# Nonparametric Analysis of Univariate Heavy-Tailed Data

## Research and Practice

**Natalia Markovich**
*Institute of Control Sciences,*
*Russian Academy of Sciences,*
*Moscow, Russia*

BICENTENNIAL
1807
WILEY
2007
BICENTENNIAL

John Wiley & Sons, Ltd

# Nonparametric Analysis of Univariate Heavy-Tailed Data

WILEY SERIES IN PROBABILITY AND STATISTICS

A complete list of the titles in this series appears at the end of this volume.

# Nonparametric Analysis of Univariate Heavy-Tailed Data

## Research and Practice

**Natalia Markovich**
*Institute of Control Sciences,*
*Russian Academy of Sciences,*
*Moscow, Russia*

John Wiley & Sons, Ltd

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wiley.com

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

***Other Wiley Editorial Offices***

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 42 McDougall Street, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 6045 Freemont Blvd, Mississauga, ONT, L5R 4J3

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Anniversary Logo Design: Richard J. Pacifico

***British Library Cataloguing in Publication Data***

A catalogue record for this book is available from the British Library

ISBN 978-0470-51087-2

Typeset in 10/12pt Times by Integra Software Services Pvt. Ltd, Pondicherry, India
Printed and bound in Great Britain by TJ International, Padstow, Cornwall
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

*To my parents and daughter*

# Contents

**Appendices**

# Preface

Heavy-tailed distributions are typical of phenomena in complex multi-component systems such as biometry, economics, ecological systems, sociology, Web access statistics and Internet traffic, bibliometrics, finance and business. Typical examples of such distributions are Pareto, Weibull with shape parameter less than 1, Cauchy, and Zipf–Mandelbrot law. Heavy-tailed distributions have been accepted as realistic models for various phenomena: WWW session and TCP flow characteristics (e.g., sizes and durations), on/off-periods of packet traffic, file sizes, service time and input in queuing models, flood levels of rivers, major insurance claims, extreme levels of ozone concentration, high wind-speed values, wave heights during a storm, and low and high temperatures. Examples of applications can be found in the books by Embrechts et al. (1997), Adler et al. (1998), Coles (2001), Beirlant et al. (2004), Reiss and Thomas (2005), McNeil et al. (2005), and Castillo et al. (2006). In both populations of living individuals and inanimate objects such as automobile motors a common tendency has been discovered: the mortality risk for living objects (or the hazard rate for inanimate objects) decreases at infinity, which corresponds to heavy-tailed distributions (Yashin et al., 1996). Insurance company disasters caused by large claims, the overloading of computers by large files and of energy networks by strong deviations of weather and climate phenomena from the average behavior are rare and dangerous events. The methodology described in the book is therefore of current interest.

The analysis of heavy-tailed distributions requires special methods of estimation because of their specific features: slower than exponential decay to zero, violation of Cramér's condition, possible nonexistence of some moments, and sparse observations at the tail domain of the distribution. For example, the central limit theorem, which states the convergence of sums of independent and identically distributed (i.i.d.) random variables (r.v.s) to a Gaussian limit distribution, holds for a large variety of distributions: all we need is a finite variance of the summands.

If this variance is infinite, then we get so-called stable distributions as limit distributions of the normalized sums (Lévy, 1925; Khintchine and Lévy, 1936). Cramér's condition, which states the existence of the moment generating function, is violated for heavy-tailed distributions. Therefore, many results of the large deviation theory that require Cramér's condition (e.g., Cramér's theorem, which states the convergence of the tail of the finite sum of i.i.d. r.v.s to a Gaussian tail) are violated (Petrov, 1975). A linear approximation of the renewal function (RF) for large time intervals of observation changes for an infinite second moment as well.

The statistical analysis of heavy-tailed distributions requires special methods that differ from classical tools due to the sparse observations in the tail domain of the distribution. For example, the histogram is a powerful tool of visual statistical data analysis. Small isolated bars often arise in histogram plots. The data which correspond to such bars are called 'outliers' and the compact mass of the bars is called the 'body' of the distribution. In classical textbooks the 'outliers' are considered as trash, deemed to be present in the sample as a result of some mistake. The usual recommendation is to remove them before any serious analysis or to use robust methods which are stable with respect to contamination of the data. But in many cases the 'outliers' are a vital part of the data; for example, the size of files transported by a network during the transfer of some firm's home page may vary from kilobytes to megabytes (see Crovella et al., 1998). In a histogram large sizes will be viewed as apparent 'outliers'. A network administrator who controls the operation of the network must take into account the existence of such files to avoid network overload. Theoretically, those data where the 'outliers' play a significant role are described by heavy-tailed distributions (Sigman, 1999).

For compactly supported and light-tailed distributions (i.e., those without heavy tails) the histogram is a good estimate of the corresponding probability density function (PDF). But if the distribution is heavy-tailed, the histogram provides misleading peaks in the 'tail' domain or oversmoothes the 'body' of the PDF. The same is true for most of the common nonparametric PDF estimates such as kernel, projection and spline estimates (Čencov, 1982; Silverman, 1986; Devroye and Györfi, 1985).

Usually, quantiles can be estimated by means of an empirical distribution function or weighted estimators based on sample order statistics. However, high quantiles (e.g., 99% or 99.9%) cannot be calculated in the usual way, since the empirical distribution function is equal to 1 outside the range of the sample.

The hazard rate function decays to zero at infinity for heavy-tailed distributions, whereas it increases at infinity for light-tailed distributions and is constant for the exponential distribution. Hence, its estimation has to be different for various classes of distributions.

Ignoring heavy tails in the data may lead to serious distortions of the estimation and errors in system control.

This book focuses mainly on nonparametric methods of the statistical analysis of univariate heavy-tailed i.i.d. r.v.s from samples of moderate sizes. However, the methods are widely useful for dependent data. Dependence detection, the estimation of the PDF from dependent data and elements of bivariate analysis are therefore also considered.

The estimation of the PDF from empirical data is a central problem in mathematical statistics. The PDF is used for the description of the sample, classification, failure time detection, the construction of generators of random numbers, and the estimation of different functionals of the PDF such as the hazard rate function. The estimation of marginal distributions is the first step towards a multivariate analysis.

Traditionally, two main sets of methods, the block maxima method and the peaks over thresholds (POT) method have been developed to estimate tail measures of the risk such as probabilities of exceeding high levels, high quantiles (called value-at-risk (VaR) in finance), and expected shortfall (Embrechts et al., 1997; Coles, 2001; Beirlant et al., 2004; McNeil et al., 2005). The block maxima (i.e., a set of maximal values selected in the blocks of data) are modelled by a generalized extreme value (GEV) distribution with distribution function (DF) $G(x) = \exp\{-(1+\gamma(x-\mu)/\sigma)_+^{-1/\gamma}\}$. In the POT method the values which are larger than some thresholds are modelled by the generalized Pareto distribution (GPD) with DF $\Psi_{\sigma,\gamma}(x) = 1-(1+\gamma x/\sigma)_+^{-1/\gamma}$. The parameters in these models (in particular the tail index $1/\gamma$) are estimated from a sample using nonparametric methods (e.g., Hill's method) or parametric methods (e.g., maximum likelihood).

In practice, we often need an estimate of the whole PDF or DF, both the 'tail' and the 'body', for example for classification or the estimation of the expectation. Another example is given by the copula technique (and, generally, multivariate analysis) which suggests the estimation of marginal distributions based on all data (Mikosch, 2006). The parametric tail models considered are not a good fit for the whole DF and the PDF and, hence, are not appropriate for such aims. Therefore, in this book, much attention is devoted to the nonparametric estimation of heavy-tailed PDFs.

We consider three sets of estimators of the whole heavy-tailed PDF that are purely or partly nonparametric. These are variable bandwidth kernel estimators, combined estimators that fit the 'tail' and the 'body' of the PDF by parametric and nonparametric models respectively, and estimators based on the transformation approach.

The need for different amounts of smoothing at different locations of heavy-tailed PDFs leads to the usage of kernel estimators with window width (or, roughly speaking, the 'width' of the kernel) varying from one point to another, that is, variable bandwidth kernel estimators (Abramson, 1982; Hall, 1992; Silverman, 1986). However, these estimators, at least with compactly supported kernels, are not intended for the estimation of a heavy-tailed PDF in the 'tail' domain, where the observations are sparse. This is because the latter estimators are defined on finite intervals. These are approximately the same as the ranges of the samples.

Application of heavy-tailed kernels for variable bandwidth kernel estimators has yet to be investigated in the literature.

It is obvious that nonparametric PDF estimates with good behavior in the ‘tail’ domain are required. This feature is significant for classification (pattern recognition) purposes when the PDFs of many populations are compared. If one uses an empirical Bayesian classification algorithm, then the observations will be classified by the comparison of the corresponding PDF estimates of each class. Since the object can arise in the ‘tail’ domain as well as in the ‘body’, a tail estimator with good properties is of primary importance for classification.

To improve the PDF estimation at infinity a transform–retransform scheme is considered here. This scheme implies a preliminary transformation of the data to a finite interval, that is, to a sample with a PDF that is more convenient for the estimation. Then one can estimate the PDF of a new r.v. obtained by the transformation by means of some nonparametric method and get the PDF of the original data by the reverse transformation of the PDF estimate of the transformed data. Furthermore, the back-transformed PDF estimates with fixed smoothing parameters work like location-adaptive estimates and allow the estimation of the PDF to be improved on the entire domain on which it is defined. Logarithmic transformations are a popular choice with this approach.

In this book, combinations of data transformations and nonparametric estimates are considered that provide accurate PDF estimation and have decay rates at infinity close to those of the original PDFs. In this respect, a good deal of attention is devoted to a so-called adaptive transformation to a finite interval, which uses essentially the asymptotic distribution of the maximum of the sample as a model of the distribution behavior at infinity. The latter idea is followed throughout the book: an adaptive transformation may be applied to the PDF, and high quantile and hazard rate estimation to classification.

A parametric–nonparametric estimation combines the advantages of parametric tail models to describe the ‘tail’ well enough and nonparametric methods to describe the ‘body’ domain (i.e., that limited area of relatively small values of an underlying r.v.) better. A similar idea was proposed in Barron et al. (1992), where a parametric model of the ‘tail’ of the PDF is superimposed on a histogram estimate of the ‘body’. Despite its ease of application, it is extremely sensitive to the correct choice of the parametric family and may provide a poor fit of the ‘body’ of a PDF in the case of moderate sample sizes. In practice, we often observe r.v.s governed by multimodal heavy-tailed distributions. Hence, it is important to use combined estimators aimed at accurately fitting both the multimodal ‘body’ and the ‘tail’ of the PDF.

For practical needs, it is more important to provide such estimates of the PDF that are more suited to the tasks in hand. That is why another topic of the book concerns the investigation of the capacities of the PDF estimates considered with regard to the pattern recognition problem. Many methods of classification that use PDF estimates are known (Silverman, 1986; Aivazyan et al., 1989). We consider a procedure that allows increased influence of ‘outliers’ in the tail domain on the quality of the classification, thus preventing large misclassification losses by rare events.

High quantile estimates for heavy-tailed distributions are applied to determine the values of characteristics of observed objects that may lead to rare but large losses. High quantiles indicate the VaRs in finance or the thresholds of parameters in complex systems such as the Internet (e.g., the 99.9% quantile can provide the maximal threshold for the file size) or atomic power stations. In this book, we discuss some but not all known high quantile estimators.

The tail index is a key characteristic of heavy-tailed data. It shows the shape of the tail of the distribution without making any assumption regarding the parametric form of the tail. By means of the tail index, one can identify a heavy tail in measurements and the number of finite moments. All characteristics of heavy-tailed r.v.s are based on the tail index. In this book, many well-known estimators of the tail index such as Hill's, POT, moment, UH, and ratio estimators are considered. Furthermore, a relatively new tail index estimator, proposed in Davydov et al. (2000) – called the group estimator here – is described. It has the essential advantage that it can be calculated recursively. The latter property is convenient for on-line estimation.

The mortality risk function plays a significant role in population analysis. It is connected with the finding of causes of certain events in the population such as morbidity and mortality. This function is called the hazard rate if the reliability of technical systems is under investigation. Hitherto, most analysts have used the parametric approach for mortality risk estimation from empirical data. This means that before carrying out the estimation one decides what kind of function the mortality risk is expected to be. However, it might be difficult to describe the data by means of these models sufficiently accurately applying the cause factors as parameters. The parametric approach is problematic for the analysis of population processes by means of semi-Markov models when the intensity of the appearance of events is interpreted as an intensity of the transition from one state to another. An alternative approach is to use nonparametric models, when only general information about the estimated function is available. For the estimation of the hazard rate, however, the nonparametric approach is rarely used: in the literature, the preliminary estimation of the PDF and the DF by kernel or histogram-type estimators (Prakasa Rao, 1983) and regularized estimates (Stefanyuk, 1992) has been considered. One reason for this is a specific difficulty arising from the different asymptotic behavior of this function in the right-hand part of its domain for light- and heavy-tailed distributions. Hence, its estimation has to be different for various classes of distributions. In this book, the data transformation approach to a finite interval is considered to estimate the hazard rates corresponding to compactly supported distributions by nonparametric methods. The estimation of the hazard rate is presented as an inverse ill-posed problem involving Volterra's integral equation, and a so-called regularization method, (Tikhonov and Arsenin, 1977) is used to find its approximation. The estimation of the hazard rate and the hazard rate ratio is considered for a biological application (the problem of hormesis detection) and for teletraffic problems.

For the purposes of warranty control, reliability analysis of technical systems, and particularly of telecommunication networks, one often needs to estimate the RF. This function is equal to the mean number of arrivals of the relevant events before a fixed time. Usually, measurement facilities count the events of interest, for example, the number of requested and transferred Web pages, incoming or outgoing calls in consecutive time intervals of fixed length. To estimate the RF, several realizations of the counting process (e.g., observations of number of calls over several days) may be required, with further averaging inside the corresponding time interval. However, it may be that the RF has to be estimated using only one set of inter-arrival times between events. This applies particularly to warranty control or when it would be too expensive to obtain numerous observations of the process. Explicit forms of the RF are obtained only for a few inter-arrival time distributions such as the uniform, exponential, Erlang or normal (Asmussen, 1996). The preliminary estimation of the DF or the PDF, if the latter exists, may become a more complicated problem than direct estimation of the RF. Here, the main attention is devoted to the nonparametric estimation of the RF from a sample of the i.i.d. inter-arrival times between events of moderate size. A few known results in this area (Frees, 1986a, 1986b; Grübel and Pitts, 1993; Schneider et al., 1990; Markovitch and Krieger, 2002b; Markovich and Krieger, 2006a) are discussed in this book. The well-known Frees estimate requires a huge amount of calculation even if one operates with samples as small as 20–30 observations. A sufficiently accurate estimate of the RF from empirical data is discussed that is also feasible for large samples. As always, the key problem of nonparametric estimates is the choice of the parameter that is responsible for the smoothing. Hence, the data-dependent selection of a smoothing parameter of the RF estimates is the main object of interest here.

## The main methodology

The statistical tools considered are based on the results of probability theory, mathematical statistics, extreme value theory, and the theory of the solution of ill-posed operator equations. The statistical methodology considered in this book is elaborated for the evaluation of characteristics of heavy-tailed r.v.s from samples of moderate size.

Due to the lack of information beyond the range of the sample, nonparametric statistical estimation is based essentially on the asymptotic distribution of sample maxima as a model of the distribution behavior at infinity. The basic result of extreme value theory concerning the asymptotic behavior of the marginal distribution of the sample maxima (a GEV distribution) was provided by Gnedenko (1943). This result was extended to multivariate extreme value distributions by Galambos (1987).

The asymptotical tail distribution is the only realistic knowledge regarding the behavior of the distribution beyond the range of the sample. A data transformation approach that is discussed at length in the book essentially uses these asymptotic results. This approach allows us to transform the initial r.v. that is assumed to be GEV distributed into a new one. The latter may be located in a finite interval. That may both simplify the estimation (e.g., the estimation of the PDF) and allow us to apply some relevant estimators such as the histogram, or projection estimators that are applicable just for distributions with compact supports. The data transformations can be useful for the further development and the identification of models of multivariate distributions. It is known that such tools as copulas are invariant with respect to monotone transformations of r.v.s. That may give rise to construct dependence measures and models for 'conveniently distributed' r.v.s just using reliable transformations.

Another methodology considered in the book is given by a statistical regularization method. This has evolved from Tikhonov's regularization theory (Tikhonov and Arsenin, 1977). The latter theory was intended for the solution of deterministic linear and nonlinear operator equations. Due to the uncertainties in the availability of an operator and the right-hand part of the operator equation, the solution may be related to an ill-posed problem. Unlike Tikhonov's method the method considered deals with stochastic operator equations. This approach was elaborated in Vapnik and Stefanyuk (1979), Vapnik (1982), and Stefanyuk (1986), and applied to population analysis in Markovich and Michalski (1995) and Markovich (1995, 2000) and to the analysis of teletraffic systems in Markovitch and Krieger (1999). Regularization is a developing area and is not restricted by the framework of Tikhonov's scheme. The next step could be a wider application of other regularization schemes to statistical applications.

In this book, the nonparametric estimation of characteristics of r.v.s plays a significant role. A smoothing of nonparametric estimates, for instance, the choice of the bin width in a histogram or the bandwidth in kernel estimators of the PDF, is key to an accurate approximation. The values of smoothing parameters recommended by theory usually minimize the mean squared error of the estimate or its asymptotic analog. This gives the values that are functions of a sample size. In practice, where one deals with samples of moderate sizes such values of parameters can provide unsatisfactory estimates. That is why, in this book, much attention is focused on data-dependent methods such as a cross-validation (Wahba, 1981) and the discrepancy method (Markovich, 1989; Vapnik et al., 1992). The stochastic version of the discrepancy method has evolved from the discrepancy method for deterministic operator equations (Morozov, 1984).

Another approach is based on the minimization of an empirical bootstrap estimate of the mean squared error of the estimate by an unknown parameter. Bootstrapping is a tool for obtaining a reasonable value of an unknown smoothing parameter.

## What is new?

The book contains many results from the author's advanced research material that are presented for the first time. These are:

(i) the combined parametric–nonparametric estimator of a PDF;

(ii) the adaptive data transformation that allows the PDF to be fitted at infinity better than a pure nonparametric estimate;

(iii) the discrepancy method as a data-dependent smoothing tool of nonparametric PDF estimates;

(iv) the application of the retransformed PDF estimates for classification;

(v) on-line recursive estimation of the tail index;

(vi) a modification of Weissman's estimator of high quantiles that has smaller mean squared error;

(vii) regularized estimates of the hazard rate function and hazard rate ratio;

(viii) the estimator of the RF at finite time intervals from samples of inter-arrival times of moderate sizes;

(ix) the bootstrap and plot methods as data-dependent smoothing tools for selecting a smoothing parameter in the RF estimator.

Many practical recommendations for the implementation of the presented estimators are given, namely:

(i) the use of nonparametric PDF estimates in finance, telecommunication, population analysis, and multivariate analysis;

(ii) the usage of the classification methodology for the clustering of Internet data and Web prefetching;

(iii) the usage of high quantile estimates in finance and the identification of parameter bounds in technical systems;

(iv) the application of the hazard rate function in teletraffic (e.g., retrial call rate estimation);

(v) the application of the hazard rate ratio in population analysis (e.g., hormesis detection) and for failure time detection;

(vi) the application of RF estimates for overload control of telecommunication systems and warranty control;

(vii) the rough detection of heavy tails and dependence in data and the application of these methods to Web traffic and TCP flow data by way of illustration.

The reader can easily learn how to do a rough and more advanced statistical analysis of the data.

## Content and general outline of the book

The book gives a detailed survey of classical results and recent developments in the theory of nonparametric estimation of the PDF, the tail index, high quantiles, the hazard rate and the renewal function assuming the data come from i.i.d. random variables with heavy tails. Both asymptotic results such as convergence rates of the estimates and results for samples of moderate sizes supported by Monte Carlo investigation are considered. Special comments are also made on the application of the methods considered to dependent data. Observations that serve to clarify the main line of the exposition are located in footnotes.

In Chapter 1 definitions and basic properties of classes of heavy-tailed distributions are considered. Tail index estimation and methods for the selection of the number of largest order statistics in Hill's estimator are presented. Rough methods for the detection of heavy tails and the number of finite moments as well as dependence detection and simple bivariate analysis provide the ideas for a preliminary statistical data analysis. The methods considered are applied to measurements of Web traffic and TCP flows.

Chapter 2 is devoted to PDF estimation. The main principles and the links between them are presented. Classical nonparametric estimators of the PDFs and smoothing methods are considered. PDF estimation using dependent data is discussed. Examples of the applications of PDF estimates are given.

Chapter 3 describes three classes of heavy-tailed PDF estimation methods. These are methods that 'paste' together the parametric tail models and nonparametric estimates of the main part of the PDF (e.g., the combined parametric–nonparametric method and Barron's estimator), the variable bandwidth kernel estimators, and the retransformed nonparametric estimators that use transformations of the data.

In Chapter 4 so-called fixed and adaptive transformations are proposed. The difference between them is that fixed transformations do not depend on the distribution, in contrast to adaptive transformations. These transformations are applied to improve the estimation of heavy-tailed PDFs. Special boundary kernels are considered to improve the behavior of retransformed kernel estimates at infinity. The key problem of any nonparametric estimator is the choice of a smoothing parameter that determines the accuracy of the estimation. Data-dependent discrepancy methods are investigated both for nonvariable and variable bandwidth kernel estimators as well as for a projection estimator. The mean squared errors of these estimates are proved to be optimal.

In Chapter 5 the application of the retransformed PDF estimates described in the previous chapter to the classification problem is considered. An empirical Bayesian algorithm is used. Then any new observation is classified by the comparison of the corresponding PDFs of each class. The retransformed kernel and polygram estimators are used to estimate heavy-tailed PDFs of each class. The accuracy of the classifiers obtained is compared by a simulation study. Possible applications of this classification technique to Web traffic data analysis and Web prefetching are considered.

Chapter 6 contains estimators of the high quantiles for heavy-tailed distributions. The estimates are compared by a Monte Carlo study using simulated r.v.s. The distribution of the logarithm of the ratio of Weissman's estimate to the true value of the quantile is proved to be asymptotically normal. The same result is obtained for the modification of Weissman's estimate. An application to WWW traffic data is considered.

Chapter 7 elaborates the nonparametric estimation of the hazard rate function in light- and heavy-tailed cases. The statistical regularization method and its theoretical background are presented. The application of the hazard rate and hazard rate ratio to telecommunication and population analysis is discussed.

Finally, Chapter 8 includes the estimation of the renewal function within finite and infinite time intervals. Nonparametric estimators for finite intervals, their asymptotical theoretical properties and smoothing methods are considered.

The companion website for the book is http://www.wiley.com/go/nonparametric

## Audience

This book is intended as a practical manual on the statistical theory of heavy-tailed data. The exposition is accompanied by numerous illustrations and examples motivated by applications in telecommunication, population analysis, and finance. Each chapter is provided with exercises. These may help the reader to understand the application of the statistical methods presented. The book assumes only an elementary knowledge of probability theory and statistical methods. Sometimes the subject requires the use of intermediate mathematical techniques such as probability theory, statistics, and mathematical analysis.

The book is aimed at a relatively broad audience including students, practitioners, and engineers who are faced with analyzing heavy-tailed empirical data and are interested in the rough methodology and algorithms for numerical calculations related to the analysis of heavy-tailed data, as well as researchers and PhD students who are looking for new approaches and fundamental results, supported by proofs. Readers are expected to have diverse backgrounds including computer science, performance evaluation engineering, statistics, economics, demography, and population analysis. Readers with an interest in applied areas can skip the proofs of the theorems located in the appendices.

## Acknowledgments

This text was mainly developed in my habilitation thesis entitled 'Estimation of characteristics of heavy-tailed random variables by limited samples' and in a PhD course entitled 'Analysis methods of heavy-tailed data'. It was given as a part of

the European project ‘EuroNGI. Design and engineering of the Next Generation Internet’. I am grateful to all my students for their corrections to the exercises.

In particular, I would like to express my sincere thanks to Prof. A. Yagola of Lomonosov State University in Moscow, Prof. A. Kukush and Prof. R. Maiboroda of Shevchenko State University in Ukraine, and Dr. Sci. A. Dobrovidov and Dr. A. Stefanyuk of the Institute of Control Sciences, Moscow, for reviewing my habilitation thesis and useful discussions of my results. Prof. Dr. U. Krieger of Otto-Friedrich University, Bamberg, Germany, undertook to read the whole manuscript and made useful comments. I take pleasure in thanking Prof. P. Tran-Gia and Dr. Norbert Vicari of the Julius-Maximilians-University at Würzburg, Jorma Kilpi of VTT, Helsinki, who have provided me with teletraffic data. I express my special thanks to my PhD supervisor, Prof. V. Vapnik of Columbia University, who laid down the path for me to follow in my scientific life and the opportunity to become a working member of the Institute of Control Sciences of the Russian Academy of Sciences.

The partial financial support I have received from the European research project ‘EuroNGI. Design and engineering of the Next Generation Internet’ and its continuation ‘EuroFGI’ (contracts no. 507613 and 028022) is also greately appreciated.

I want to thank the entire Wiley team and, in particular, my copy-editor, Richard Leigh.

# 1

# Definitions and rough detection of tail heaviness

In this chapter, the basic definitions and properties of heavy-tailed distributions are presented. Tail index estimation and methods for selecting the number of largest order statistics in the Hill estimator are discussed. Rough methods for the detection of heavy tails, the number of finite moments, dependence and long-range dependence are described. Elements of bivariate analysis are presented: estimation of the Pickands function and bivariate quantiles. The latter methods are applied to the analysis of telecommunication data.

## 1.1 Definitions and basic properties of classes of heavy-tailed distributions

We start with the common definitions.

**Definition 1** *The set* $(\Omega, \mathcal{A}, P)$ *is called the probability space, where* $\Omega$ *is the space of elementary events,* $\mathcal{A}$ *is a* $\sigma$*-algebra of subsets of* $\Omega$*, and* $P$ *is a probability measure on* $\mathcal{A}$*.*

Let $(\Omega, \mathcal{A})$ be some measurable space, $(R, \mathcal{B}(R))$ be the real line with the $\sigma$-algebra $\mathcal{B}(R)$ of Borelian sets on $R$.

**Definition 2** *The real-valued function $X = X(\omega)$ defined on $(\Omega, \mathcal{A})$, is called a random variable (r.v.), if for any $B \subseteq \mathcal{B}(R)$ $\{\omega : X(\omega) \in B\} \subseteq \mathcal{A}$ holds.*

**Definition 3** *The function $F_X(x) = P\{\omega : X(\omega) \leq x\}$, $x \in R$, is called the distribution function (DF) of the r.v. $X$.*

**Definition 4** *Let a nonnegative real-valued function $f(t)$, $t \in R$, exist such that for all $x \in R$,*

$$F_X(x) = \int_{-\infty}^{x} f(t)dt.$$

*The function $f(t)$, $t \in R$, is called the probability density function (PDF) of r.v. $X$.*

**Definition 5** *The r.v.s $X_1, X_2, \dots, X_n$ ($X_i \in B_i \subseteq R$, $B_i$ is a finite set) are called independent if, for any $x_1, x_2, \dots, x_n \in R$,*

$$P\{X_1 = x_1, \dots, X_n = x_n\} = P\{X_1 = x_1\} \dots P\{X_n = x_n\}$$

*or equivalently, for any $B_1, \dots, B_n \in \mathcal{B}(R)$,*

$$P\{X_1 \in B_1, \dots, X_n \in B_n\} = P\{X_1 \in B_1\} \dots P\{X_n \in B_n\}.$$

In terms of DFs and PDFs, independence means that

$$F(x_1, x_2, \dots, x_n) = F_1(x_1)F_2(x_2) \dots F_n(x_n),$$

and

$$f(x_1, x_2, \dots, x_n) = f_1(x_1)f_2(x_2) \dots f_n(x_n),$$

where $F_k(x_k)$ and $f_k(x_k)$ are the DF and PDF of the r.v. $X_k$.

The definition of heavy-tailed distributions may be derived from the extreme value theory. Let $X^n = \{X_1, \dots, X_n\}$ be a sample of independent and identically distributed (i.i.d.) r.v.s with DF $F(x) = P\{X_1 \leq x\}$ and $M_n = \max(X_1, X_2, \dots, X_n)$. It is known (Gnedenko, 1943; David, 1981) that if the limit distribution of maxima $M_n$ exists then there exist normalizing constants $a_n, b_n$ such that

$$P\{(M_n - b_n)/a_n \leq x\} = F^n(b_n + a_n x) \to_{n\to\infty} H_\gamma(x), \qquad x \in R, \tag{1.1}$$

and an extreme value DF $H_\gamma(x)$ belongs to one of the following types of distribution function:[1]

$$H_\gamma(x) = \begin{cases} \exp(-x^{-1/\gamma}), & x > 0, \gamma > 0 \quad \text{(Fréchet)}, \\ \exp(-(-x)^{-1/\gamma}), & x < 0, \ \gamma < 0 \quad \text{(Weibull)}, \\ \exp(-e^{-x}), & \gamma = 0, x \in R \quad \text{(Gumbel)}. \end{cases} \tag{1.2}$$

The distribution $H_\gamma(x)$ can also be rewritten as

$$H_\gamma(x) = \begin{cases} \exp(-(1 + \gamma x)^{-1/\gamma}), & \gamma \neq 0, \\ \exp(-e^{-x}), & \gamma = 0, \end{cases} \tag{1.3}$$

[1] This result remains true if $X_1, \dots, X_n$ are weak dependent (Leadbetter et al., 1983).

where $1+\gamma x > 0$ (Jenkinson–von Mises representation). $H_\gamma(x)$ is called a standard generalized extreme value (GEV) distribution.

**Example 1** **(Coles, 2001)** *If $X^n$ is a sequence of independent standard exponential r.v.s with DF $F(x) = 1 - \exp(-x)$ for $x > 0$ then, letting $a_n = 1$ and $b_n = n$ in* (1.1), *the limit distribution of $M_n$ as $n \to \infty$ is the Gumbel distribution. In the case of standard Fréchet r.v.s with DF $F(x) = \exp(-1/x)$ and $a_n = n$ and $b_n = 0$, the limit distribution of $M_n$ is precisely the standard Fréchet distribution with $\gamma = 1$ in* (1.2). *Let $X^n$ be a sequence of independent uniform r.v.s on $[0, 1]$ with DF $F(x) = x$ for $x \in [0, 1]$ and $a_n = 1/n$ and $b_n = 1$. Then the limit distribution of $M_n$ is of Weibull type with $\gamma = -1$.*

**Definition 6** *The parameter $\gamma$ is called the extreme value index (EVI) and defines the shape of the tail of the r.v. X. The parameter $\alpha = 1/\gamma$ is called the tail index.*

**Definition 7** *We say that the r.v. X and its distribution F belong to the maximum domain of attraction of $H_\gamma(x)$ if (1.1) is fulfilled. We write $X \in \mathrm{MDA}(H_\gamma)$ $(F \in \mathrm{MDA}(H_\gamma))$.*

We shall consider only nonnegative r.v.s.

**Definition 8** *A DF $F(x)$ (or the r.v. X) is called heavy-tailed if its tail $\bar{F}(x) = 1 - F(x) > 0$, $x \geq 0$, satisfies, for all $y \geq 0$,*

$$\lim_{x\to\infty} P\{X > x + y | X > x\} = \lim_{x\to\infty} \bar{F}(x+y)/\bar{F}(x) = 1.$$

This intuitively implies that if $X$ exceeds a large value then it will most probably exceed any larger value, too.

Roughly speaking, heavy-tailed distributions belong to the class of those long-tailed distributions whose tails decay to 0 slower than an exponential tail (Figure 1.1). The exponential distribution is often considered as a boundary between classes of heavy-tailed and light-tailed distributions. Typical examples of heavy- and light-tailed distributions are given in Table 1.1.

The class of heavy-tailed distributions comprises the subexponential class of distributions ($S$) and its subset, that is, distributions with regularly varying tails.

**Definition 9** *The DF $F(x)$ (or the r.v. X), defined on $(0, \infty)$, is called subexponential ($F \in S$ ($X \in S$)), if*

$$P\{S_n > x\} \sim nP\{X_1 > x\} \sim P\{M_n > x\} \qquad \text{as } x \to \infty,^2$$

*for some $n \geq 2$, where $S_n = X_1 + \ldots + X_n$, $M_n = \max_{i=1,\ldots,n}\{X_i\}$.*

[2] For any positive functions $f$ and $g$, $f \sim g$ as $x \to x_1$ means that $\lim_{x\to x_1} f(x)/g(x) = 1$.

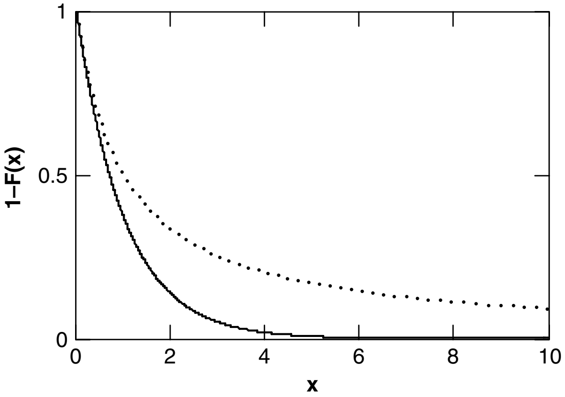

**Figure 1.1** Comparison of tail behavior: exponential distribution (solid line), Pareto distribution (dotted line).

**Table 1.1** Examples of heavy- and light-tailed distributions.

| | |
|---|---|
| *Heavy-tailed distributions* | *Subexponential:*<br>Pareto, lognormal, Weibull with shape parameter less than 1<br>*With regularly varying tails:*<br>Pareto, Cauchy, Burr, Fréchet, Zipf–Mandelbrot law |
| *Light-tailed distributions* | exponential, gamma, Weibull with shape parameter greater than 1, normal, compactly supported distributions |

Intuitively, subexponentiality means that the only way the sum can be large is by one of the summands getting large (in contrast to the light-tailed case, where all summands are large if the sum is so).

**Definition 10** *The DF $F$ (or r.v. $X$) is called a regularly varying distribution at infinity of index $\alpha = 1/\gamma$, $\gamma > 0$ ($X \in R_{-1/\gamma}$), if*

$$P\{X > x\} = x^{-1/\gamma}\ell(x), \ \forall x > 0, \tag{1.4}$$

*where $\ell(x)$ is called a slowly varying function ($\ell(x) \in R_0$).*

**Definition 11** *A positive, Lebesgue measurable function $\ell(x)$ on $(0, \infty)$ is called a slowly varying function at infinity if* $\lim_{x\to\infty} \ell(tx)/\ell(x) = 1, \forall t > 0$ *(Feller, 1968; Sigman, 1999).*

Examples of $\ell(x)$ are given by $c \ln x$, $c \ln(\ln x)$ and all functions converging to positive constants. Using different functions $\ell(x)$, one can get a great variety of tails.

For light-tailed distributions all moments $E[(X^+)^k]$ exist and are finite. In contrast, for regularly varying distributions the moments $EX^\beta$ are finite only if $\beta < 1/\gamma$.

Basic properties of regularly varying distributions (Breiman, 1965; Bingham et al., 1987; Feller, 1971; Mikosch, 1999; Resnick, 2006) are summaryzed in the following lemma.

**Lemma 1** *Let $X \in R_{-\alpha}$. Then,*

$(i)$ $X \in S$.

$(ii)$ $E\{X^{\beta}\} < \infty$ *if* $\beta < \alpha$, $E\{X^{\beta}\} = \infty$ *if* $\beta > \alpha$.

$(iii)$ *If* $\alpha > 1$, *then* $X^{r} \in R_{1-\alpha}$ *and* $P\{X^{r} > x\} \sim \ell(x)x^{1-\alpha}/((\alpha-1)E\{X\})$ *as* $x \to \infty$.

$(iv)$ *If $Y$ is nonnegative and independent of $X$ such that* $P\{Y > x\} = \ell_2(x)x^{-\alpha_2}$, *then* $X+Y \in R_{-\min(\alpha,\alpha_2)}$ *and* $P\{X+Y > x\} \sim P\{X > x\} + P\{Y > x\}$ *as* $x \to \infty$.

$(v)$ *(Breiman's theorem) If $Y$ is nonnegative and independent of $X$ such that* $E\{Y^{\alpha+\varepsilon}\} < \infty$ *for some* $\varepsilon > 0$, *then* $XY \in R_{-\alpha}$ *and*

$$P\{XY > x\} \sim E\{Y^{\alpha}\}P\{X > x\} \qquad as \quad x \to \infty.$$

Heavy-tailed distributions differ strongly from the normal or exponential distributions; for example, the exponential distribution function $F(x) = 1 - e^{-\lambda x}$, $x \geq 0$, satisfies

$$\overline{F}(x+y)/\overline{F}(x) = \exp(-\lambda y), \qquad x \geq 0, \quad y \geq 0,$$

and hence it is not heavy-tailed.

An important property of heavy-tailed distribution is given by the violation of Cramér's condition. This means that the moment generating function does not satisfy $E(e^{\varepsilon x}) < \infty$, $\varepsilon > 0$. Many results of the large deviation theory require the fulfillment of Cramér's condition. Otherwise, for example, Cramér's theorem on the convergence of $P\{S_n > x\}$ ($S_n$ is the sum of $n$ independent r.v.s) to the tail of a normal distribution is violated. Intervals of normal convergence of heavy-tailed distributions are presented in Mikosch and Nagaev (1998).

In practice, a tail function $\bar{\mathrm{F}}(\mathrm{x})$ is often fitted by the generalized Pareto distribution. The latter is based on Pickands' theorem (Pickands, 1975):

**Theorem 1** *Let $X_1, \ldots, X_n$ be an i.i.d. random sequence. The limit distribution of the excess of the $X_i$ over the threshold $u$ is necessarily of generalized Pareto form,*

$$\lim_{u \uparrow x_F, u+x<x_F} P\left(X_1 - u > x | X_1 > u\right) \to (1+\gamma x)_{+}^{-1/\gamma}, \qquad x \in R,$$

*where*

$$x_F = \sup\{x \in R : F(x) < 1\}$$

*is the right endpoint of the distribution $F(x)$, the shape parameter $\gamma \in R$, and*

$$(x)_+ = \begin{cases} x, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

## 1.2 Tail index estimation

The tail index reflects the shape of the distribution tail (with no assumption on the parametric form of the tail) and, therefore, plays a key role in the analysis of heavy-tailed measurements. The tail index is used for the estimation of high (99%, 99.9%) quantiles of observed r.v.s, the estimation of the PDF of the r.v. (Markovitch and Krieger, 2002a) and, hence, for classification (Maiboroda and Markovich, 2004). It allows one to identify roughly whether the distribution is heavy-tailed or not as well as to determine the number of finite moments.

There are numerous estimators of the EVI $\gamma$. Let $X^n = \{X_1, \ldots, X_n\}$ be i.i.d. r.v.s with common DF $F(x)$.

### 1.2.1 Estimators of a positive-valued tail index

**Hill's estimator for $\gamma = 1/\alpha > 0$**

We assume that $F(x)$ belongs to the class of regularly varying distributions (see Definition 10). For many applications, it is important to know $\alpha$. For example, if $\alpha < 2$, than $EX_1^2 = \infty$ holds. Hill's estimator (Hill, 1975), used for $\gamma = 1/\alpha > 0$, is determined by

$$\widehat{\gamma}^H(n, k) = \frac{1}{k} \sum_{i=1}^{k} \log X_{(n-i+1)} - \log X_{(n-k)}, \tag{1.5}$$

where $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ are the order statistics of the sample $X^n = \{X_1, X_2, \ldots, X_n\}$ and $k$ is a further smoothing parameter.

It is a remarkable feature that the estimator (1.5) may be obtained in several ways – for example, by the maximum likelihood (ML) method assuming $F \in R_{-1/\gamma}$ (Hill, 1975), by the regularly varying approach (de Haan, 1994), by the regression approach (Beirlant et al., 1999), or by using quantiles (Beirlant et al., 2004). For detailed discussion, see Embrechts et al. (1997) and Resnick (2006, Section 4.4).

Hill's estimator is weakly consistent if

$$k \to \infty, \qquad k/n \to 0 \quad \text{as} \quad n \to \infty \tag{1.6}$$

(Mason, 1982), and asymptotically normal with mean $\gamma$ and variance $\gamma^2/k$,

$$\sqrt{k}\left(\widehat{\gamma}^H(n, k) - \gamma\right) \to^d N(0, \gamma^2)$$

(Häusler and Teugels, 1985). In practice, the accuracy of the estimate depends on the selection of $k$. If the r.v. $X \in R_{-1/\gamma}$, then the slowly varying function $\ell(x)$, which is usually unknown, influences the estimation. Hill's estimator does not work well if the r.v. $X$ does not belong to class $R_{-1/\gamma}$. Plots of Hill's estimates against $k$ are

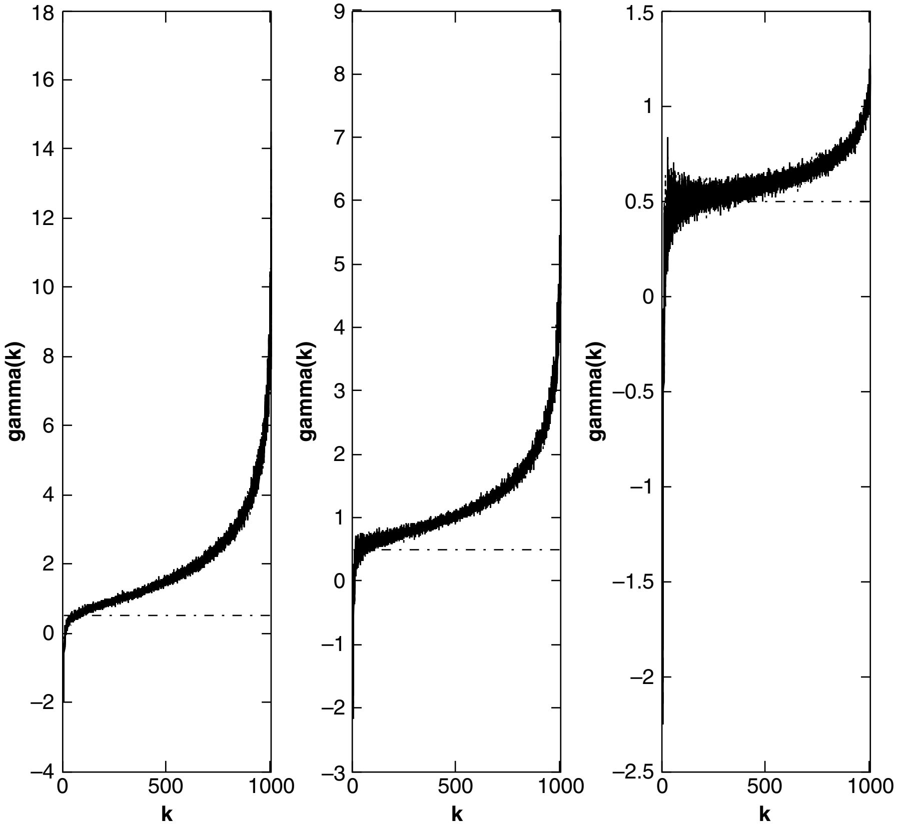

**Figure 1.2** Hill's estimate against $k$ for 15 realizations of the Weibull distribution (left), Pareto (middle) and Fréchet (right) distributions, each with parameter $\alpha = 0.5$ (dotted line). The sample size is $n = 1000$.

shown in Figure 1.2 for 15 realizations of Weibull, Pareto and Fréchet distributions, each with parameter $\alpha = 0.5$.

### The ratio estimator

The ratio estimator

$$a_n = a_n(x_n) = \sum_{i=1}^{n} \ln(X_i/x_n)\mathbf{1}\{X_i > x_n\} / \sum_{i=1}^{n} \mathbf{1}\{X_i > x_n\} \tag{1.7}$$

is a generalization of Hill's estimator in the sense that we use an arbitrary threshold level $x_n$ instead of an order statistic $x_n = X_{(n-k)}$ in (1.5) (Goldie and Smith, 1987). Here, $\mathbf{1}(A)$ is the indicator function of the event $A$. The statistic (1.7) seems to be among a few tail index estimators whose bias and mean squared error (MSE) asymptotics are known (Novak, 1996).

Note that Hill's estimator and the ratio estimator may also be applied to dependent data (Novak, 2002; Resnick and Stărică, 1999). Hill's estimator is very sensitive with respect to dependence in the data (see Ebmrechts et al., 1997). The asymptotic normality of the ratio estimator under the specific mixing condition that is fulfilled in many parametric models (e.g., ARCH and GARCH) is proved in Novak (2002).

## 1.2.2 The choice of *k* in Hill's estimator

### Visual choice of *k*

The parameter $k$ may be estimated visually by means of the exceedance plot, that is, the plot $\{(u, e(u)) : X_{(1)} < u < X_{(n)}\}$. Here

$$e(u) = \sum_{i=1}^{n}(X_i - u)\mathbf{1}\{X_i > u\} / \sum_{i=1}^{n}\mathbf{1}\{X_i > u\} \tag{1.8}$$

is the empirical mean excess function over threshold $u$ of a given sample $X^n$. The linearity of $e(u)$ over some level $u$ corresponds to a Pareto mean $e^P(u) = (1 + \gamma u)/(1 - \gamma)$. Then the number of the order statistic that is the closest to $u$ is accepted as the estimate of $n - k$.

Alternatively, one can estimate $k$ from the Hill plot $\{k, \hat{\gamma}^H(n, k) : k = 1, \ldots, n - 1\}$. The estimate of $k$ is selected from the interval $[k_-, k_+]$ of stability of the function $\hat{\gamma}^H(n, k)$. The latter approach is based on the consistency of Hill's estimator. One may take the mean estimate (1.5) in $[k_-, k_+]$ as the estimate of $\gamma$, that is, $\hat{\gamma}^H(n, k) \approx \gamma$ for all $k \in [k_-, k_+]$, and $k$ corresponding to this $\gamma$ as the optimal value.

Methods of selecting $k$ from empirical data are mostly based on the choice of a trade-off between the bias and the variance of Hill's estimate. The bias increases and the variance decreases, as $k$ increases.

It was proved in Hall and Welsh (1985) that the asymptotical MSE of Hill's estimate is minimal for

$$k_n^{\mathrm{opt}} \sim \left(\frac{C^{2\rho}(\rho + 1)^2}{2D^2\rho^3}\right)^{1/(2\rho+1)} n^{2\rho/(2\rho+1)},$$

if the distribution function satisfies the so-called Hall's condition

$$1 - F(x) = Cx^{-1/\beta}\left(1 + Dx^{-\rho/\beta} + o(x^{-\rho/\beta})\right).$$

Since parameters $\rho > 0$, $C > 0$ and $D \neq 0$ are unknown, this result cannot be applied directly to estimate $k$.

Among adaptive procedures for the automatic choice of $k$ one can mention the bootstrap methods (Hall, 1990; Danielsson et al., 1997; Caers and Van Dyck, 1999), which minimize the asymptotic MSE of the EVI, and the so-called sequential procedure (Drees and Kaufmann, 1998), based on the fact that the maximal deviation of the statistic $\sqrt{i}(\hat{\gamma}^H(n,i)-\gamma)$, $2 \le i \le k$, is of order $(\log\log n)^{1/2}$, that is,

$$\max_{2\le i\le k_n} \sqrt{i}(\hat{\gamma}^H(n,i)-\gamma-b_{n,i}) = O((\log\log n)^{1/2})$$

in probability, for all intermediate sequences $k_n$, where $b_{n,i} \in R$ are Hill estimator bias terms (Mason and Turova, 1994).

### Bootstrap method for selection of *k*

The number $k$ of retained data that are fitted to the tail corresponds to the minimum of the mean squared error (MSE),

$$\text{MSE}(\hat{\gamma}) = E\,(\hat{\gamma}-\gamma)^2 = \text{bias}^2(\hat{\gamma}) + \text{variance}(\hat{\gamma}) \to \min_k .$$

Here the bias is given by

$$b(n,k) = E\widehat{\gamma}^H(n,k) - \gamma,$$

and the variance is determined by

$$\text{var}(n,k) = E\left(\widehat{\gamma}^H(n,k) - E\widehat{\gamma}^H(n,k)\right)^2 .$$

We assume that Hill's estimate $\widehat{\gamma}^H(n,k)$ is used as $\hat{\gamma}$.

Since $\gamma$ is unknown and MSE cannot be evaluated, the bootstrap approach proposes replacing $\gamma$ in the MSE by an average calculated over some amount of resamples. These resamples are drawn from the initial sample $X^n$ randomly with replacement. This implies that some observations from $X^n$ will be represented in a resample with repetitions and others will not be represented at all.

As a result, in order to estimate $k$ one takes the value that minimizes a bootstrap empirical estimate of the MSE. More precisely, the bootstrap estimate of the bias is given by

$$b^*(n_1,k_1) = E\{\widehat{\gamma}^{*H}(n_1,k_1)|X^n\} - \widehat{\gamma}^H(n,k),$$

and the bootstrap estimate of the variance is determined by

$$\text{var}^*(n_1,k_1) = E\left\{\left(\widehat{\gamma}^{*H}(n_1,k_1) - E\{\widehat{\gamma}^{*H}(n_1,k_1)|X^n\}\right)^2 |X^n\right\}.$$

To construct these estimates, a smaller sample size $n_1 \le n$ is used and

$$\widehat{\gamma}^{*H}\,(n_1,k_1) = \frac{1}{k_1}\sum_{i=1}^{k_1} \log X^*_{(n_1-i+1)} - \log X^*_{(n_1-k_1)}$$

is Hill's estimate of $\gamma$. It is determined by the resample $X_*^{n_1} = \{X_1^*,\ldots,X_{n_1}^*\}$ drawn randomly from $X^n$ with replacement, where $X^*_{(1)} \le \ldots \le X^*_{(n_1)}$ are the order statistics of the sample $X_*^{n_1}$. In the bootstrap estimates considered $X^n$ is fixed and

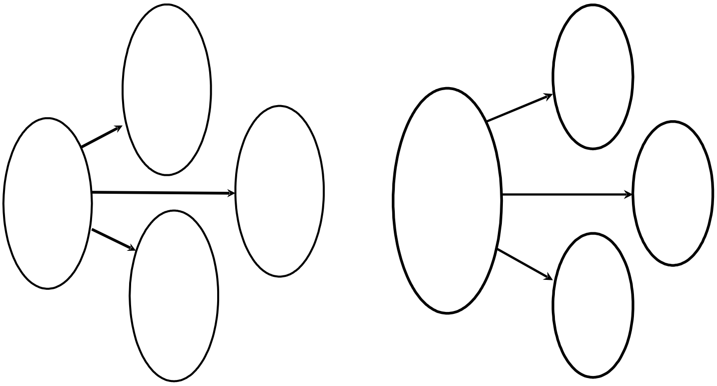

**Figure 1.3** Classical bootstrap: resamples of the same size $n$ as the sample $X^n$ are used (left). Nonclassical bootstrap: resamples of smaller size $n_1 = n^\beta$, $0 < \beta < 1$, than $n$ are used (right).

the expectation is calculated among all theoretically possible resamples $X_*^{n_1}$. In practice, the expectation is replaced by the average over the underlying resamples.

The reason for using smaller resamples is that the classical bootstrap with resamples of the same size $n$ as the initial sample leads to underestimates of the bias. Using a smaller sample size $n_1 \le n$ and $k_1$ data may help to avoid the situation where the bootstrap estimate of the bias is equal to zero regardless of the true bias of the estimate (Figure 1.3). Such situations arise particularly when linear estimates such as linear regressions or kernel estimates are used (Hall, 1990).[3]

**Example 2 (Hall, 1990)** *Suppose* $\hat{\vartheta}$ *is a linear function* $\hat{\vartheta} = \sum_{i=1}^n \varphi(X_i)$ *of data* $X_1, \ldots, X_n$, *and* $\vartheta^* = \sum_{i=1}^n \varphi(X_i^*)$ *is the same function constructed from the resample* $X_1^*, \ldots, X_n^*$. *Then* $E\{\vartheta^*|X^n\} = nE\{\varphi(X_i^*)|X^n\} = n\sum_{i=1}^n n^{-1}\varphi(X_i) = \hat{\vartheta}$, *since* $X_i^*$ *may be selected in n ways from* $X^n$. *This implies that the bias of the bootstrap estimate is* $\text{bias}^* = E\{\vartheta^*|X^n\} - \hat{\vartheta} = 0$, *but the bias of* $\hat{\vartheta}$ *is* $E\{\hat{\vartheta}\} - \vartheta \ne 0$. *Note that* $\text{bias}^*$ *is random. Hence, it is not a bias in the usual sense.*

[3] It seems that the problems with the classical bootstrap are even greater. It is proved in Bickel and Sakov (2002) that the statistic

$$a_n(F_n)\,(\max(X_1^*, \ldots, X_n^*) - b_n(F_n))$$

(where $a_n, b_n$ are normalyzed constants, see (1.1)) does not converge to $H_\gamma(x)$ for the bootstrap with resamples of size $n$. If resamples of smaller size $n_1 < n$ are used, $n_1 \to \infty$, $n_1/n \to 0$ and von Mises' condition

$$x\frac{f(x)}{1-F(x)} \to_{x\to\infty} \frac{1}{\gamma}$$

is satisfied, then

$$a_{n_1}(F_n)\left(\max(X_1^*, \ldots, X_{n_1}^*) - b_{n_1}(F_n)\right) \to H_\gamma(x).$$