computationally stable alternatives. We believe this to be a more difficult task than that of improving the procedures in the case of the symmetric stable distribution; see Section 2.5.

R estimates the density and distribution functions from integral representations given in Zolotarev (1986). For $\alpha \in (0.98, 1)$ and $\alpha \in (1, 1.02)$, the integrand functions are highly peaked and the adaptive integration procedure fails and reports an error. For $\alpha \in (0, 1)$ and $\beta = 1$, the adaptive integration procedure, which evaluates the integrals by splitting the integration region in two parts, encounters problems when the lower and upper integration bounds are equal, because the *integrate* function in R does not know how to handle such cases. We were able to resolve this by a slight modification to the existing R commands.

4.5.2 Location parameter estimation

We consider cases $\alpha \in (0,1)$ and $\alpha \in (1,2)$ separately, and apply the method of L-estimation to log-transformed variables as in Section 4.4.2. For $\alpha \in (0,1)$, the entries in $\mathbf{P}$ are independent, positive, strictly stable random variables following the distribution $S(x;\alpha)$, which is equivalent to $\big[\cos(\frac{\pi}{2}\alpha)\big]^{1/\alpha} F\big(x;\alpha,1,1,0\big)$. Consider $u, v \in V$, with coresponding rows a and b in $\mathbf{B} = \mathbf{VP}$. Then the difference between projected points has entries of the form

$$x_j := a_j - b_j = \sum_{i=1}^{m} (u_i - v_i) p_{ij} \sim F\big(x;\alpha,1,\gamma = [\cos(\pi/2\alpha)]^{1/\alpha} d_\alpha(u,v),0\big) = \gamma S(x;\alpha),$$

and the goal is to estimate $d_\alpha(u,v)$ based on the random sample $(x_1,\ldots,x_k)$ by the method of L-estimation. Consider the transformation

$$y_i := \log x_i \stackrel{\mathcal{D}}{=} \log\gamma + z_i := \mu + z_i, \quad i = 1,\ldots,k,$$

where $y_i \sim f_\mu(y) = f_0(y-\mu)$, satisfying

$$f_0(z) = e^z f\big(e^z;\alpha,1,\gamma = \big[\cos(\pi/2\alpha)\big]^{1/\alpha},0\big), \quad -\infty < z < \infty. \tag{4.25}$$

As $\alpha \to 1^-$, the distribution $S(x;\alpha) \to 1$, as shown in Figure 2.1, and $x_i \to \gamma$. So the scale parameter γ can be estimated exactly by $k^{-1}\sum_{i=1}^{k} x_i$, which agrees intuitively with the fact that the Cramér-Rao lower bound tends to 0 as $\alpha \to 1$ (see Figure 4.21). The latter can be explained by the fact that as $\alpha \to 1$, $y_i \to \log\gamma$, so $I_\mu(F_\mu)$, the Fisher information about $\mu = \log\gamma$ contained in y_i, is infinite, i.e., μ is known exactly.

Proposition 4.5.1. *The density function $f_0(z)$ defined in* (4.25) *satisfies the conditions for L-estimation.*

Proof. As $z \to \infty$, $f_0(z) \sim \exp(-\alpha z) \to 0$. As $z \to -\infty$, $\exp(z) \to 0$, so by converting to Zolotarev's parameterisation (B) and applying Theorem 2.5.2 (p. 99) from Zolotarev (1986), we obtain that $f_0(z) \sim \exp\big\{0.5\alpha z(\alpha-1)^{-1} - \exp(\alpha z(\alpha-1)^{-1})\big\} \to 0$. It remains to show that $z^2 f_0'(z) \to 0$ as $z \to \pm\infty$. In the limit as $z \to \infty$, this is obvious. As $z \to -\infty$, we use

the approximation $f_0'(z) \sim \exp\big\{-\exp(\alpha z(\alpha-1)^{-1})\big\}$, and the limiting result follows since $0 < \alpha < 1$. □

For $\alpha \in (1,2)$, the entries in $\mathbf{P}$ are independent strictly stable random variables of maximal skew, i.e., $p_{ij} \sim F(x;\alpha,1,1,0)$. Then, $x_j \sim F\big(x;\alpha,1,\gamma = d_\alpha(u,v),0\big)$, and the goal is to estimate the scale parameter γ. Consider the transformation

$$y_i := \log|x_i| \stackrel{\mathcal{D}}{=} \log\gamma + z_i := \mu + z_i, \quad i = 1,\ldots,k,$$

where $y_i \sim f_\mu(y) = f_0(y-\mu)$, satisfying

$$f_0(z) = e^z f\big(e^z;\alpha,1,1,0\big) + e^z f\big(-e^z;\alpha,1,1,0\big), \quad -\infty < z < \infty. \tag{4.26}$$

Proposition 4.5.2. *The density function $f_0(z)$ defined in* (4.26) *satisfies the conditions for L-estimation.*

Proof. As $z \to \infty$, $f(e^z;\alpha,1,1,0) \sim \exp(-z(\alpha+1))$, and $f(-e^z;\alpha,1,1,0) \sim \exp\big\{0.5(2-\alpha)z(\alpha-1)^{-1} - \exp(\alpha z(\alpha-1)^{-1})\big\}$ by converting to parameterisation (B) and applying Theorem 2.5.2. So, $f_0(z) \sim \exp(-\alpha z) + \exp\big\{0.5\alpha z(\alpha-1)^{-1} - \exp(\alpha z(\alpha-1)^{-1})\big\} \to 0$. In the limit as $z \to -\infty$, we show that $f(e^z;\alpha,1,1,0) \sim \exp(-z(1-\alpha^2))$, and similarly for $f(-e^z;\alpha,1,1,0)$, so $f_0(z) \sim \exp(z\alpha^2) \to 0$. It is now straighforward to show that $z^2 f_0'(z) \to 0$ as $z \to \pm\infty$. □

In both cases, we estimate μ by the improved bias-corrected estimator $\hat{\mu}_{BC}$ given, in finite samples, by

$$\hat{\mu}_{BC} = \sum_{i=1}^{k} w_{ik}\left(y_{(i)} - F_0^{-1}\Big(\frac{i}{k+1}\Big) + \frac{i(k-i+1)}{2(k+1)^2(k+2)}\frac{\ell'\big(F_0^{-1}(i/(k+1))\big)}{\big[f_0\big(F_0^{-1}(i/(k+1))\big)\big]^2}\right),$$

and weights

$$w_{ik} = \frac{\ell''(F_0^{-1}(i/(k+1)))}{\sum_{j=1}^{k}\ell''(F_0^{-1}(j/(k+1)))},$$

where, if $z = F_0^{-1}(i/(k+1))$,

$$\frac{i}{k+1} = \begin{cases} \mathbb{P}(W \le e^z) & \text{for } \alpha \in (0,1), \text{ and } W \sim S(w;\alpha), \\ \mathbb{P}(-e^z \le W \le e^z) & \text{for } \alpha \in (1,2), \text{ and } W \sim F(w;\alpha,1,1,0). \end{cases}$$

Proposition 4.5.3. *The L-estimator $\sum_{i=1}^k w_{ik} Y_{(i)}$ has finite variance.*

Proof. Recall that $y_i = \log|x_i| = \mu + z_i$, where $y_i \sim f_0(y-\mu)$. We have

$$\Big|\sum_{i=1}^k w_{ik} Y_{(i)}\Big| \le \sum_{i=1}^k |w_{ik} Y_{(i)}| \le k \max\{|w_{ik}|, i = 1,\ldots,k\} \max\{|Y_{(i)}|, i = 1,\ldots,k\}.$$

So we are interested in the behaviour of the order statistic $Y_{(i)}$ in the tails as $y \to \pm\infty$. If $Y_{(i)}$ has finite variance, then so does the L-estimator.

For $\alpha < 1$, we have from Proposition 4.5.1

$$f_\mu(y) \sim \begin{cases} \exp(-y\alpha) & \text{as } y \to \infty \\ \exp\big\{\alpha y(\alpha-1) - \exp(y\alpha/(\alpha-1))\big\} & \text{as } y \to -\infty. \end{cases}$$

The density of $Y_{(i)}$ is $f_{(i)}(y) = k!/[(i-1)!(k-i)!] \times \big[F_\mu(y)\big]^{i-1}\big[1-F_\mu(y)\big]^{k-i} f_\mu(y)$. As $y \to \infty$, $F_\mu(y) \sim 1 - \alpha^{-1}\exp(-\alpha y) \to 1$, so it makes no significant contribution to the density in the right tail. Hence, $f_{(i)}(y) \sim \exp(-\alpha y(k-i+1))$, and $\int_0^\infty y^2 e^{-\alpha y(k-i+1)} dy = 2[\alpha(k-i+1)]^{-3} < \infty$. As $y \to -\infty$, $F_\mu(y) \sim \exp\{-\exp(y\alpha(\alpha-1)^{-1})\} \to 0$, so $f_{(i)}(y) \sim \exp\{-i\exp(y\alpha(\alpha-1)^{-1})\}$, and

$$\begin{aligned} \int_{-\infty}^0 y^2 \exp\{-ie^{y\alpha/(\alpha-1)}\} dy &= \Big(\frac{1-\alpha}{\alpha}\Big)^3 \int_1^\infty \frac{1}{u} (\log u)^2 e^{-iu} du \\ &= \frac{1}{3}\Big(\frac{1-\alpha}{\alpha}\Big)^3 \int_1^\infty (\log u)^3 e^{-iu} du \\ &< \frac{1}{3}\Big(\frac{1-\alpha}{\alpha}\Big)^3 \int_1^\infty u^3 e^{-iu} du < \infty. \end{aligned}$$

So $Y_{(i)}$ has finite variance, and so does the L-estimator for $\alpha < 1$.

For $\alpha > 1$, the density function is

$$f_\mu(y) \sim \begin{cases} \exp(y\alpha^2) & \text{as } y \to -\infty \\ \exp\big\{0.5\alpha y(\alpha-1)^{-1} - \exp(y\alpha(\alpha-1)^{-1})\big\} & \text{as } y \to \infty. \end{cases}$$

Similarly, it can be shown that $Y_{(i)}$ has finite variance. □

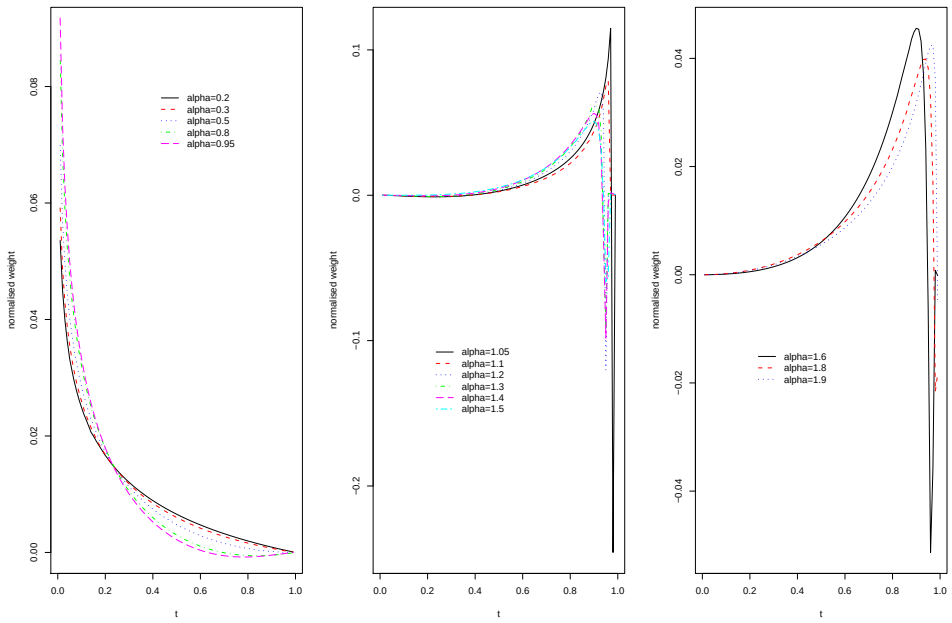


Figure 4.22: Approximate weight functions for L-estimator of location parameter from maximally skewed stable projections.

Figure 4.22 presents the weight functions for various values of α. For $\alpha < 1$, the weight function is continuous with overall decreasing trend. For $\alpha > 1$, the weight function increases smoothly in value until about $t = i/(k+1) = 0.9$, then drops sharply below zero, and increases again for t close to 1. For $\alpha \in (1.0, 1.3]$, the change in w_{ik} is particularly sharp; we suspect this to be due to a numerical instability in fBasics commands, that translates into lack of smoothness in the function $\ell''(F_0^{-1}(i/(k+1)))$. Figure 4.23 shows this behaviour in more detail; we suspect that this will result in poor small sample performance of the estimator for $\alpha \in (1.0, 1.3]$.

The asymptotically efficient L-estimator of $\theta = \gamma^\alpha$ is

$$\hat{\theta}_{BC} = \exp\left(\alpha\hat{\mu}_{BC}\right)\left(1 - \frac{\alpha^2}{2kI_\mu(F_\mu)}\right),$$

where $I_\mu(F_\mu)$ is the Fisher information about μ contained in y_1. The estimator is unbiased up to terms of order $O\left(k^{-2}\right)$, and satisfies $\sqrt{k}\left(\hat{\theta}_{BC} - \theta\right) \xrightarrow{\mathcal{D}} \text{Normal}\left(0, \alpha^2\theta^2/I_\mu(F_\mu)\right)$ as $k \to \infty$.

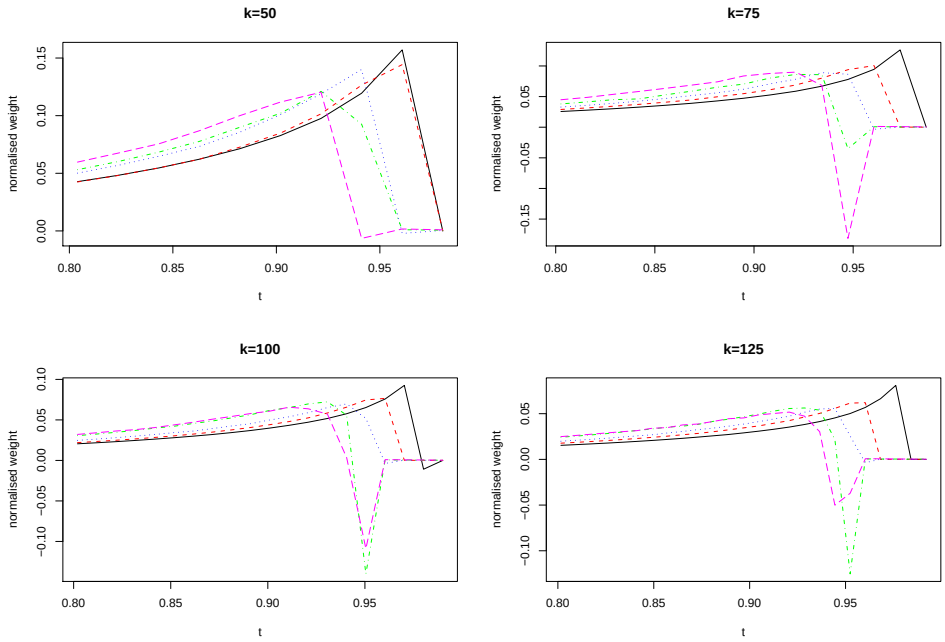


Figure 4.23: Weight function w_{ik} versus $t = i/(k+1)$, $k = 50, 75, 100, 125$, and $\alpha = 1.05$ (black), 1.1 (red), 1.15 (blue), 1.2 (green), 1.25 (magenta).

Table 4.6 gives approximations to the Fisher information

$$I_\mu(F_\mu) = -\int_0^1 \ell''(F_0^{-1}(t))dt,$$

obtained by estimating the integral via the quadrature rule with $n = 1000$ intervals of width $h = (1-2\delta)/n$, $\delta = 0.001$. We use the smaller width $h/10$ for $\alpha < 1$ and $t < 0.1$, and $h/20$ for $\alpha > 1$ and $t > 0.8$. Moreover, we use these approximations to estimate the Cramér-Rao lower bound.

Figures 4.24 ($k = 50$) and 4.25 ($k = 100$) compare the optimal fractional power, and quantile estimators of θ with the L-estimators in terms of mean square error in small samples with 10^5 replicates. Trimming (asymmetrical, removing the 5 largest order statistics) and winsorising ($p = 0.8$) for $\alpha > 1$ are applied as described in Subsection 4.4.4.

In the range $\alpha < 1$, the untrimmed L-estimator and the optimal fractional power estimator have very good performance, while the optimal quantile does not, particularly for small

α	$I_\mu(F_\mu)$	α	$I_\mu(F_\mu)$	α	$I_\mu(F_\mu)$	α	$I_\mu(F_\mu)$
0.2	0.0443	0.65	1.4810	1.15	15.9893	1.6	1.6473
0.25	0.0735	0.7	2.2433	1.2	9.0836	1.65	1.5723
0.3	0.1138	0.75	3.5671	1.25	5.9226	1.7	1.5325
0.35	0.1688	0.8	6.1151	1.3	4.2364	1.75	1.5224
0.4	0.2442	0.85	11.8643	1.35	2.8853	1.8	1.5388
0.45	0.3481	0.9	29.0110	1.4	2.6253	1.85	1.5825
0.5	0.4936	0.95	125.9015	1.45	2.2206	1.9	1.6573
0.55	0.7021	1.05	142.3186	1.5	1.9500	1.95	1.7761
0.6	1.0095	1.1	35.7971	1.55	1.7677		

Table 4.6: Approximate Fisher information values $I_\mu(F_\mu)$ tabulated for values of $\alpha \in [0.2, 2)$.

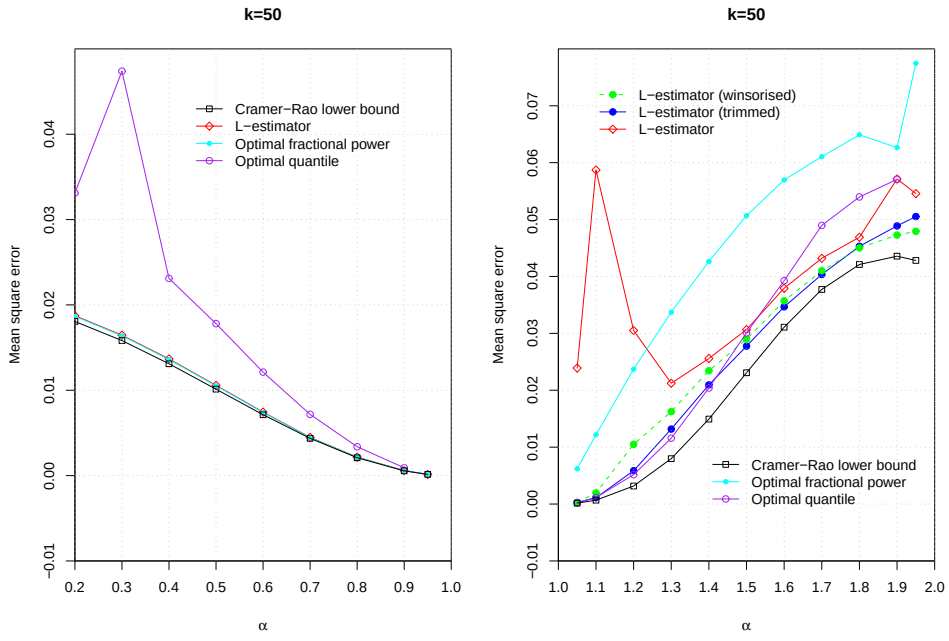


Figure 4.24: Comparison in terms of m.s.e. of L-estimators, optimal fractional power, and optimal quantile estimators of θ. The Cramér-Rao lower bound is plotted for comparison.

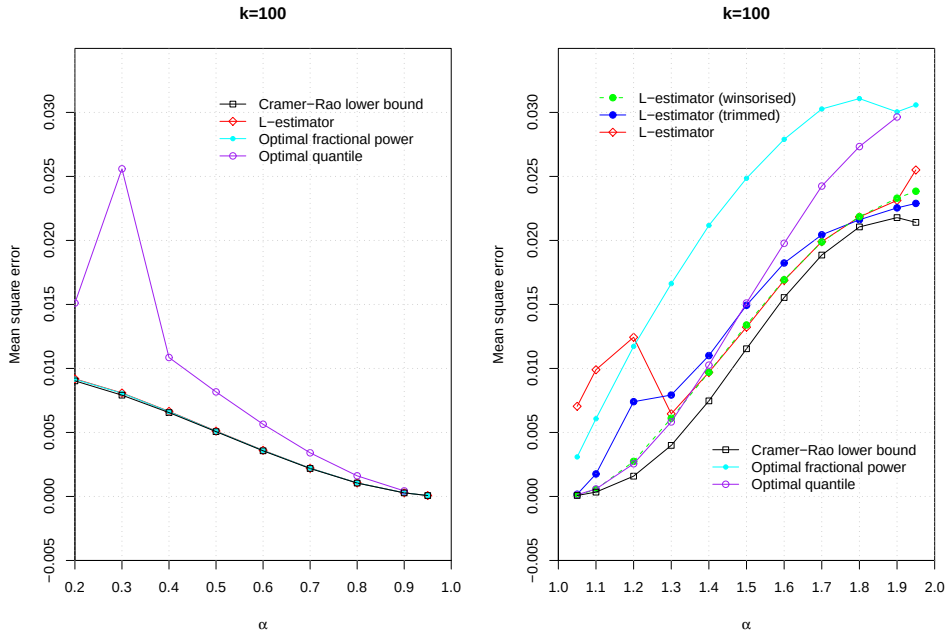


Figure 4.25: Comparison in terms of m.s.e. of L-estimators, optimal fractional power, and optimal quantile estimators of θ. The Cramér-Rao lower bound is plotted for comparison.

α and α close to 1. For $\alpha = 0.95$ we obtain a mean square error of 63 for the optimal quantile estimator and exclude the point from the plot; this is similar to observations in Li (2008d).

For $\alpha > 1$, the trimmed and winsorised L-estimators have good small sample performance with $k = 50$, as does the optimal quantile estimator for $\alpha \in (1.0, 1.5]$. The untrimmed L-estimator performs poorly for $\alpha < 1.3$ due to the behaviour of the weight function observed in Figure 4.23.

For $k = 100$, the L-estimator ($\alpha \geq 1.3$) and the winsorised L-estimator have very good performance, surpassed slightly by that of the trimmed L-estimator for $\alpha \in [1.8, 2.0)$. For $\alpha \in (1.0, 1.3)$, the L-estimator continues to have poor performance, but the m.s.e. values are not as large as observed when $k = 50$. For $\alpha > 1$, the quantile estimator is defined with values $q \in [0.778, 0.855]$, which suggests that trimming the top 5 order statistics, or winsorising with $p = 0.8$ will not result in very large loss in efficiency.

Figure 4.26 compares the length of the 95% confidence interval for the L-estimators

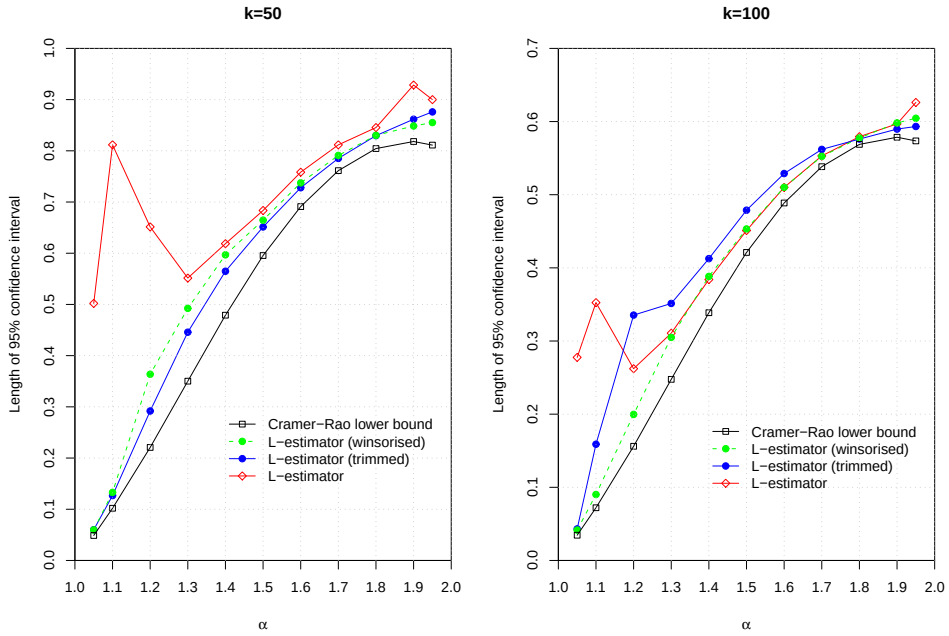


Figure 4.26: Comparison of length of 95% confidence interval for L-estimators (trimmed and winsorised) of θ (10^5 replicates).

(trimmed and winsoried) computed from 10^5 simulations, against the lower bound of $2 \times 1.96 \times \alpha/\sqrt{k \times I_\mu(F_\mu)}$. For $k = 50$, the winsorised and trimmed estimators have confidence intervals of comparable length, whereas for $k = 100$, the winsorised estimator has shorter confidence intervals than the trimmed one. We explain the latter situation by pointing out that when $k = 100$, the trimmed estimator removes only the top 5% of order statistics, whereas the winsorised estimator brings in the top 10%.

4.6 Summary

In this chapter we consider the problem of distance-preserving dimension reduction with the aim of recovering l_α distances (quasi-distances) for $\alpha \in (0, 2]$ via data sketching with weighted linear combinations of α-stable random variables. We discuss data sketching to symmetric, as well as maximally skewed stable random variables, and apply different transformations,

reducing the problem to one of location or scale estimation. We propose asymptotically efficient estimators based on the method of L-estimation, and analyse in great detail the small sample performance of these estimators, showing that it can be improved by trimming or winsorising with minimal loss of efficiency. Furthermore, we show via simulations that our estimators outperform competing estimators in terms of performance in small samples. This is particularly important for dimension reduction where the goal is to find low-dimensional projections from which the quantities of interest, in our case, l_α distances and quasi-distances, can be well approximated.

Chapter 5

Entropy estimation and MCMC convergence assessment

Entropy estimation is a useful tool for summarising changes in distribution of streaming data over time, or across fixed windows of time, and many algorithms have been developed for this purpose (Bhuvanagiri and Ganguly, 2006; Chakrabarti et al., 2006; Harvey et al., 2008; Li, 2008e). Areas of application include network traffic monitoring (Lall et al., 2006), analysis of sequences of neural signals, known as spike trains (Shlens et al., 2007), and visual tracking in video streams (Schraudolph, 2004).

We propose two estimators for empirical entropy over streaming data based on skewed stable random projections: the first is an asymptotically efficient estimator based on the method of L-estimation, and the second is a near-optimal, equally-weighted mean estimator. Both estimators have good small sample performance, and improve upon existing estimators of entropy.

In the second part of the chapter, we consider the problem of convergence assessment of MCMC algorithms for simulating from complex, high-dimensional, discrete distributions. The MCMC output chain can be viewed as a data stream defining a vector of empirical probabilities over the observations, and it is interesting to describe the evolution of the

chain by summary statistics. We argue that monitoring the behaviour of summary statistics such as cardinality, entropy, and l_α distances between independent chains with different starting points may be a valuable tool for convergence assessment.

5.1 Entropy and streaming data

The classical concept of entropy, known as Shannon's entropy, measures the randomness or uncertainty associated with a discrete random variable X. Let $X \sim p$ have probability mass function $p_i = \mathbb{P}(X = i) > 0$, $i = 1, \ldots, n$, satisfying $\sum_{i=1}^n p_i = 1$. Denote the entropy of X by $H(p)$, defined by

$$H(p) = \sum_{i=1}^{n} p_i \log \frac{1}{p_i}, \tag{5.1}$$

where $H(p) \leq \log n$ is maximal when $X \sim \text{Unif}(\{1, \ldots, n\})$. This concept was introduced in the field of information theory by the paper of Shannon and Weaver (1949) as a measure of the average information content of a message X that is missing when the value of X is unknown. It is related to notions of entropy existing in other areas of science and engineering, most notably in thermodynamics and quantum mechanics.

Rényi (1961) characterises Shannon's entropy and proposes another equivalent measure:

$$H_\alpha(p) = \frac{1}{1-\alpha} \log \Big(\sum_{i=1}^{n} p_i^\alpha \Big), \tag{5.2}$$

for $\alpha > 0$, where Shannon's entropy is the limiting case as $\alpha \to 1$, i.e., $H(p) = \lim_{\alpha \to 1} H_\alpha(p)$. Tsallis (1988) also proposes an equivalent measure of information defined as

$$S_\alpha(p) = \frac{1}{\alpha - 1} \Big(1 - \sum_{i=1}^{n} p_i^\alpha \Big), \tag{5.3}$$

satisfying $H(p) = \lim_{\alpha \to 1} S_\alpha(p)$.

In the context of non-negative data streams, the accumulation vector $\mathbf{a}$ of the stream defines the empirical distribution of the data types observed up to time T. Define the empirical probabilities $p_i = a_i / \sum_{j \in \mathcal{D}} a_j$, and $H(p)$ is the corresponding empirical entropy.

Expressions (5.2) and (5.3) show that the problem of estimating Shannon's entropy is closely related to that of estimating the l_α norm (quasi-norm) of the vector of empirical probabilities. Li (2008e) proposes a two-stage approximation procedure for estimating the entropy over streaming data: first, estimate the l_α norm (quasi-norm) by compressed counting and projection to maximally skewed, α-stable random variables (Li, 2009), and second, estimate Shannon's entropy in (5.1) by either $H_\alpha(p)$ or $S_\alpha(p)$ with $\alpha \approx 1$. The second stage poses the question of how small should α be in practice, and introduces an additional source of error in the approximation procedure. We propose an alternative approach to estimating the entropy not in the limit as $\alpha \to 1$, but directly with $\alpha = 1$, that reduces the problem to that of estimating the location parameter of transformed, maximally skewed stable random variables of index 1.

For further motivation, consider an experiment resulting in data x, and let θ parameterise the distribution on the space of the experiment, denoted by $p(x \mid \theta)$. Lindley (1956) proposes a measure of the information provided by the experiment with prior distribution $p(\theta)$: $H(p(\theta \mid x)) - H(p(\theta))$. Assume that the prior and posterior are discrete distributions defined over prohibitively large spaces such that their entropies cannot be computed exactly. Suppose, furthermore, that it is possible to simulate independent draws from these distributions. Then, imagine running two independent streams of random draws from $p(\theta \mid x)$ and $p(\theta)$, respectively. The accumulation vectors of the streams define the empirical distributions, that, as the streams run to infinity, converge to the true distributions, and the difference in empirical entropies is an estimator of the information provided by the experiment, $H(p(\theta \mid x)) - H(p(\theta))$.

We consider the problem of assessing convergence of MCMC algorithms designed to simulate from a complex, high dimensional discrete distribution π, in particular, to assess whether the output is representative of the high density regions of π. By definition, an MCMC chain whose elements converge rapidly to weakly correlated draws from the stationary distribution

π is said to possess good mixing speed. See, for example, Robert and Casella (2004). One of the challenges of MCMC convergence assessment is detecting when a chain is trapped in regions of high local density. If π is low dimensional, the failure to escape from a particular region of the support of π can usually be detected from plots of marginal density functions from several independent MCMC output chains; however, in high dimensions, there is no equivalent graphical method for assessing lack of convergence. We propose to compare the output of several MCMC samplers with different starting points by computing and monitoring the behaviour of summary statistics such as cardinality, entropy, and l_α distances. Intuitively, unreasonably small cardinality estimates may be an indication that the chain is trapped in a region of high local density, whereas convergence of estimates of entropy and l_α distances, over long runs of the algorithm, to a common value may be indicative of stationarity. In Section 5.3, we illustrate this with simulations of the posterior distribution of a decomposable Gaussian graphical model.

5.2 Estimating the empirical entropy

5.2.1 Introduction

Let $X_\alpha \sim S(x;\alpha)$, for fixed $0 < \alpha < 1$, or equivalently, $X_\alpha \sim F\big(x;\alpha,1,\gamma = [\cos(\frac{\pi}{2}\alpha)]^{1/\alpha},0\big)$. Let $p = (p_1,\ldots,p_n)$ be a vector of probabilities with $\sum_{i=1}^n p_i = 1$. Let $\big(X_\alpha^{(1)},\ldots,X_\alpha^{(n)}\big)$ be a vector of independent copies of X_α. Since

$$\sum_{i=1}^n p_i X_\alpha^{(i)} \stackrel{\mathcal{D}}{=} \Big(\sum_{i=1}^n p_i^\alpha\Big)^{1/\alpha} X_\alpha,$$

and, as $\alpha \to 1$

$$\frac{1}{\alpha-1}\Big(1-\sum_{i=1}^n p_i^\alpha\Big) \to H(p_1,\ldots,p_n) = -\sum_{i=1}^n p_i \log p_i, \tag{5.4}$$

Li (2008e) estimates the entropy by approximating the left hand side of (5.4) with $\alpha \approx 1$.

Our method is based on the limiting distribution of a transformed variable X_α as $\alpha \to 1$.

We work with parameterisation S0 which is continous in all parameters, and has the property that δ and γ have the interpretation of location and scale for all values of α and β.

Under parameterisation (A), the characteristic function of X_α has the form

$$\begin{aligned}\phi(t) &= \exp\Big\{\cos(\pi/2\alpha)\big[-|t|^\alpha + it|t|^{\alpha-1}\tan(\pi/2\alpha)\big]\Big\} \\ &= \exp\Big\{\cos(\pi/2\alpha)\big[-|t|^\alpha + it\omega_M(t,\alpha,1) + it\tan(\pi/2\alpha)\big]\Big\},\end{aligned}$$

equal to the c.f. under parameterisation (M) with parameters $\alpha_M = \alpha$, $\beta_M = 1$, $\gamma_M = \cos(\pi/2\alpha)$, and $\delta_M = \tan(\pi/2\alpha)$. To transform to parameterisation S0, we use expression (2.5) and (2.6) to obtain $\alpha_0 = \alpha$, $\beta_0 = 1$, $\gamma_0 = \big[\cos(\pi/2\alpha)\big]^{1/\alpha}$, and $\delta_0 = \big[\cos(\pi/2\alpha)\big]^{\alpha^{-1}} \times \tan(\pi/2\alpha)$. Let the subscript 0 denote parameterisation S0, and write $X_\alpha \sim F_0(x;\alpha,1,\gamma_0,\delta_0)$.

We transform X_α such that the entropy $-\sum_{i=1}^n p_i \log p_i$ can be recoved as the location parameter of a sum of independent tranformed variables weighted by p_i. Consider the following transformation:

$$Y_\alpha = \big[\pi/2\sin(\pi/2\alpha)\big]^{1/\alpha} \times \frac{X_\alpha - \delta_0}{\gamma_0} - \log(2/\pi) \sim F_0\big(x;\alpha,1,[\pi/2\sin(\pi/2\alpha)]^{1/\alpha}, -\log(2/\pi)\big),$$

with limiting distribution $F_0\big(x;1,1,\pi/2,-\log(2/\pi)\big)$ as $\alpha \to 1$, and characteristic function

$$\phi(t) = \exp\big(-|t|\pi/2 - it\log|t|\big) = (-it)^{-it}.$$

If we can show that Y_α has a moment generating function $M_\alpha(t)$ defined for t is a neighbourhood of 0, then $Y_1 = \lim_{\alpha\to 1} Y_\alpha$ has a moment generating function satisfying $M_1(t) = \phi(-it)$.

$$\begin{aligned}M_\alpha(t) &= \exp\Big(-t\big[\big(\pi/2\sin(\pi\alpha/2)\big)^{1/\alpha}\delta_0\gamma_0^{-1} + \log(2/\pi)\big]\Big)\mathbb{E}\exp\Big(t\big(\pi/2\sin(\pi\alpha/2)\big)^{1/\alpha}\frac{X_\alpha}{\gamma_0}\Big) \\ &= \exp\Big(\tan(\pi\alpha/2)\big[-t\big(\pi/2\sin(\pi\alpha/2)\big)^{1/\alpha} - \pi/2(-t)^\alpha\big] - t\log(2/\pi)\Big),\end{aligned}$$

provided $t < 0$, where the second equality holds using the expression of the Laplace transform from Theorem 2.2.1. In the limit, $M_1(t) = \lim_{\alpha\to 1} M_\alpha(t) = (-t)^{-t} = \phi(-it)$.

The problem of estimating the entropy reduces to that of approximating the location parameter of a stable random variable as follows. Suppose it is possible to simulate exactly

from p. Start a stream of i.i.d. observations from p, and define a vector $(x_1, \ldots, x_k) = (0, \ldots, 0)$, for some pre-determined k. At time t, we observe $i_t \sim p$, and use i_t to set the seed of the pseudo-random number generator. Then, generate independent random variables $h_j(i_t) \sim F_0\big(x; 1, 1, \pi/2, -\log(2/\pi)\big)$, and update the vector by $x_j \rightarrow x_j + h_j(i_t)$, for $j = 1, \ldots, k$.

At time $t = T$, we want to approximate the empirical entropy $H(\hat{p})$ as estimator of $H(p)$, where $\hat{p}_i = T^{-1} \sum_{t=1}^{T} \mathbb{I}(i_t = i)$ is the empirical probability, $i = 1, \ldots, n$. For $j = 1, \ldots, k$, define

$$y_j = \frac{1}{T} x_j = \frac{1}{T} \sum_{i=1}^{n} \sum_{t=1}^{T} \mathbb{I}(i_t = i) h_j(i) = \sum_{i=1}^{n} \hat{p}_i h_j(i);$$

$(y_1, \ldots, y_k)$ forms a random sample of maximally skewed, strictly stable random variables with $\alpha = 1$ and c.f.

$$\begin{aligned}
\phi(t) &= \mathbb{E} \exp\Big(it \sum_{i=1}^{n} \hat{p}_i h_j(i) \Big) \\
&= \prod_{i=1}^{n} \exp\Big(-|\hat{p}_i t| \pi/2 - i \hat{p}_i t \log(|\hat{p}_i t|) \Big) \\
&= \exp\Big(-\pi/2 |t| \sum_{i=1}^{n} \hat{p}_i - it \sum_{i=1}^{n} \hat{p}_i [\log|t| + \log \hat{p}_i] \Big) \\
&= \exp\Big(\pi/2 \big[-|t| - it 2/\pi \log|t| \big] - it \sum_{i=1}^{n} \hat{p}_i \log \hat{p}_i \Big),
\end{aligned}$$

since $\sum_{i=1}^{n} \hat{p}_i = 1$. Therefore, $\big(y_1, \ldots, y_k\big)$ is a random sample of stable variables following the distribution $F\big(y; 1, 1, \pi/2, -\sum_{i=1}^{n} \hat{p}_i \log \hat{p}_i\big)$, where the location parameter $\delta = -\sum_{i=1}^{n} \hat{p}_i \log \hat{p}_i$ is the entropy estimate based on the empirical distribution of the observations up to time T. This procedure also applies if the observations come from a simple data stream in the form $(i_t, 1)$, and can be easily extended to non-negative data streams.

5.2.2 L-estimator for location parameter

We estimate the entropy $-\sum \hat{p}_i \log \hat{p}_i$ by the method of L-estimation from a random sample $y_1, \ldots, y_k \sim F\big(y; 1, 1, \pi/2, \delta = -\sum \hat{p}_i \log \hat{p}_i\big)$. Write $y_i = \delta + z_i$, where $z_i \sim f\big(z; 1, 1, \pi/2, 0\big)$ with support on $(-\infty, \infty)$. Let $f_0(z)$ and $F_0(z)$ denote the density and distribution functions of z_i; then $y_i \sim f_\delta(y) = f_0(y - \delta)$ with c.d.f. $F_\delta(y)$.

Proposition 5.2.1. *The density function $f_0(z) = f\big(z; 1, 1, \pi/2, 0\big)$ satisfies the conditions of L-estimation.*

Proof. As $z \to \infty$,

$$f_0(-z) = f\big(-z; 1, 1, \pi/2, 0\big) = f\big(z; 1, -1, \pi/2, 0\big) = f_A\big(z; 1, -1, \pi/2, 0\big) = f_B\big(z; 1, -1, 1, 0\big),$$

where the subscripts A and B denote parameterisations A and B, respectively, in Zolotarev (1986). From Theorem 2.5.2 of this book (p. 99), $f_B(z; 1, -1, 1, 0) \sim (2\pi)^{-1/2} \exp\big((z - 1)/2 - \exp(z - 1)\big)$ as $z \to \infty$. Letting $z \to -\infty$, $f_0(z) \sim \exp\big(-(z + 1)/2 - \exp(-z - 1)\big)$, which is dominated by $\exp(-\exp(-z)) \to 0$. Similarly, $z^2 f_0'(z) \sim z^2 \exp\big((-z - 1)/2 - \exp(-z - 1)\big)\big[-1/2 + \exp(-z - 1)\big]$ is dominated by $z^2 \exp(-\exp(-z))$ which converges to 0 as $z \to -\infty$. Moreover, as $z \to \infty$,

$$f_0(z) = f\big(z; 1, 1, \pi/2, 0\big) = f_A\big(z; 1, 1, \pi/2, 0\big) = f_B\big(z; 1, 1, 1, 0\big) \sim z^{-2}$$

by Theorem 2.5.4 in Zolotarev (1986) (p. 101), so $f_0(z) \to 0$ and $z^2 f_0'(z) \to 0$ as $z \to \infty$. □

Following Subsection 4.4.2, the bias-corrected L-estimator of δ is given by:

$$\hat{\delta}_{BC} = \sum_{i=1}^{k} w_{ik} \left(y_{(i)} - F_0^{-1}\Big(\frac{i}{k+1}\Big) + \frac{i(k-i+1)}{2(k+1)^2(k+2)} \frac{\ell'\big(F_0^{-1}(i/(k+1))\big)}{\big[f_0\big(F_0^{-1}(i/(k+1))\big)\big]^2} \right), \tag{5.5}$$

where $\ell(y) = \log f_0(y)$, and the weights are approximated as in equation (4.8). Furthermore, $\sqrt{k}\big(\hat{\delta}_{BC} - \delta\big) \xrightarrow{\mathcal{D}} \text{Normal}(0, 1/I_{11})$ as $k \to \infty$, where I_{11} is the Fisher information about δ contained in y_1.

Proposition 5.2.2. *The L-estimator* $\sum_{i=1}^{k} w_{ik} Y_{(i)}$ *has infinite variance, unless* $w_{kk} = 0$.

Proof. We are interested in the tail behaviour of the ith order statistic $Y_{(i)}$. From Proposition 5.2.1, the density function satisfies

$$f_\delta(y) \sim \begin{cases} y^{-2} & \text{as } y \to \infty \\ \exp\big\{ -(y+1)/2 - \exp(-(y+1)) \big\} & \text{as } y \to -\infty. \end{cases}$$

As $y \to -\infty$, $F_\delta(y) \sim \exp\{-e^{-(y+1)}\} \to 0$, so $f_{(i)}(y) \sim \exp\{-ie^{-(y+1)}\}$, and

$$\int_{-\infty}^{0} y^2 \exp\{-ie^{-(y+1)}\} dy = \int_{e^{-1}}^{\infty} \big[(\log u)^2 + 2\log u + 1 \big] \frac{1}{u} e^{-iu} du < \infty.$$

As $y \to \infty$, $F_\delta(y) \sim 1 - 1/y \to 1$, so $f_{(i)} \sim y^{-(k-i+2)}$, and $\int_0^\infty y^2 y^{-(k-i+2)} dy < \infty$, provided $i < k$. So, if $w_{kk} = 0$, i.e., the largest order statistic is trimmed, then the estimator has finite variance. □

We compute the Fisher information in two ways:

$$I_{11} = \int_{-\infty}^{\infty} \Big(\frac{\partial}{\partial \delta} \log f_\delta(y) \Big)^2 f_\delta(y) dy = \int_0^1 \Big[\ell'\big(F_0^{-1}(t)\big) \Big]^2 dt, \tag{5.6}$$

and

$$I_{11} = \int_{-\infty}^{\infty} -\frac{\partial^2}{\partial \delta^2} \log f_\delta(y) f_\delta(y) dy = -\int_0^1 \ell''\big(F_0^{-1}(t)\big) dt. \tag{5.7}$$

Figure 5.1 displays the integrand functions in expressions (5.6) (top row) and (5.7) (bottom row), alongside the corresponding transformed integrands according to transformation $t \to t^{1/3}$. We estimate the Fisher information by integrating the transformed function over the region $[0.00001, 0.3)$, and the untransformed one over $[0.3, 0.998]$ using quadrature rule with 5000 equally spaced points. The approximations to (5.6) and (5.7) agree to within 4 significant digits, giving $I_{11} = 0.3445$.

Figure 5.2 displays the normalised weight function w_{ik} and the quantile q defining the optimal quantile estimator:

$$\hat{\delta}_q = q - Quantile\{y_i, i = 1, \ldots, k\} - F_0^{-1}(q) - B_k,$$

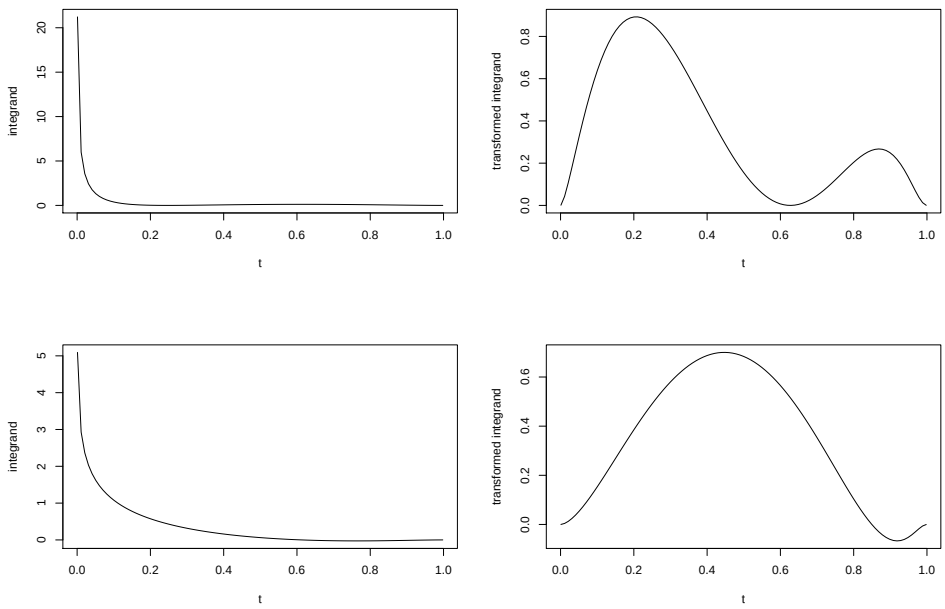


Figure 5.1: Left column: approximations to integrand in expressions (5.6) (top row) and (5.7) (bottom row). Right column: approximations to transformed integrand according to transformation $t \to t^{1/3}$.

where B_k is the small sample bias estimated via simulations at $\delta = 0$. The optimal q is chosen to minimise the asymptotic variance $q(1-q)\big[f_0(F_0^{-1}(q))\big]^{-2}$ (David and Nagaraja, 2003); this value is $q = 0.09466$, and the corresponding estimator has 67% relative efficiency compared to the MLE. The optimal quantile estimator for the location parameter is similar to the optimal quantile estimator for the scale parameter in (4.6), proposed by Li (2008a).

We remark that the weight w_{ik} is a continuous function with overall decreasing trend. Moreover, the small value of the optimal quantile, $q = 0.09466$, indicates that little information would be lost by trimming or winsorising in the upper tail of the distribution of the data $y_1, \ldots, y_k$. Unlike previously discussed cases of L-estimation where sharp oscillations in weight function corresponding to large order statistics resulted in perturbations in m.s.e., in this situation we have a smooth weight function, but the weighted sum $\sum_{i=1}^{k} w_{ik} y_{(i)}$ has infinite variance (see Proposition 5.2.2). Hence, the m.s.e. computed from simulations is

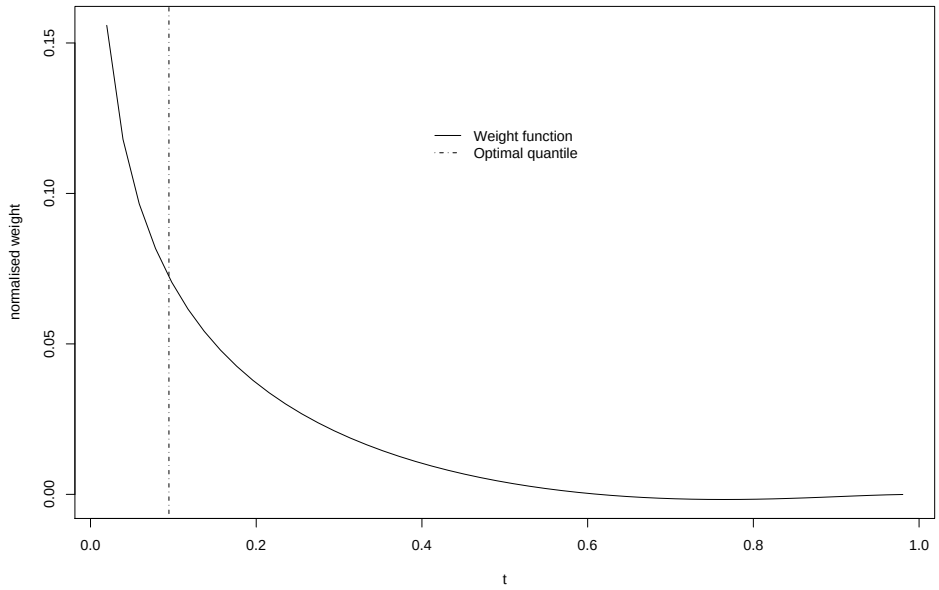


Figure 5.2: This plot displays the normalised weight function w_{ik} versus $t := i/(k+1)$, $i = 1, \ldots, k$ with $k = 50$. The vertical line at $t = 0.09466$ indicates the quantile defining the optimal quantile estimator.

very large, but trimming and winsorising is shown to reduce the m.s.e. to the Cramér-Rao lower bound. Proposition 5.2.2 implies that trimming the largest order statistic sufficies to make the variance of the L-estimator finite.

Figure 5.3 displays the m.s.e. of the L-estimator (trimmed and winsorised versions), and of the optimal quantile estimator, for sample sizes $k \in [10, 150]$. For $k = 10$, we obtain best results by trimming 70% of the largest observations, and winsorising with $p = 0.7$. For $k = 20$, we use trimming with $p = 0.8$. For all other values of k, we trim the largest 10% of observations, and winsorise with $p = 0.9$. Overall, the L-estimator performs better than the optimal quantile estimator; for $k \geq 30$, the m.s.e. of the trimmed L-estimator is very close to the theoretical lower bound.

Figure 5.4 compares the length of the 95% confidence interval of the L-estimators, showing that we can expect the estimates to lie close to the true value, particularly for sample sizes

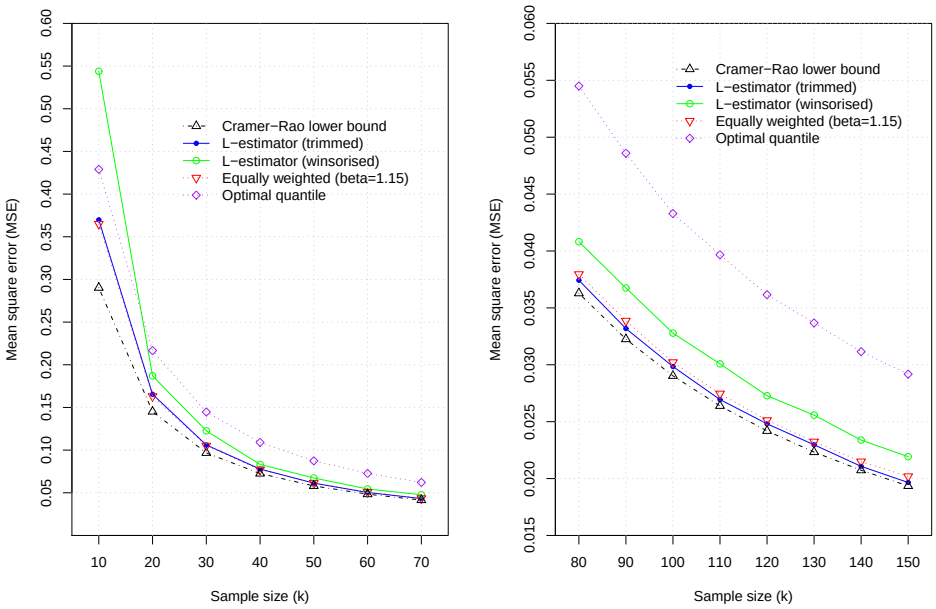


Figure 5.3: Comparison in terms of m.s.e. of L-estimator of δ (trimmed and winsorised), optimal quantile, and equally weighted estimator (10^5 replicates), alongside the Cramér-Rao lower bound. The latter estimator is introduced in Subsection 5.2.3.

$k \geq 30$. Again, the trimmed L-estimator outperforms the winsorised one.

5.2.3 Equally weighted estimator for scale parameter

Consider again the random sample $(y_1, \ldots, y_k)$ with $y_i = \delta + z_i$, where $z_i \sim f\big(z; 1, 1, \pi/2, 0\big)$. We exploit the property that, for $t < 0$, $\mathbb{E}\big(e^{tZ}\big) = (-t)^{(-t)}$, shown in Subsection 5.2.1, to derive a near-optimal, equally-weighted estimator of δ.

Consider the transformation $w_i = \exp(-y_i) = \exp(-(\delta + z_i))$, that reduces the problem to one of estimating the scale parameter $\gamma = \exp(-\delta)$. The mean $k^{-1}\sum_{i=1}^{k} w_i$ is an unbiased estimator of $\exp(-\delta)$ with variance $3\exp(-2\delta)/k$. The Fisher information about $\exp(-\delta)$ contained in w_1 is $I_{11} \times \big(\partial\delta/\partial\gamma\big)^2 = 0.3445\exp(-2\delta)$, where I_{11} is the Fisher information about δ in y_1. Therefore the asymptotic efficiency of the mean estimator relative to the MLE is $(0.3445 \times 3)^{-1} \approx 0.967$.

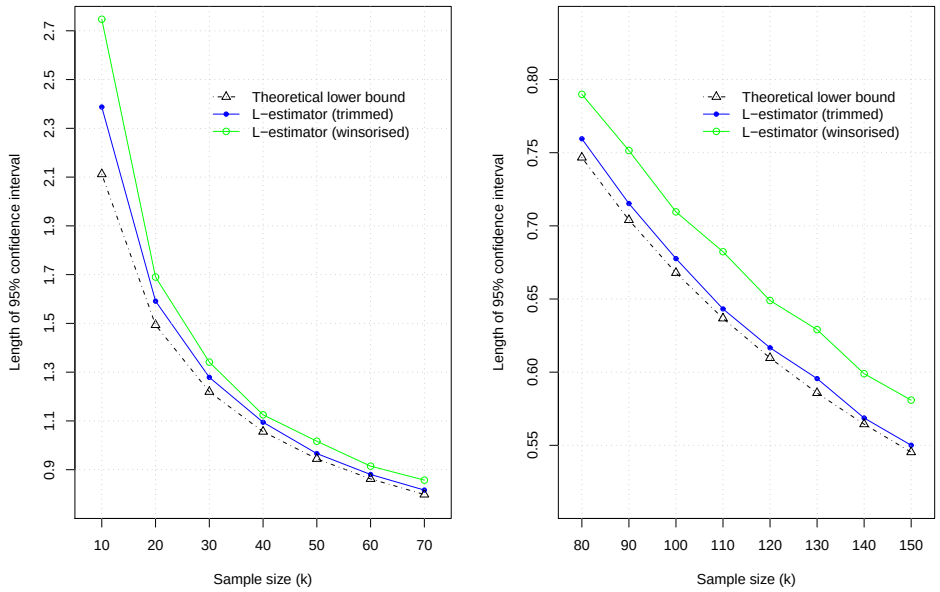


Figure 5.4: Comparison in terms of length of 95% confidence interval for L-estimator of δ (trimmed and winsorised), alongside the theoretical lower bound.

Similarly, consider the transformation

$$w_i = \exp(-\beta y_i) = \exp(-\beta(\delta + z_i)),\ \beta > 0,$$

and find the value of β that maximises the asymptotic relative efficiency. Define $\gamma = \exp(-\beta\delta)$. Since $\mathbb{E}W = \beta^{\beta}\gamma$, the estimator $\hat{\gamma} = \beta^{-\beta}k^{-1}\sum_{i=1}^{k} w_i$ is unbiased for γ, having variance $k^{-1}\gamma^2(4^{\beta} - 1)$. The Fisher information about γ contained in w_i equals $\beta^{-2}\gamma^{-2}I_{11}$, so the asymptotic efficiency of $\hat{\gamma}$ compared to the MLE is $\beta^2(4^{\beta} - 1)^{-1}0.3445^{-1}$. The latter is maximised when $\beta \approx 1.15$, attaining a value of 0.978.

In the sequel, $\beta = 1.15$. Let $\hat{\delta} = -\beta^{-1}\log\hat{\gamma}$ be the estimator of δ. In small samples, this estimator has an additive bias

$$BC = \mathbb{E}\hat{\delta} - \delta = -\frac{1}{\beta}\mathbb{E}\log\left(\beta^{-\beta}k^{-1}\sum_{i=1}^{k} e^{-\beta Z_i}\right)$$

that can be estimated from simulations. We propose the bias-corrected estimator $\hat{\delta}_{BC} = \hat{\delta} - BC$. Figure 5.3 shows that the equally weighted estimator $\hat{\delta}_{BC}$ performs nearly as well as

the trimmed L-estimator in terms of mean square error in small samples. Moreover, by the Central Limit Theorem, $\sqrt{k}\Big(k^{-1}\sum_{i=1}^{k} w_i - \beta^{\beta}\gamma\Big) \xrightarrow{\mathcal{D}} \text{Normal}\big(0, \gamma^2\beta^{2\beta}(4^{\beta}-1)\big)$ as $k \to \infty$, and by the Delta method, it follows that $\sqrt{k}\big(\hat{\delta} - \delta\big) \xrightarrow{\mathcal{D}} \text{Normal}\big(0, \beta^{-2}(4^{\beta}-1)\big)$ as $k \to \infty$.

5.3 MCMC convergence assessment

We consider the problem of convergence assessment of MCMC methods, in particular the Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970), for simulating from the posterior distribution in Gaussian graphical models. We begin by introducing terminology and basic results on graphical models; for more details, see Lauritzen (1996).

5.3.1 Decomposable Gaussian graphical models

Let $\mathbf{y} = (y_1, \ldots, y_n)^T$ be a random sample from the multivariate Gaussian distribution Normal$(\mathbf{0}, \Sigma)$ in p-dimensions, where $\mathbf{0} = (0, \ldots, 0)^T$, and Σ is a $p \times p$ positive-definite matrix known as the variance-covariance matrix. Let $X_1, \ldots, X_p$ represent the variables whose joint distribution is multivariate Gaussian. The conditional independence structure of this distribution can be represented graphically by an undirected graph $G = (V, \mathcal{E})$, where $V = \{1, \ldots, p\}$ is the set of vertices or nodes, where node i represents variable X_i, and $\mathcal{E}$ is the set of edges. By definition, an edge $(i, j) \in \mathcal{E}$ if and only if X_i is independent of X_j given variables X_k, $k \neq i, j$. Moreover, the edges of the graph correspond to nonzero elements in the precision matrix $K = \Sigma^{-1}$.

Let $\mathcal{M}(G) = \{K; \text{positive definite } p \times p \text{ matrix } K = (k_{ij}) \text{ satisfying } k_{ij} = 0 \iff (i, j) \notin \mathcal{E}\}$. The likelihood of the data $\mathbf{y}$ is given by

$$p(\mathbf{y}|\Sigma, G) = (2\pi)^{-np/2}(\det \Sigma)^{-n/2} \operatorname{etr}\Big(-\frac{1}{2}\Sigma^{-1}\mathbf{y}\mathbf{y}^T\Big), \quad \text{for } \Sigma^{-1} \in \mathcal{M}(G),$$

where $\operatorname{etr}(A) = \exp(\operatorname{trace}(A))$. If the graph G is decomposable, then the likelihood factorises over the prime components and separators of the graph, rendering the distribution

particularly tractable.

In Bayesian statistics, interest lies in simulating from the posterior distribution

$$p(G, \Sigma|\mathbf{y}) = p(\Sigma|G, \mathbf{y})p(G|\mathbf{y}) \propto p(\mathbf{y}|\Sigma, G)p(G, \Sigma), \tag{5.8}$$

for particular prior distribution $p(G, \Sigma) = p(\Sigma|G)p(G)$. We follow the paper of Jones et al. (2005) in setting the priors on decomposable graph G with corresponding variance-covariance matrix Σ. In particular, we place a hyper-inverse Wishart prior $p(\Sigma|G)$, denoted $\text{HIW}(G, \delta, \Phi)$, where Φ is a positive-definite $n \times n$ matrix and $\delta > 0$. This prior factorises over the prime components and separators of G. Moreover, it is conjugate, and the posterior $p(\Sigma|G, \mathbf{y})$ is $\text{HIW}(G, \delta^\star, \Phi^\star)$ with $\delta^\star = \delta + n$, and $\Phi^\star = \Phi + \mathbf{y}\mathbf{y}^T$. Following Roverato (2000), it is possible to sample directly from $p(\Sigma|G, \mathbf{y})$.

Jones et al. (2005) place a prior on G that encourages sparseness by penalizing the number of edges; in particular, a graph with $|\mathcal{E}|$ edges has prior probability $\beta^{|\mathcal{E}|}(1-\beta)^{\binom{p}{2}-|\mathcal{E}|}$ with parameter $\beta = 2/(p-1)$. Now, the marginal likelihood of G simplifies

$$p(\mathbf{y}|G) = \frac{p(\mathbf{y}|G, \Sigma)p(\Sigma|G)}{p(\Sigma|G, y)} = (2\pi)^{-np/2}\frac{h(G, \delta, \Phi)}{h(G, \delta^\star, \Phi^\star)},$$

where $h(G, \delta, \Phi)$ and $h(G, \delta^\star, \Phi^\star)$ are the HIW prior and posterior normalising constants, respectively, and exist in closed form for decomposable graphs G. Hence, to sample from the joint posterior distribution in (5.8), it remains to have an algorithm for simulating from $p(G|\mathbf{y}) \propto p(\mathbf{y}|G)p(G)$.

As in Jones et al. (2005), we use the add-delete MH algorithm; at each iteration, this algorithm chooses with equal probability between adding an edge to, or deleting an edge from the current graph, and the addition or deletion is performed such that the proposal graph is again decomposable. Then the proposal graph is accepted according to the Metropolis-Hastings acceptance probability

$$\alpha(G', G^c) = \min\left\{1, \frac{p(\mathbf{y}|G')p(G')}{p(\mathbf{y}|G^c)p(G^c)} \times \frac{q(G^c|G')}{q(G'|G^c)}\right\},$$

where $G^c = (V, \mathcal{E}^c)$, and $G' = (V, \mathcal{E}')$ are the current and proposal graphs, respectively. If G' is formed by adding an edge to G^c, then $q(G'|G^c) = \left(\binom{p}{2} - |\mathcal{E}^c|\right)^{-1}$; else, $q(G'|G^c) = |\mathcal{E}^c|^{-1}$.

5.3.2 Implementation

We use the C++ code implementing the Metropolis-Hastings algorithm for decomposable Gaussian graphical models available from the authors of Jones et al. (2005) at the website: www.stat.duke.edu/research/software/west/ggm.html. We consider the first simulated example from that paper with an underlying decomposable graph on 15 nodes ($p = 15$) and a data set consisting of 250 observations. The prior on $\Sigma|G$ has parameters $\delta = 3$ and $\Phi = \tau I$ with $\tau = 0.0004$; the prior on G has parameter $\beta = 1/7$. See Jones et al. (2005) for discussion on the choice of τ.

The algorithm is started at the empty graph (no edges present), the full graph (all edges present), and a decomposable non-empty graph (edges $(1, 2)$, $(1, 3)$, $(2, 3)$, $(3, 4)$, $(4, 1)$ present), and run for 2 million iterations. At each iteration, the current graph is represented by a 105-bit binary sequence, where $p(p-1)/2 = 105$ is the maximum number of edges present in the graph; this sequence is hashed to a 64-bit long integer that seeds a random number generator. We then simulate from the following stable distributions and update the corresponding sketch vectors: $X \sim f(x; 0.02, 1, [\cos(\pi \times 0.02/2)]^{1/0.02}, 0)$ for cardinality estimation, $X \sim f(x; 1, 1, \pi/2, 0)$ for entropy estimation, and $X \sim f(x; \alpha, 0, 1, 0)$ for l_α distance estimation with $\alpha = 0.5, 1.0, 1.5$.

In this implementation, we use two random number generators (RNGs), the Mersenne Twister (MT) (Matsumoto and Nishimura, 1998) for the MH algorithm (i.e., for deciding whether to add or delete an edge, and whether to accept the proposal graph), and the Ranq1 combined RNG recommended in Section 7.1.3 of *Numerical Recipes* (Press et al., 2007) for data sketching (i.e., for simulating from the stable distribution). The Mersenne Twister RNG has a very large prime period of $2^{19937} - 1$ so it is well suited for simulating

long sequences of random numbers (in our case, 4 million long); the C++ implementation of the code is taken from www.bedaux.net/mtrand. The Ranq1 RNG takes the output of a 64-bit Xorshift generator and runs it through a multiplicative linear congruential generator (LCG) of the form $I_i = aI_{i-1} \mod 2^{64}$ with $a = 2685821657736338717$. This RNG has period approximately 1.8×10^{19}. In comparison, the authors of Jones et al. (2005) use the built-in generator from the C++ standard library (*srand* and *drand*) for the MH algorithm; according to Press et al. (2007), these generators have no standard implementation and are often badly flawed.

For hashing the binary sequence representation of the graph, we use the hash object function Hashfn2 from Section 7.6.1 of *Numerical Recipes* (Press et al., 2007). This function is designed to work on arbitrary sized inputs by incorporating them into a final hash value a byte at a time, returning a 64-bit unsigned long integer. The ith byte of the input is incorporated by running the current hash value through a multiplicative LCG and adding to the result a random, but fixed, 64-bit value from a lookup table of length 256, corresponding to the ith byte value in $0 \ldots 255$.

5.3.3 Convergence assessment

We present plots of estimators of cardinality, entropy, and l_α distances between chains with different starting graphs. We run the MH-algorithm for 2 million iterations, and obtain the same graph with highest log posterior probability as that reported in Jones et al. (2005); the authors of this paper run the algorithm for 1,050,000 iterations.

For cardinality estimation ($\alpha = 0.02$), we compute the bias-corrected maximum likelihood estimator $\hat{c}_{DS,BC} = (k-1)/k\hat{c}_{DS}$, where k is the length of the data sketch, and $\hat{c}_{DS}$ appears in equation (3.6). This estimator has approximate 95% confidence interval $\hat{c}_{DS,BC} \pm 1.96 \times \hat{c}_{DS,BC}/\sqrt{k-2}$.

For entropy estimation, we compare the trimmed bias-corrected L-estimator $\hat{\delta}_{BC}$ from

equation (5.5) to the equally weighted estimator from Subsection 5.2.3 ($\beta = 1.15$). Note that we are estimating the entropy of the empirical distribution on the graphs visited by the MH-chain, not the entropy of the posterior distribution $p(G|\mathbf{y})$, but as the number of iterations increases to infinity, the former converges to the latter. The approximate 95% confidence interval is $\hat{\delta}_{BC} \pm 1.96\sqrt{\text{MSE}}$, where MSE = 0.06131 for the L-estimator, and MSE = 0.06163 for the equally weighted estimator, obtained by simulations, as shown in Figure 5.3.

For l_α distance approximation, we employ symmetric, strictly stable random projections, and take the location parameter estimation approach, i.e., transforming by taking the logarithm of the absolute value. We compute the trimmed L-estimators $\hat{\gamma}$ of $\gamma = \left(\sum_{i=1}^{n} |p'_i - p''_i|^\alpha \right)^{1/\alpha}$ for $\alpha \geq 1$, and $\hat{\theta}$ of $\theta = \sum_{i=1}^{n} |p'_i - p''_i|^\alpha$ for $\alpha < 1$, where $(p'_1, \ldots, p'_n)$ and $(p''_1, \ldots, p''_n)$ are the empirical distribution functions from two independent MH-chains. The approximate 95% confidence intervals are $\hat{\gamma} \pm 1.96\hat{\gamma}\sqrt{\text{MSE}}/\alpha$, and $\hat{\theta} \pm 1.96\hat{\theta}\sqrt{\text{MSE}}$, respectively, where MSE = 0.02974 for $\alpha = 0.5$, MSE = 0.04115 for $\alpha = 1.0$, and MSE = 0.04921 for $\alpha = 1.5$, obtained by simulations, as shown in Figure 4.13.

Cardinality estimators are obtained from data sketches of length $k = 200$, while entropy and l_α distance estimators are obtained from data sketches of length $k = 50$. Cardinality estimators are plotted on log scale. MT indicates that the Mersenne Twister RNG was used for the MH-algorithm with specified seed. We compare the output chains from different starting graphs, as well as from the same starting graph, but with different seeds for the Mersenne Twister RNG.

For the purpose of comparison, we compute 'exact' values for the cardinality and entropy, which appear in black in the following graphs. We call these values 'exact' because they are computed from the hashed value representation of the graphs in the MH-chain, so due to the low probability of collision of the hash function, these values are very close, but not identical to the true values.

Figure 5.5 compares cardinality and entropy estimators for three MH-chains started from

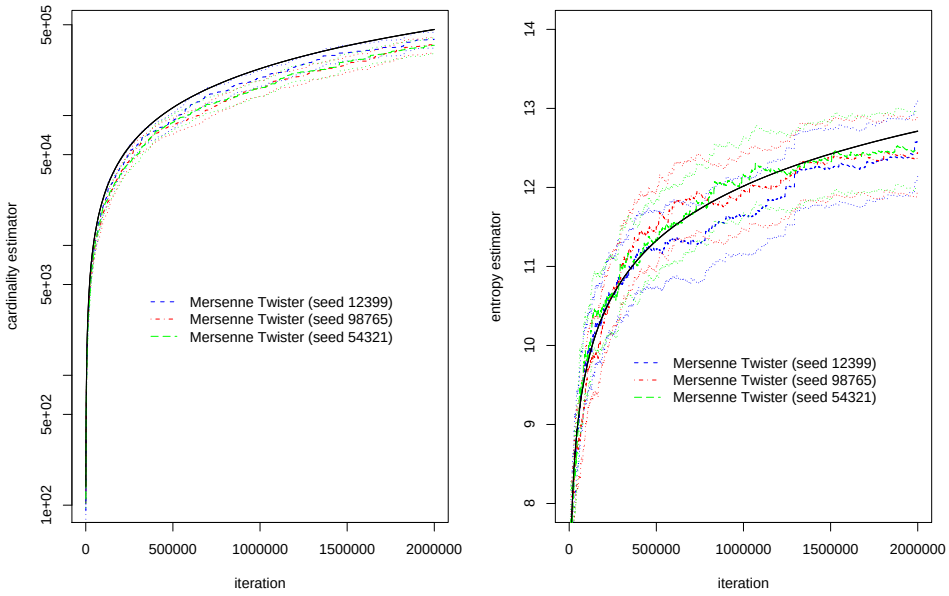


Figure 5.5: Left: Cardinality estimator on log scale for chains started from the empty graph. Right: L-estimator of entropy for chains started from the empty graph. 'Exact' values appear in black. Approximate 95% confidence intervals are drawn with finer line width.

the empty graphs, but with different seeds for the Mersenne Twister RNG. On a log scale, the cardinality estimators are close to the 'exact' value drawn in black. For the initial 500,000 iterations, the estimators increase sharply in value, and, in the case of entropy, the estimators vary across the three chains. After the first million iterations, the latter estimators converge to a common value, which underestimates slightly the 'exact' entropy. Both estimators and 'exact' values continue to increase, but at a significantly slower rate than in the initial part of the run, indicating that the chains are moving freely.

Figure 5.6 displays cardinality estimators on a log scale for three chains with different starting graphs. The behaviour is similar to that observed in Figure 5.5; the 'exact' cardinality values are consistently above the upper bound of the approximate 95% confidence intervals, indicating that the estimators underestimate the 'exact' cardinality. Cardinality estimation is a difficult problem in terms of computing time and storage requirements. Data

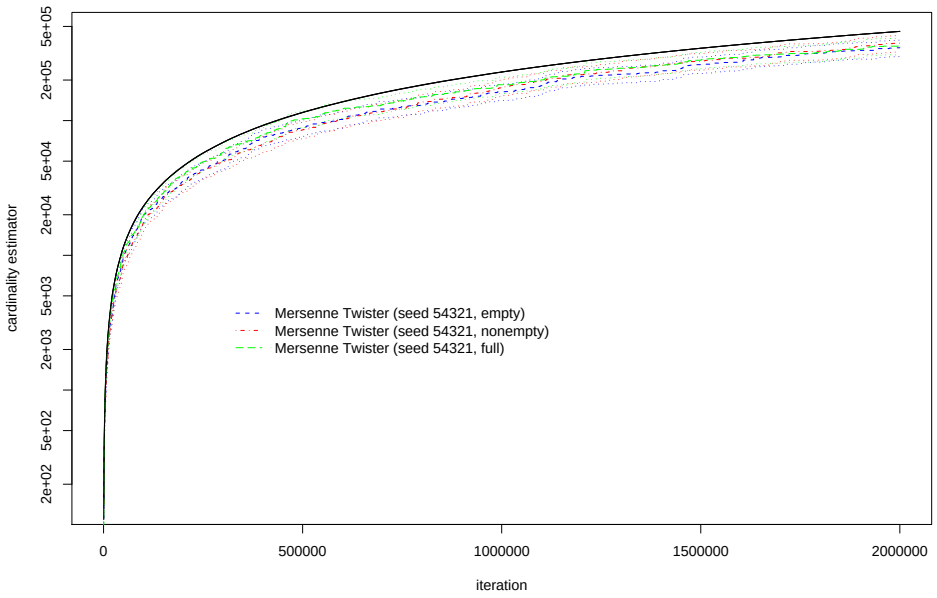


Figure 5.6: Cardinality estimator on log scale for chains with different starting graphs. The 'exact' value appears in black. Approximate 95% confidence intervals are drawn with finer line width.

sketch vectors of length far exceeding 200 are required to estimate accurately cardinalities on the order of 600,000.

Figure 5.7 compares entropy estimators for three chains with different starting graphs: L-estimator versus equally weighted estimator with $\beta = 1.15$. The value of both estimators increases sharply in the first 500,000 iterations, and this behaviour continues for the remainder of the run, but at a significantly slower rate. The 'exact' entropy value falls within the approximate 95% confidence intervals for all chains and estimators; the equally weighted estimator appears to underestimate the 'exact' value more than the L-estimator. Figure 5.8 shows the relative change in entropy as the number of iterations increases. This plot is more informative than the previous one; it shows that after the first million iterations, the increases in entropy estimates are less sharp, decreasing in magnitude, and less frequent.

Figure 5.9 compares l_α distance estimates for pairs of chains with different starting graphs.

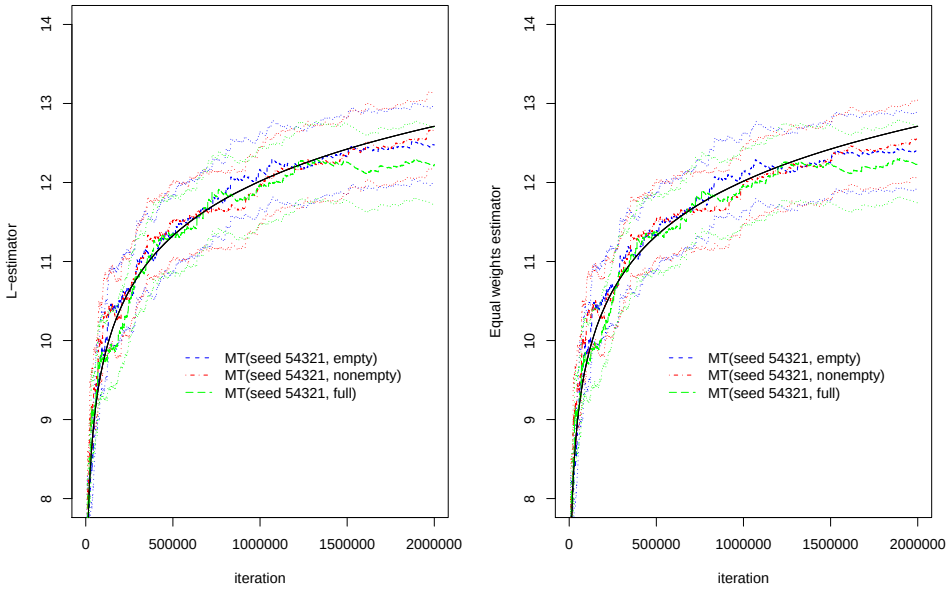


Figure 5.7: Entropy estimators for chains with different starting graphs: L-estimator (left) versus equally weighted estimator with $\beta = 1.15$ (right). The 'exact' value appears in black. Approximate 95% confidence intervals are drawn with finer line width.

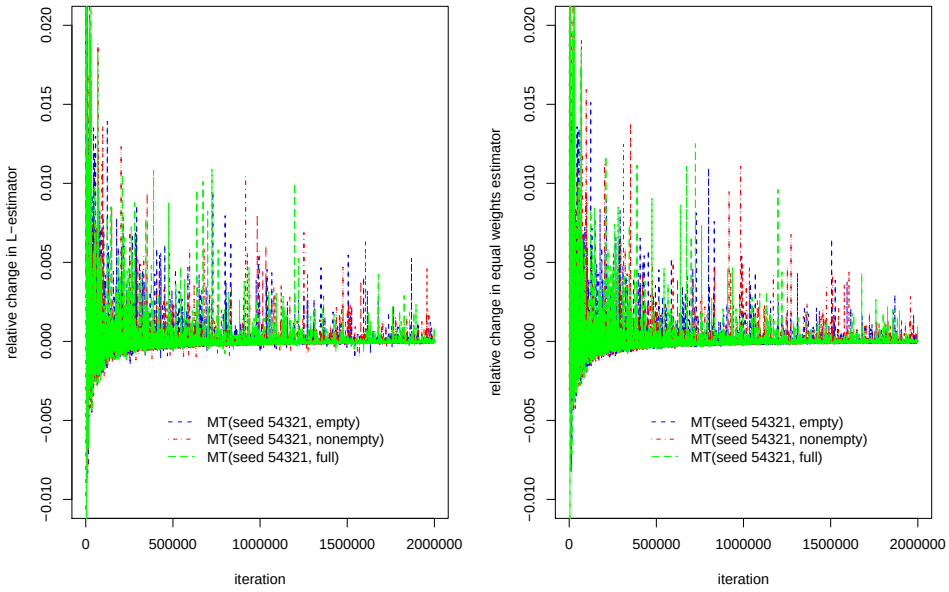


Figure 5.8: Relative change in entropy for estimators in Figure 5.7.

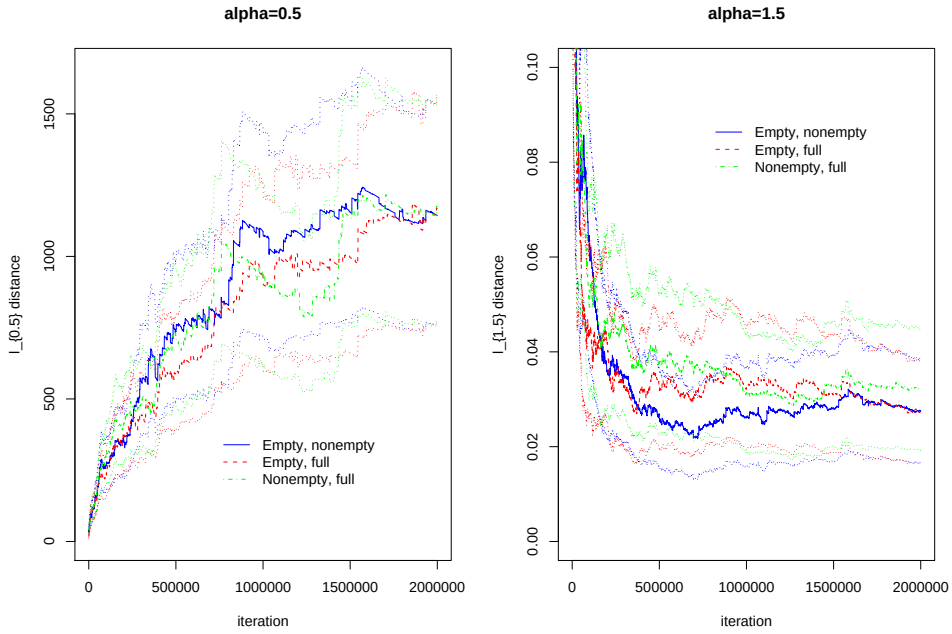


Figure 5.9: l_α distance estimators: $\hat{\theta}$ for $\alpha = 0.5$ (left) and $\hat{\gamma}$ for $\alpha = 1.5$ (right). Approximate 95% confidence intervals are drawn with finer line width. Mersenne Twister RNG is started with seed 54321.

The plot on the left shows great variability in estimates across the three pairs of chains. For α small, $\hat{\theta}$ is close to the Hamming distance; hence, for chains with cardinality on the order of 600,000, we would expect the former to be large and variable, so not very informative as a possible indicator of lack of convergence. However, for $\alpha > 1$, $\gamma = \theta^{1/\alpha}$ is small, and $\hat{\gamma}$ is more informative (Figure 5.9, right). The estimators vary greatly for the first million iterations, but eventually stabilize and converge towards a common value.

5.4 Summary of results

The cardinality estimates increase steadily and consistently indicating that the Markov chain is effective at exploring the extensive sample space. Entropy and distance estimates with stable trend, converging to a common value across different chains or pairs of chains, re-

spectively, indicate that the chains have explored the high density regions of the equilibrium distribution.

In the simulated example on 15-node decomposable graphs drawn from the posterior distribution in a Gaussian graphical model, the Metropolis-Hastings algorithm performs well in the sense that it explores all the high density regions of the sample space, returning a MAP estimate that closely matches the true underlying graph. Moreover, after 1.5 million iterations, entropy and l_α distance estimates from chains with different starting graphs converge to a common value, indicating that all chains have explored the sample space under the posterior distribution.

5.5 Conclusion

In conclusion, for the purpose of MCMC convergence assessment, we recommend observing the behaviour of cardinality, entropy (equally-weighted mean estimator), and l_α distance estimates ($\alpha > 1$) across chains with different starting points as the number of iterations increases. In higher-dimensional problems, for example on graphs with hundreds of nodes, assessing the convergence of a given MCMC method from the output chain is a very difficult task. For these problems, monitoring the behaviour of one-dimensional statistics such as cardinality, entropy, and l_α distances between independent chains with different starting points may be a valuable qualitative tool for convergence assessment. Moreover, computing the value of one-dimensional summary statistics from low-dimensional sketches offers a significant reduction in complexity.

Chapter 6

Conclusion

This thesis has considered the problem of estimating summary statistics of interest such as cardinality, l_α distances and quasi-distances, and entropy, over streaming data with full statistical efficiency from low dimensional data sketches. These sketches are vectors of linear combinations of α-stable random variables, updated on the fly as data elements are observed, processed, and discarded. We have shown that our algorithms have modest computational complexity and storage requirements.

For cardinality estimation we have identified two approaches to constructing data sketches involving indirect record keeping; the first using pseudo-random variates and storing selected order statistics, and the second using random projections. We propose recursively computable, maximum likelihood estimators of the cardinality. We recommend the first approach involving hashing to Geometric$(1 - \rho^{-1})$ random variables ($\rho = 1.1$) and storing maximum order statistics. Algorithms based on this approach attain the tight lower bound on space complexity for cardinality estimation in data streams. The main result of Chapter 3 demonstrates a previously unsuspected connection between these two approaches.

For l_α distance and quasi-distance estimation with $\alpha \in (0, 2]$, we have investigated data sketches based on weighted linear combinations of α-stable random variables, and shown that the problem of recovering these distances can be reduced to that of estimating location and

scale parameters of transformed stable random variables. We have shown how to estimate these parameters by the method of L-estimation, i.e., via weighted linear combinations of order statistics, resulting in asymptotically efficient estimators. For certain values of α, the weight functions display sharp oscillations in small samples, perturbing the performance of the estimators, as shown in plots of mean square error compared to the theoretical lower bound. We have explored trimming and winsorising to improve the small sample performance in such instances, and shown that our estimators outperform existing ones.

For entropy estimation, we propose two estimators based on projections to transformed stable random variables of index $\alpha = 1$. These improve upon the competing estimator of Li (2008e) that approximates the entropy in the limit as $\alpha \to 1$. The first estimator, based on the method of L-estimation, is asymptotically efficient, but has the disadvantage of requiring trimming or winsorising to reduce its infinite variance. We recommend the second estimator since it has finite variance, 97% asymptotic relative efficiency compared to the maximum likelihood estimator, and good small sample performance.

Monitoring the behaviour of summary statistics such as cardinality, entropy, and l_α distances between MCMC output chains with different starting points may be a useful qualitative tool for detecting lack of convergence. We have illustrated this with simulations of the posterior distribution of a decomposable Gaussian graphical model via the Metropolis-Hastings algorithm.

Whereas the focus of this thesis has been on comparing the performance of these estimators in terms of Fisher efficiency, current literature in computer science and engineering compares performance in terms of bounds on tail probabilities, i.e., the probability that the estimator will exceed the true value by a factor of $1 \pm \epsilon$ for fixed sketch size k. If these bounds are decreasing in ϵ, then, by requiring that they not exceed some fixed $\delta > 0$, arbitrarily small, one obtains lower bounds on the sketch size k. This, in turn, results in a lower bound on the space complexity of the algorithm for obtaining an (ϵ, δ)-approximation

to the quantity of interest. We have proved exponentially decreasing tail bounds for the maximum likelihood estimators of the cardinality, but do not have equivalent results for the estimators based on the method of L-estimation. Proving tail bounds for linear combinations of order statistics is complicated, but it might be possible, and useful, to bound the rate of convergence of the distribution of these estimators to normality.

For the future, I am interested in extending these techniques to the near-neighbour problem in high-dimensional spaces (Shakhnarovich et al., 2006). Locality-sensitive hashing, that maps points to low-dimensional sketches, reduces the exponential dependence of the query time on the dimension, known as the curse of dimensionality, to a near-linear dependence with relatively modest storage complexity (Indyk, 2004; Andoni et al., 2006).

Appendix A

Proofs

Proof of Theorem 2.2.2. Let F and f denote the cumulative distribution and density function of X_i, $i = 1, \ldots, n$, and let $\omega_n(\lambda)$ be the Laplace transform of $S_n/X_{(n)}$. Let $F^\star(x) = 1 - F(x)$. For $\lambda > 0$,

$$\begin{aligned}
\omega_n(\lambda) &= \mathbb{E}\big\{e^{-\lambda S_n/X_{(n)}}\big\} \\
&= \mathbb{E}\Big[\mathbb{E}\big\{e^{-\lambda S_n/X_{(n)}}|X_{(n)}\big\}\Big] \\
&= ne^{-\lambda}\bigg\{\int_0^\infty \int_0^y \ldots \int_0^y e^{-\lambda(x_2+\ldots+x_n)/y} F(dx_2)\ldots F(dx_n)F(dy)\bigg\} \\
&= ne^{-\lambda}\int_0^\infty \bigg\{\int_0^y e^{-\lambda x/y}F(dx)\bigg\}^{n-1} F(dy).
\end{aligned}$$

Now, letting $t = x/y$, we have

$$\begin{aligned}
\int_0^y e^{-\lambda x/y}F(dx) &= y\int_0^1 e^{-\lambda t}f(yt)dt \\
&= 1 - F^\star(y) - \lambda\int_0^1 e^{-\lambda t}\big[F^\star(ty) - F^\star(y)\big]dt
\end{aligned}$$

By (2.8), for $t > 0$, $(ty)^\alpha F^\star(ty)[y^\alpha F^\star(y)]^{-1} \to 1$ as $y \to \infty$, so for y large, $F^\star(ty) \approx t^{-\alpha} F^\star(y)$. Therefore, as $y \to \infty$,

$$\begin{aligned}
\lambda \int_0^1 e^{-\lambda t}\big[F^\star(ty) - F^\star(y)\big]dt &\approx \lambda \int_0^1 e^{-\lambda t}\big[t^{-\alpha}F^\star(y) - F^\star(y)\big]dt \\
&= \lambda F^\star(y) \int_0^1 e^{-\lambda t}\big[t^{-\alpha} - 1\big]dt \\
&= \alpha F^\star(y) \int_0^1 \big(1 - e^{-\lambda t}\big)t^{-\alpha-1}dt,
\end{aligned}$$

by integration by parts and by noting that $\alpha \int_0^1 t^{-\alpha-1}dt = \lim_{t\to 0}\big(t^{-\alpha} - 1\big)$. Let $\phi = \alpha \int_0^1 (1 - e^{-\lambda t})t^{-\alpha-1}dt$. Hence, we have that

$$\begin{aligned}
\int_0^y e^{-\lambda x/y}F(dx) &\approx 1 - F^\star(y)(1+\phi) \quad \text{so,} \\
\omega_n(\lambda) &\approx n e^{-\lambda} \int_0^\infty \big[1 - F^\star(y)(1+\phi)\big]^{n-1} F(dy).
\end{aligned}$$

Let $z = nF^\star(y)$, so $F(dy) = -n^{-1}dz$. Then,

$$\begin{aligned}
\omega_n(\lambda) &= e^{-\lambda} \int_0^n \Big[1 - \frac{z}{n}(1+\phi)\Big]^{n-1} dz \\
&\to e^{-\lambda} \int_0^\infty e^{-(1+\phi)z}dz = \frac{e^{-\lambda}}{1+\phi} \quad \text{as } n \to \infty.
\end{aligned}$$

The mean and variance of W then follow by differentiating the Laplace transform. □

Proof of Lemma 2.3.2. By representation in (2.7), we have, for $\alpha < 1$, that

$$|X|^\alpha \stackrel{\mathcal{D}}{=} \frac{|\sin\big(\alpha(U_0+U)\big)|^\alpha}{|\cos(\alpha U_0)\cos U|} \times \left|\frac{\cos\big(\alpha U_0 + (\alpha-1)U\big)}{L}\right|^{1-\alpha},$$

where $L \sim \text{Exp}(1)$, $U \sim \text{Unif}(-\pi/2, \pi/2)$ independently, and $U_0 = \alpha^{-1}\arctan(\beta\tan(\pi\alpha/2))$. As $\alpha \to 0$,

$$\frac{1}{\big|\cos(\alpha U_0)\cos U\big|} \times \left|\frac{\cos\big(\alpha U_0 + (\alpha-1)U\big)}{L}\right|^{1-\alpha} \to \frac{1}{L},$$

since $\alpha U_0 = \arctan\big(\beta\tan(\pi\alpha/2)\big) \to 0$, and $\cos(\alpha U_0) \to 1$ as $\alpha \to 0$. So it remains to show that $|\sin(\alpha(U_0+U))|^\alpha \to 1$ as $\alpha \to 0$. Define $Y = |\sin(\alpha(U_0+U))|$ and show that $\lim_{\alpha\to 0}\log Y^\alpha = 0$ by l'Hôpital's rule.

$$\begin{aligned}\frac{d}{d\alpha}\log Y &= \frac{\cos\big(\arctan(\beta\tan(\pi\alpha/2))+\alpha U\big)}{\sin\big(\arctan(\beta\tan(\pi\alpha/2))+\alpha U\big)}\times\frac{d}{d\alpha}\Big[\arctan(\beta\tan(\pi\alpha/2))+\alpha U\Big]\\ &= \frac{\beta\tan(\pi\alpha/2)+\tan(\alpha U)}{1-\beta\tan(\pi\alpha/2)\tan(\alpha U)}\times\left[\frac{\pi/2\beta\sec^2(\pi\alpha/2)}{1+\beta^2\tan^2(\pi\alpha/2)}+U\right]\end{aligned}$$

tends to 0 as $\alpha\to 0$, and the result follows. □

Proof of Theorem 2.3.1. Let $L\sim\text{Exp}(1)$, $U\sim\text{Unif}(0,\pi)$, $L\perp\!\!\!\perp U$. Using the result in Kanter (1975), we have for fixed $y>0$

$$\begin{aligned}\mathbb{P}\left(\frac{\big[X_{(n)}\big]^{\alpha_n}}{n}\le y\right) &= \left[\mathbb{P}\Big(a(U)\times(ny)^{(\alpha_n-1)^{-1}}\le L\Big)\right]^n\\ &= \left\{\int_0^\pi\frac{1}{\pi}\int_{a(x)(ny)^{(\alpha_n-1)^{-1}}}^\infty e^{-l}dldx\right\}^n\\ &= \left\{\frac{1}{\pi}\int_0^\pi\exp\Big\{-a(x)\times(ny)^{(\alpha_n-1)^{-1}}\Big\}dx\right\}^n \qquad (A.1)\end{aligned}$$

For $x\in(0,\pi)$ and fixed n, define

$$\begin{aligned}f_n(x,y) &:= \exp\Big\{-a(x)\times(ny)^{(\alpha_n-1)^{-1}}\Big\}\\ &= \exp\Big\{-(ny)^{(\alpha_n-1)^{-1}}\big[\sin(\alpha_n x)\big]^{\alpha_n/(1-\alpha_n)}\sin\big((1-\alpha_n)x\big)\big[\sin(x)\big]^{-(1-\alpha_n)^{-1}}\Big\}.\end{aligned}$$

Since $\lim_{x\to\pi^-}a(x)=\infty$, and $\lim_{x\to 0^+}a(x)=\alpha_n^{\alpha_n/(1-\alpha_n)}(1-\alpha_n)$, we define

$$f_n(0,y):=\exp\Big\{-\alpha_n^{\alpha_n/(1-\alpha_n)}\times(1-\alpha_n)\times(ny)^{(\alpha_n-1)^{-1}}\Big\}\text{ and }f_n(\pi,y):=0.$$

Hence $f_n(x,y)$, with $y>0$ and n fixed, is continuous in the interval $[0,\pi]$ and differentiable in $(0,\pi)$. By the Mean Value Theorem for Integration applied to (A.1), $\exists x_n\in(0,\pi)$ such that

$$\int_0^\pi\exp\big\{-a(x)\times(ny)^{(\alpha_n-1)^{-1}}\big\}dx=\pi\exp\big\{-a(x_n)\times(ny)^{(\alpha_n-1)^{-1}}\big\}=\pi f_n(x_n,y).$$

Therefore

$$\begin{aligned}\mathbb{P}\big(n^{-1}[X_{(n)}]^{\alpha_n}\le y\big) &= \exp\Big\{-n(ny)^{(\alpha_n-1)^{-1}}\times\big[\sin(\alpha_n x_n)\big]^{\alpha_n/(1-\alpha_n)}\times\sin\big((1-\alpha_n)x_n\big)\\ &\quad\times\big[\sin(x_n)\big]^{-(1-\alpha_n)^{-1}}\Big\}\\ &= \big[f_n(x_n,y)\big]^n. \qquad (A.2)\end{aligned}$$

The goal is to show that expression (A.2) converges to $e^{-1/y}$ as $n \to \infty$ provided that $\alpha_n \log(n) \to 0$ as $n \to \infty$.

Claim 1. *The function $a(x)$ is increasing on the interval $(0, \pi)$ for any $\alpha_n \in (0, 1)$ fixed.*

Proof. For fixed $p \in (0, 1)$, define the function $s_p(x) = \sin(px)/\sin(x)$. It can easily be shown that $s_p(x)$ is increasing on the interval $(0, \pi)$. Now, $a(x) = \big[s_{\alpha_n}(x)\big]^{\alpha_n/(1-\alpha_n)} \times s_{1-\alpha_n}(x)$, is the product of two non-negative, increasing functions on $(0, \pi)$ for fixed $0 < \alpha_n < 1$; hence $a(x)$ is increasing on $(0, \pi)$. □

If follows from Claim 1 that for fixed $y > 0$, the function $f_n(x, y)$ is decreasing on the interval $(0, \pi)$. So, $\forall x \in (0, \pi)$,

$$\big[f_n(x, y)\big]^n \leq \big[f_n(0, y)\big]^n = \exp\Big\{ - n\alpha_n^{\alpha_n/(1-\alpha_n)} \times (1 - \alpha_n) \times (ny)^{(\alpha_n - 1)^{-1}} \Big\}. \tag{A.3}$$

Taking the logarithm on both sides of (A.3) and letting $n \to \infty$, we obtain that

$$n \log\big(f_n(x, y)\big) \leq -\alpha_n^{\alpha_n/(1-\alpha_n)} \times (1 - \alpha_n) \times n^{\alpha_n/(\alpha_n - 1)} \times y^{(\alpha_n - 1)^{-1}} \to -\frac{1}{y},$$

provided that $\alpha_n \log(n) \to 0$ as $n \to \infty$. From (A.2) and (A.3), it follows that as $n \to \infty$

$$\mathbb{P}\Big(n^{-1}\big[X_{(n)}\big]^{\alpha_n} \leq y\Big) \leq \exp\Big\{ - \alpha_n^{\alpha_n/(1-\alpha_n)} \times (1 - \alpha_n) \times n^{\alpha_n/(\alpha_n - 1)} \times y^{(\alpha_n - 1)^{-1}} \Big\} \to e^{-1/y}.$$

Let $n \geq \lceil 1/y \rceil$. Since, $ny > 1$ and $a(x) > 0$, we have that

$$f_n(x, y) = \exp\Big\{ - a(x) \times (ny)^{(\alpha_n - 1)^{-1}} \Big\} \geq \exp\Big\{ - \frac{a(x)}{ny} \Big\}, \quad \forall x \in (0, \pi).$$

Define the sequence $\big\{\epsilon_n = \pi\sqrt{\alpha_n}/2\big\}$ in the range $(0, \pi/2)$. As $n \to \infty$, $\epsilon_n \to 0$, $\alpha_n/\epsilon_n = 2\sqrt{\alpha_n}/\pi \to 0$ and $\alpha_n \log(\epsilon_n/\alpha_n) = \alpha_n \log(\pi/2) - 1/2\alpha_n \log(\alpha_n) \to 0$. By Claim 1, $a(x) \leq a(\pi - \epsilon_n)$ for $x \in (0, \pi - \epsilon_n]$, so

$$\int_0^{\pi - \epsilon_n} f_n(x, y) dx \geq \int_0^{\pi - \epsilon_n} \exp\Big\{ - \frac{a(\pi - \epsilon_n)}{ny} \Big\} dx = (\pi - \epsilon_n) \exp\Big\{ - \frac{a(\pi - \epsilon_n)}{ny} \Big\}. \tag{A.4}$$

Now consider values of x in the interval $(\pi - \epsilon_n, \pi - \alpha_n/n)$; since $\alpha_n/\epsilon_n \to 0$ as $n \to \infty$, then for n sufficiently large, $\epsilon_n \geq \alpha_n$ and $\pi - \epsilon_n \leq \pi - \alpha_n/n$.

Using the facts that $x\cos x \le \sin x \le x\ \forall x \in (0,\pi/2)$ and $\sin(x) = \sin(\pi - x)\ \forall x$, we find an upper bound for $a(x)$ on $(\pi - \epsilon_n, \pi - \alpha_n/n)$ as follows.

$$\begin{aligned} a(x) &\le \frac{(\alpha_n x)^{\alpha_n/(1-\alpha_n)}\sin(\pi-(1-\alpha_n)x)}{\big[\sin(\pi-x)\big]^{(1-\alpha_n)^{-1}}} \\ &\le \frac{(\alpha_n x)^{\alpha_n/(1-\alpha_n)}[\pi-(1-\alpha_n)x]}{(\pi-x)^{(1-\alpha_n)^{-1}}\big[\cos(\pi-x)\big]^{(1-\alpha_n)^{-1}}} \\ &\le \frac{(\alpha_n \pi)^{\alpha_n/(1-\alpha_n)}[\pi-x+\alpha_n\pi]}{(\pi-x)^{(1-\alpha_n)^{-1}}\big[\cos(\epsilon_n)\big]^{(1-\alpha_n)^{-1}}}, \end{aligned}$$

since $\pi - x \le \epsilon_n < \pi/2$ and $\cos(x)$ is strictly decreasing on $(0,\pi/2)$. Moreover, we note that if $\pi - x > \alpha_n/n$, then $(\pi-x)^{(1-\alpha_n)^{-1}} > (\pi-x)(\alpha_n/n)^{\alpha_n/(1-\alpha_n)}$, and

$$\begin{aligned} a(x) &\le \frac{(\alpha_n\pi)^{\alpha_n/(1-\alpha_n)}[\pi-x+\alpha_n\pi]}{(\pi-x)(\alpha_n/n)^{\alpha_n/(1-\alpha_n)}\big[\cos(\epsilon_n)\big]^{(1-\alpha_n)^{-1}}} \\ &= \frac{(\pi n)^{\alpha_n/(1-\alpha_n)}[\pi-x+\alpha_n\pi]}{\big[\cos(\epsilon_n)\big]^{(1-\alpha_n)^{-1}}(\pi-x)} \\ &= b_n\frac{\pi-x+\alpha_n\pi}{\pi-x}, \end{aligned}$$

where $\Big\{b_n = (\pi n)^{\alpha_n/(1-\alpha_n)}[\cos(\epsilon_n)\big]^{-(1-\alpha_n)^{-1}}\Big\}$ is a sequence converging to 1 as $n\to\infty$ and $\alpha_n\log(n)\to 0$. Let $c_n = b_n/(ny)$. We have

$$\begin{aligned} \int_{\pi-\epsilon_n}^{\pi-\alpha_n/n} f_n(x,y)dx &\ge \int_{\pi-\epsilon_n}^{\pi-\alpha_n/n}\exp\Big\{-\frac{c_n(\pi-x+\alpha_n\pi)}{\pi-x}\Big\}dx \\ &= \exp\{-c_n\}\int_{\alpha_n/n}^{\epsilon_n}\exp\big\{-c_n\alpha_n\pi/u\big\}du \\ &\ge \exp\{-c_n\}\int_{\alpha_n/n}^{\epsilon_n}\big(1-c_n\alpha_n\pi/u\big)du \\ &= \exp\{-c_n\}[\epsilon_n-\delta_n], \end{aligned} \tag{A.5}$$

where $n\delta_n = \alpha_n\big[1+\pi b_n y^{-1}\log(n\epsilon_n/\alpha_n)\big] \to 0$ as $n\to\infty$. Using inequalities (A.4) and (A.5) and the fact that $ny > 1$, we obtain the following lower bound

$$\begin{aligned} \int_0^\pi f_n(x,y)dx &\ge (\pi-\epsilon_n)\exp\Big\{-\frac{a(\pi-\epsilon_n)}{ny}\Big\} + (\epsilon_n-\delta_n)\exp\Big\{-\frac{b_n}{ny}\Big\} \\ &\ge \exp\Big\{-\frac{1}{ny}\Big\}\Big[(\pi-\epsilon_n)\exp\Big\{-\frac{a(\pi-\epsilon_n)}{ny}\Big\} + (\epsilon_n-\delta_n)\exp\Big\{-\frac{b_n}{ny}\Big\}\Big], \end{aligned}$$

Let $\delta_n' = (\pi - \epsilon_n)\exp\big\{-a(\pi-\epsilon_n)/(ny)\big\} + (\epsilon_n - \delta_n)\exp\big\{-b_n/(ny)\big\} - \pi$. We now show that $n \mid \delta_n' \mid \to 0$ as $n \to \infty$.

Claim 2. *If $a_n \to 1$ as $n \to \infty$, then $n\Big[\exp\big\{-a_n/(ny)\big\} - \exp\big\{-1/(ny)\big\}\Big] \to 0$.*

Proof.

$$\begin{aligned} n\Big[\exp\big\{-a_n/(ny)\big\} - \exp\big\{-1/(ny)\big\}\Big] &= n\exp\big\{-1/(ny)\big\}\Big[\exp\big\{(1-a_n)/(ny)\big\} - 1\Big] \\ &= \exp\big\{-1/(ny)\big\} \times \left[\sum_{i=1}^{\infty} \frac{(-1)^i(1-a_n)^i}{n^{i-1}y^i(i!)}\right]. \end{aligned}$$

We want to show that the limit of the infinite sum as $n \to \infty$ is 0. We have that

$$\lim_{n\to\infty}\sum_{i=1}^{\infty}\frac{(-1)^i(1-a_n)^i}{n^{i-1}y^i(i!)} = \lim_{n\to\infty}\lim_{S\to\infty}\sum_{i=1}^{S}\frac{(-1)^i(1-a_n)^i}{n^{i-1}y^i(i!)} = \lim_{S\to\infty}\lim_{n\to\infty}\sum_{i=1}^{S}\frac{(-1)^i(1-a_n)^i}{n^{i-1}y^i(i!)}.$$

Let $\epsilon > 0$ be arbitrary. Then $\exists N \in \mathbb{N}$ such that $n \geq N$ implies that $\mid 1 - a_n \mid \leq \epsilon$, so

$$\left|\sum_{i=1}^{S}\frac{(-1)^i(1-a_n)^i}{n^{i-1}y^i(i!)}\right| \leq \sum_{i=1}^{S}\frac{\mid 1-a_n\mid^i}{n^{i-1}y^i(i!)} \leq \sum_{i=1}^{S}\frac{\epsilon^i}{n^{i-1}y^i(i!)} \to 0 \text{ as } n \to \infty,$$

and the claim is proved. □

Claim 3. *$a(\pi - \epsilon_n) \to 1$ as $n \to \infty$.*

Proof.

$$\begin{aligned} a(\pi-\epsilon_n) &= \Big[\sin\big(\alpha_n(\pi-\epsilon_n)\big)\Big]^{\alpha_n/(1-\alpha_n)} \sin\big((1-\alpha_n)(\pi-\epsilon_n)\big)\big[\sin(\pi-\epsilon_n)\big]^{-(1-\alpha_n)^{-1}} \\ &= \cos\big(\alpha_n(\pi-\epsilon_n)\big)\left[\frac{\sin\big(\alpha_n(\pi-\epsilon_n)\big)}{\sin(\pi-\epsilon_n)}\right]^{\alpha_n/(1-\alpha_n)} \\ &\quad -\cos(\pi-\epsilon_n)\left[\frac{\sin\big(\alpha_n(\pi-\epsilon_n)\big)}{\sin(\pi-\epsilon_n)}\right]^{(1-\alpha_n)^{-1}}. \end{aligned}$$

As $n \to \infty$,

$$\left[\frac{\sin\big(\alpha_n(\pi-\epsilon_n)\big)}{\sin(\pi-\epsilon_n)}\right]^{\alpha_n/(1-\alpha_n)} = \left\{\frac{\sin\big(\alpha_n(\pi-\epsilon_n)\big)}{\alpha_n(\pi-\epsilon_n)} \times \frac{\alpha_n}{\epsilon_n} \times (\pi-\epsilon_n) \times \frac{\epsilon_n}{\sin(\epsilon_n)}\right\}^{\alpha_n/(1-\alpha_n)} \to 1,$$

since $\alpha_n/(1-\alpha_n)\log(\alpha_n/\epsilon_n) \to 0$. Similarly, $\big\{\sin(\alpha_n(\pi-\epsilon_n))/\sin(\pi-\epsilon_n)\big\}^{(1-\alpha_n)^{-1}} \to 0$ as $n \to \infty$. Finally, since $\cos\big(\alpha_n(\pi-\epsilon_n)\big) \to 1$ and $\cos(\pi-\epsilon_n) \to -1$ as $n \to \infty$, the claim follows. □

By Claims 2 and 3 and the fact that $n\delta_n \to 0$ as $n \to \infty$, we have that

$$\begin{aligned} n \mid \delta_n' \mid \;\leq\; & (\pi - \epsilon_n) n \Big| \exp\big\{ -a(\pi - \epsilon_n)/(ny) \big\} - \exp\big\{ -1/(ny) \big\} \Big| + \\ & n|\epsilon_n - \delta_n| \times \Big| \exp\big\{ -b_n/(ny) \big\} - \exp\big\{ -1/(ny) \big\} \Big| + \\ & \pi n \Big| \exp\big\{ -1/(ny) \big\} - 1 \Big| + n\delta_n \exp\big\{ -1/(ny) \big\} \to 0, \end{aligned}$$

so, as $n \to \infty$,

$$\mathbb{P}\left(n^{-1} \big[X_{(n)} \big]^{\alpha_n} \leq y \right) \geq \Big[\pi^{-1}(\pi + \delta_n') \exp\big\{ -1/(ny) \big\} \Big]^n \to e^{-1/y}.$$

Therefore, as $n \to \infty$,

$$\mathbb{P}\left(n^{-1} \big[X_{(n)} \big]^{\alpha_n} \leq y \right) \to e^{-1/y},$$

and we have shown that $n^{-1} \big[X_{(n)} \big]^{\alpha_n} \xrightarrow{\mathcal{D}} 1/L$, provided that $\alpha_n \log(n) \to 0$ as $n \to \infty$. Finally, by Lemmma 2.3.1,

$$\left[\frac{X_{(n)}}{S_n} \right]^{\alpha_n} \stackrel{\mathcal{D}}{=} \frac{\big[X_{(n)} \big]^{\alpha_n}}{n X^{\alpha_n}} \xrightarrow{\mathcal{P}} 1,$$

where $X \sim S(x; \alpha_n)$. □

Bibliography

Achlioptas, D. (2003) Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, **66**, 671–687.

Aggarwal, C. C. (2007) *Data streams: Models and Algorithms.* New York: Springer-Verlag.

Akella, A., Bharambe, A., Reiter, M. and Seshan, S. (2003) Detecting DDoS attacks on ISP networks. In *Proceedings of the Workshop on Management and Processing of Data Streams.*

Alon, N., Matias, Y. and Szegedy, M. (1999) The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, **58**, 137–147.

Andoni, A., Datar, M., Immorlica, N., Indyk, P. and Mirrokni, V. S. (2006) Locality-sensitive hashing using stable distributions. In *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice* (eds. G. Shakhnarovich, T. Darrell and P. Indyk), 55–66. Cambridge, MA: The MIT Press.

Arriaga, R. I. and Vempala, S. (2006) An algorithmic theory of learning: robust concepts and random projections. *Machine Learning*, **63**, 161–182.

Babcock, B., Babu, S., Datar, M., Motwani, R. and Widom, J. (2002) Models and issues in data stream systems. In *PODS '02: Proceedings of the twenty-first ACM SIGMOD-*

SIGACT-SIGART symposium on Principles of database systems, 1–16. New York, NY: ACM.

Bar Youssef, Z., Jayram, T. S., Kumar, R., Sivakumar, D. and Trevisan, L. (2002) Counting distinct elements in a data stream. *Lecture Notes in Computer Science*, **2483**, 1–10.

Besbeas, P. and Morgan, B. J. T. (2001) Integrated squared error estimation of Cauchy parameters. *Statistics & Probability Letters*, **55**, 397–401.

— (2004) Efficient and robust estimation for the one-sided stable distribution of index $\frac{1}{2}$. *Statistics & Probability Letters*, **66**, 251–257.

— (2008) Improved estimation of the stable laws. *Statistics and Computing*, **18**, 219–231.

Bhuvanagiri, L. and Ganguly, S. (2006) Estimating entropy over data streams. *Lecture Notes in Computer Science*, **4168**, 148.

Breiman, L. (1992) *Probability.* Reading, Massachusetts: Addison-Wesley Pub. Co., 2 edn.

Brinkman, B. and Charikar, M. (2005) On the impossibility of dimension reduction in l_1. *Journal of the Association for Computing Machinery (ACM)*, **52**, 766–788.

Bunge, J. and Fitzpatrick, M. (1993) Estimating the number of species: A review. *Journal of the American Statistical Association*, **88**, 364–373.

Chakrabarti, A., Do Ba, K. and Muthukrishnan, S. (2006) Estimating entropy and entropy norm on data streams. *Internet Mathematics*, **3**, 63–78.

Chambers, J. M., Mallows, C. L. and Stuck, B. W. (1976) A method for simulating stable random variables. *Journal of the American Statistical Association*, **71**, 340–344.

Cheng, R. C. H. and Liu, W. B. (1997) A continuous representation of the family of stable law distributions. *Journal of the Royal Statistical Society Series B*, **59**, 137–145.

Chernoff, H. (1952) A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, **23**, 493–507.

Chernoff, H., Gastwirth, J. L. and Johns, Jr., M. V. (1967) Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation. *Annals of Mathematical Statistics*, **38**, 52–72.

Cormode, G., Datar, M., Indyk, P. and Muthukrishnan, S. (2003) Comparing data streams using Hamming norms (how to zero in). *IEEE Transactions on Knowledge and Data Engineering*, **15**, 529–540.

Cormode, G. and Muthukrishnan, S. (2005a) An improved data stream summary: the Count-Min sketch and its applications. *Journal of Algorithms*, **55**, 58–75.

— (2005b) What's new: Finding significant differences in network data streams. *IEEE/ACM Transactions on Networking (TON)*, **13**, 1219–1232.

Cox, T. F. and Cox, M. A. A. (2001) *Multidimensional scaling*. Boca Raton: Chapman & Hall/CRC, 2 edn.

Cramér, H. (1946) *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

Cressie, N. (1975) A note on the behaviour of the stable distributions for small index α. *Probability Theory and Related Fields*, **33**, 61–64.

Darling, D. A. (1952) The influence of the maximum term in the addition of independent random variables. *Transactions of the American Mathematical Society*, **73**, 95–107.

Dasgupta, S. and Gupta, A. (2003) An elementary proof of the Johnson-Lindenstrauss lemma. *Random Structures and Algorithms*, **22**, 60–65.

David, H. A. and Nagaraja, H. N. (2003) *Order statistics*. New York: Wiley-Interscience, 3 edn.

Diaconis, P. and Freedman, D. (1984) Asymptotics of graphical projection pursuit. *The Annals of Statistics*, **12**, 793–815.

Dixon, W. J. (1960) Simplified estimation from censored normal samples. *The Annals of Mathematical Statistics*, **31**, 385–391.

DuMouchel, W. H. (1973) On the asymptotic normality of the maximum likelihood estimate when sampling from a stable distribution. *The Annals of Statistics*, **1**, 948–957.

— (1975) Stable distributions in statistical inference: 2. Information from stably distributed samples. *Journal of the American Statistical Association*, **70**, 386–393.

Durand, M. and Flajolet, P. (2003) Loglog counting of large cardinalities. *Lecture Notes in Computer Science*, **2832**, 605–617.

Fama, E. F. and Roll, R. (1968) Some properties of symmetric stable distributions. *Journal of the American Statistical Association*, **63**, 817–836.

— (1971) Parameter estimates for symmetric stable distributions. *Journal of the American Statistical Association*, **66**, 331–338.

Feller, W. (1971) *An Introduction to Probability Theory and Its Applications*, vol. 2. New York: John Wiley & Sons, Inc., 2 edn.

Feuerverger, A. and McDunnough, P. (1981a) On some Fourier methods for inference. *Journal of the American Statistical Association*, **76**, 379–387.

— (1981b) On the efficiency of empirical characteristic function procedures. *Journal of the Royal Statistical Society Series B*, **43**, 20–27.

Fisher, R. A. (1925) Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, **22**, 700–725.

Flajolet, P. (1990) On adaptive sampling. *Computing*, **34**, 391–400.

— (2004) Counting by coin tossings. *Lecture Notes in Computer Science*, **3321**, 1–12.

Flajolet, P. and Martin, G. N. (1985) Probabilistic counting algorithms for database applications. *Journal of Computer and System Sciences*, **31**, 182–209.

Frankl, P. and Maehara, H. (1988) The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *Journal of Combinatorial Theory*, **44**, 355–362.

Ganguly, S. (2004) Estimating frequency moments of data streams using random linear combinations. *Lecture notes in computer science*, 369–380.

— (2007) Counting distinct items over update streams. *Theoretical Computer Science*, **378**, 211–222.

Giroire, F. (2005) Order statistics and estimating cardinalities of massive data sets. In *2005 International Conference on Analysis of Algorithms*, DMTCS Proceedings, 157–166. Discrete Mathematics and Theoretical Computer Science.

Gnedenko, B. V. and Kolmogorov, A. N. (1954) *Limit Distributions for Sums of Independent Random Variables.* Cambridge, Massachusetts: Addison-Wesley Pub. Co.

Gradshteyn, I. S. and Ryzhik, I. M. (1980) *Table of Integrals, Series and Products.* London: Academic Press, Inc.

Haas, P. J., Naughton, J. F., Seshadri, S. and Stokes, L. (1995) Sampling-based estimation of the number of distinct values of an attribute. In *Proceedings of the International Conference on Very Large Data Bases*, 311–322.

Hammersley, J. M. (1950) On estimating restricted parameters. *Journal of the Royal Statistical Society Series B*, **12**, 192–240.

Harvey, N., Nelson, J. and Onak, K. (2008) Streaming algorithms for estimating entropy. In *IEEE Information Theory Workshop, 2008. ITW'08*, 227–231.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Haveliwala, T. H., Gionis, A. and Indyk, P. (2000) Scalable techniques for clustering the Web. In *WebDB (Informal Proceedings)*, 129–134.

Huber, P. J. (1985) Projection pursuit. *The Annals of Statistics*, **13**, 435–475.

— (2004) *Robust Statistics*. New York, Chichester: Wiley-Interscience.

Ibragimov, I. A. and Linnik, Y. V. (1971) *Independent and stationary sequences of random variables*. Groningen: Wolters-Noordhoff.

Indyk, P. (2000) Stable distributions, pseudorandom generators, embeddings and data stream computation. In *Proc. 41st IEEE Symposium on Foundations of Computer Science (FOCS)*, 189–197.

— (2004) Nearest neighbors in high-dimensional spaces. In *Handbook of Discrete and Computational Geometry* (eds. J. E. Goodman and J. O'Rourke), 877–892. Boca Raton, Florida: Chapman & Hall/CRC, 2 edn.

— (2006) Stable distributions, pseudorandom generators, embeddings, and data stream computation. *Journal of the Association for Computing Machinery (ACM)*, **53**, 307–323.

Indyk, P., Koudas, N. and Muthukrishnan, S. (2000) Indentifying representative trends in massive time series data sets using sketches. In *Proceedings of the 26th International Conference on Very Large Data Bases*, 363–372.

Indyk, P. and Woodruff, D. (2003) Tight lower bounds for the distinct elements problem. In *Annual Symposium on Foundations of Computer Science*, vol. 44, 283–289.

— (2005) Optimal approximations of the frequency moments of data streams. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, 202–208.

Johnson, W. B. and Lindenstrauss, J. (1984) Extensions of Lipschitz mapping into Hilbert space. *Contemporary Mathematics*, **26**, 189–206.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C. and West, M. (2005) Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, **20**, 388–400.

Kanter, M. (1975) Stable densities under change of scale and total variation inequalities. *The Annals of Probability*, **3**, 697–707.

Kirschenhofer, P. and Prodinger, H. (1993) A result in order statistics related to probabilistic counting. *Computing*, **51**, 15–27.

Knuth, D. E. (1998) *The Art of Computer Programming: Sorting and Searching*, vol. 3. Massachusetts: Addison-Wesley, 2 edn.

Kogon, S. M. and Williams, D. B. (1998) Characteristic function based estimation of stable distribution parameters. In *A practical guide to heavy tails: statistical techniques and applications* (eds. R. J. Adler, R. E. Feldman and M. S. Taqqu), 311–335. Cambridge, MA: Birkhauser Boston Inc.

Kolokoltsov, V., Korolev, V. and Uchaikin, V. (2001) Fractional stable distributions. *Journal of Mathematical Sciences*, **105**, 2569–2576.

Koutrouvelis, I. A. (1980) Regression-type estimation of the parameters of the stable laws. *Journal of the American Statistical Association*, **75**, 918–928.

— (1981) An iterative procedure for the estimation of the parameters of the stable laws. *Communications in Statistics: Simulation and Computation*, **10**, 17–28.

Lall, A., Sekar, V., Ogihara, M., Xu, J. and Zhang, H. (2006) Data streaming algorithms for estimating entropy of network traffic. In *Proceedings of the joint international conference on measurement and modeling of computer systems*, 145–156. ACM New York, NY, USA.

Lauritzen, S. (1996) *Graphical models.* Oxford University Press, USA.

Lauritzen, S. L. (1988) Extremal families and systems of sufficient statistics. In *Lecture Notes in Statistics*, vol. 49. Berlin: Springer-Verlag.

Lehmann, E. L. (1983) *Theory of Point Estimation.* New York: John Wiley & Sons, Inc.

Lehmann, E. L. and Scheffé, H. (1950) Completeness, similar regions and unbiased estimation. Part I. *Sankhyā*, **10**, 305–340.

Lévy, P. (1924) Théorie des erreurs. La loi de Gauss et les lois exceptionelles. *Bulletin de la Société Mathématique de France*, **52**, 49–85.

Li, P. (2008a) Computationally efficient estimators for dimension reduction using stable random projections. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, 403–412.

— (2008b) Estimators and tail bounds for dimension reduction in l_α $(0 < \alpha \leq 2)$ using stable random variables. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms (SODA)*, 10–19.

— (2008c) On approximating frequency moments of data streams with skewed projections. In *Preprint, available on arXiv:0802.0802v1*.

— (2008d) The optimal quantile estimator for compressed counting. In *Preprint, available on arXiv:0808.1766v1*.

— (2008e) A very efficient scheme for estimating entropy of data streams using compressed counting. In *Preprint, available on arXiv:0808.1771v2*.

— (2009) Compressed counting. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 412–421.

Li, P. and Hastie, T. J. (2008) A unified near-optimal estimator for dimension reduction in $l_\alpha (0 < \alpha \leq 2)$ using stable random variables. In *Advances in Neural Information Processing Systems 20* (eds. J. C. Platt, D. Koller, Y. Singer and S. Roweis). Cambridge, MA: MIT Press.

Li, P., Hastie, T. J. and Church, K. W. (2006a) Improving random projections using marginal information. *Lecture Notes in Computer Science*, 635–649.

— (2006b) Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 287–296.

— (2007) Nonlinear estimators and tail bounds for dimension reduction in l_1 using Cauchy random projections. *Lecture Notes in Computer Science*, 514–529.

Lindley, D. V. (1956) On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, **27**, 986–1005.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1997) *Multivariate Analysis*. London: Academic Press, 6 edn.

Matsui, M. and Takemura, A. (2006) Some improvements in numerical evaluation of symmetric stable density and its derivatives. *Communications in Statistics: Theory and Methods*, **35**, 149–172.

Matsumoto, M. and Nishimura, T. (1998) Mersenne Twister: a 623-dimensionally equidistributed uniform pseudo random number generator. *ACM Transactions on Modeling and Computer Simulation*, **8**, 3–30.

McCulloch, J. H. (1986) Simple consistent estimators of stable distribution parameters. *Communications in Statistics: Simulation and Computation*, **15**, 1109–1136.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, **21**, 1087–1091.

Muthukrishnan, S. (2005) *Data Streams: Algorithms and Applications*. Cambridge, Massachusetts: Now Publishers Inc, 1 edn.

Nikias, C. L. and Shao, M. (1995) *Signal Processing with Alpha-Stable Distributions and Applications*. New York: Wiley-Interscience, 1 edn.

Nolan, J. P. (1997) Numerical calculation of stable densities and distribution. *Communications in Statistics and Stochastic Models*, **13**, 759–774.

— (1998) Univariate stable distributions: Parameterizations and software. In *A practical guide to heavy tails: statistical techniques and applications* (eds. R. J. Adler, R. E. Feldman and M. S. Taqqu), 527–533. Cambridge, MA: Birkhauser Boston Inc.

— (1999) An algorithm for evaluating stable densities in Zolotarev's (M) parameterization. *Mathematical and Computer Modelling*, **29**, 229–233.

— (2001) Maximum likelihood estimation and diagnostics for stable distributions. In *Lévy Processes: Theory and Applications* (eds. O. E. Barndorff-Nielsen, T. Mikosch and S. I. Resnick), 379–400. Boston: Birkhauser.

— (2007) *Stable Distributions - Models for Heavy Tailed Data.* Boston: Birkhauser. In progress, Chapter 1 online at academic2.american.edu/∼jpnolan.

Paulson, A. S., Holcomb, E. W. and Leitch, R. A. (1975) The estimation of the parameters of the stable laws. *Biometrika*, **62**, 163–170.

Press, S. J. (1972) Estimation in univariate and multivariate stable distributions. *Journal of the American Statistical Association*, **67**.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (2007) *Numerical Recipes: The Art of Scientific Computing*. New York, NY: Cambridge University Press.

Rao, C. R. (1973) *Linear statistical inference and its applications*. New York: John Wiley & Sons, Inc., 2 edn.

Rényi, A. (1961) On measures of entropy and information. *Proc. Fourth Berkeley Symp. Math. Stat. and Probability*, **1**, 547–561.

Robert, C. P. and Casella, G. (2004) *Monte Carlo Statistical Methods*. Springer, 2 edn.

Roverato, A. (2000) Cholesky decomposition of a hyper inverse Wishart matrix. *Biometrika*, **87**, 99–112.

Samorodnitsky, G. and Taqqu, M. S. (1994) *Stable non-Gaussian random processes*. Boca Raton, Florida: Chapman & Hall CRC.

Schraudolph, N. (2004) Gradient-based manipulation of nonparametric entropy estimates. *IEEE Transactions on Neural Networks*, **15**, 828–837.

Shakhnarovich, G., Darrell, T. and Indyk, P. (2006) Introduction. In *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice* (eds. G. Shakhnarovich, T. Darrell and P. Indyk), 1–12. Cambridge, MA: The MIT Press.

Shannon, C. E. and Weaver, W. (1949) *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Shlens, J., Kennel, M., Abarbanel, H. and Chichilnisky, E. (2007) Estimating information rates with confidence intervals in neural spike trains. *Neural computation*, **19**, 1683–1719.

Torgerson, W. S. (1958) *Theory and Methods of Scaling*. New York, NY: Wiley.

Tsallis, C. (1988) Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, **52**, 479–487.

Tukey, J. W. (1962) The future of data analysis. *The Annals of Mathematical Statistics*, **33**, 1–67.

Vempala, S. (2004) *The Random Projection Method*. Providence, RI: American Mathematical Society.

Weron, R. (1996) On the Chambers-Mallows-Stuck method for simulating skewed stable random variables. *Statistics & Probability Letters*, **28**, 165–171.

Whang, K.-Y., Vander-Zanden, B. T. and Taylor, H. M. (1990) A linear-time probabilistic counting algorithm for database applications. *ACM Trans. Database Syst.*, **15**, 208–229.

Woodruff, D. P. (2004) Optimal space lower bounds for all frequency moments. In *Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms (SODA)*, 167–175.

Yang, C. (2001) MACS: music audio characteristic sequence indexing for similarity retrieval. In *Applications of Signal Processing to Audio and Acoustics*, 123–126.

Zolotarev, V. M. (1986) *One-dimensional stable distributions.* Providence, RI: American Mathematical Society.