Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice

Natalia Markovich

WILEY SERIES IN PROBABILITY AND STATISTICS
Nonparametric Analysis of Univariate Heavy-Tailed Data
Nonparametric Analysis of Univariate Heavy-Tailed Data

Research and Practice

Natalia Markovich
Institute of Control Sciences,
Russian Academy of Sciences,
Moscow, Russia
To my parents and daughter
# Contents

<table>
<thead>
<tr>
<th>Preface</th>
<th>xi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Definitions and rough detection of tail heaviness</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Definitions and basic properties of classes of heavy-tailed distributions</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Tail index estimation</td>
<td>6</td>
</tr>
<tr>
<td>1.2.1 Estimators of a positive-valued tail index</td>
<td>6</td>
</tr>
<tr>
<td>1.2.2 The choice of $k$ in Hill’s estimator</td>
<td>8</td>
</tr>
<tr>
<td>1.2.3 Estimators of a real-valued tail index</td>
<td>13</td>
</tr>
<tr>
<td>1.2.4 On-line estimation of the tail index</td>
<td>17</td>
</tr>
<tr>
<td>1.3 Detection of tail heaviness and dependence</td>
<td>27</td>
</tr>
<tr>
<td>1.3.1 Rough tests of tail heaviness</td>
<td>27</td>
</tr>
<tr>
<td>1.3.2 Analysis of Web traffic and TCP flow data</td>
<td>30</td>
</tr>
<tr>
<td>1.3.3 Dependence detection from univariate data</td>
<td>42</td>
</tr>
<tr>
<td>1.3.4 Dependence detection from bivariate data</td>
<td>49</td>
</tr>
<tr>
<td>1.3.5 Bivariate analysis of TCP flow data</td>
<td>51</td>
</tr>
<tr>
<td>1.4 Notes and comments</td>
<td>56</td>
</tr>
<tr>
<td>1.5 Exercises</td>
<td>57</td>
</tr>
</tbody>
</table>

| 2 Classical methods of probability density estimation | 61 |
| 2.1 Principles of density estimation | 61 |
| 2.2 Methods of density estimation | 70 |
| 2.2.1 Kernel estimators | 70 |
| 2.2.2 Projection estimators | 74 |
| 2.2.3 Spline estimators | 76 |
| 2.2.4 Smoothing methods | 76 |
| 2.2.5 Illustrative examples | 83 |
| 2.3 Kernel estimation from dependent data | 85 |
| 2.3.1 Statement of the problem | 86 |
| 2.3.2 Numerical calculation of the bandwidth | 89 |
| 2.3.3 Data-driven selection of the bandwidth | 91 |
| 2.4 Applications | 91 |
| 2.4.1 Finance: evaluation of market risk | 91 |
## CONTENTS

2.4.2 Telecommunications .......................................................... 93  
2.4.3 Population analysis ......................................................... 94  
2.5 Exercises ................................................................. 95  

3 Heavy-tailed density estimation ............................................. 99  
3.1 Problems of the estimation of heavy-tailed densities ............... 100  
3.2 Combined parametric–nonparametric method ....................... 101  
  3.2.1 Nonparametric estimation of the density by structural risk minimization ................................................. 103  
  3.2.2 Illustrative examples .................................................... 107  
  3.2.3 Web data analysis by a combined parametric–nonparametric method ......................................................... 109  
3.3 Barron’s estimator and $\chi^2$-optimality .............................. 111  
3.4 Kernel estimators with variable bandwidth .......................... 113  
3.5 Retransformed nonparametric estimators ........................... 117  
3.6 Exercises ................................................................. 119  

4 Transformations and heavy-tailed density estimation ............... 123  
4.1 Problems of data transformations ....................................... 123  
4.2 Estimates based on a fixed transformation ............................ 124  
4.3 Estimates based on an adaptive transformation .................... 128  
  4.3.1 Estimation algorithm .................................................. 128  
  4.3.2 Analysis of the algorithm .......................................... 129  
  4.3.3 Further remarks ...................................................... 133  
4.4 Estimating the accuracy of retransformed estimates .............. 135  
4.5 Boundary kernels .......................................................... 136  
4.6 Accuracy of a nonvariable bandwidth kernel estimator .......... 139  
4.7 The $D$ method for a nonvariable bandwidth kernel estimator . 141  
4.8 The $D$ method for a variable bandwidth kernel estimator ...... 142  
  4.8.1 Method and results .................................................. 142  
  4.8.2 Application to Web traffic characteristics ...................... 144  
4.9 The $\omega^2$ method for the projection estimator ................. 147  
4.10 Exercises ................................................................. 149  

5 Classification and retransformed density estimates ............... 151  
5.1 Classification and quality of density estimation .................. 151  
5.2 Convergence of the estimated probability of misclassification .. 154  
5.3 Simulation study .......................................................... 155  
5.4 Application of the classification technique to Web data analysis 160  
  5.4.1 Intelligent browser .................................................. 160  
  5.4.2 Web data analysis by traffic classification ...................... 161  
  5.4.3 Web prefetching ...................................................... 161  
5.5 Exercises ................................................................. 161
## 6 Estimation of high quantiles

6.1 Introduction ................................................. 163
6.2 Estimators of high quantiles .................................. 164
6.3 Distribution of high quantile estimates ......................... 167
6.4 Simulation study ............................................. 169
   6.4.1 Comparison of high quantile estimates in terms of relative
         bias and mean squared error ........................... 169
   6.4.2 Comparison of high quantile estimates in terms of
         confidence intervals ..................................... 170
6.5 Application to Web traffic data .................................. 175
6.6 Exercises .................................................... 176

## 7 Nonparametric estimation of the hazard rate function

7.1 Definition of the hazard rate function .......................... 180
7.2 Statistical regularization method .................................. 182
7.3 Numerical solution of ill-posed problems ....................... 185
7.4 Estimation of the hazard rate function of heavy-tailed
    distributions ................................................. 187
7.5 Hazard rate estimation for compactly supported distributions . 188
   7.5.1 Estimation of the hazard rate from the simplest
         equations ................................................. 188
   7.5.2 Estimation of the hazard rate from a special kernel
         equation .................................................. 193
7.6 Estimation of the ratio of hazard rates ........................ 197
   7.6.1 Failure time detection ................................ 199
   7.6.2 Hormesis detection .................................... 200
7.7 Hazard rate estimation in teletraffic theory ..................... 207
   7.7.1 Teletraffic processes at the packet level .................. 207
   7.7.2 Estimation of the intensity of a nonhomogeneous Poisson
         process .................................................. 208
7.8 Semi-Markov modeling in teletraffic engineering ............... 210
   7.8.1 The Gilbert–Elliott model .............................. 210
   7.8.2 Estimation of a retrial process .......................... 212
7.9 Exercises .................................................... 217

## 8 Nonparametric estimation of the renewal function

8.1 Traffic modeling by recurrent marked point processes .......... 220
8.2 Introduction to renewal function estimation ..................... 221
8.3 Histogram-type estimator of the renewal function ............... 224
8.4 Convergence of the histogram-type estimator ..................... 225
8.5 Selection of $k$ by a bootstrap method ........................ 228
8.6 Selection of $k$ by a plot .................................... 232
8.7 Simulation study ................................................ 234
8.8 Application to the inter-arrival times of TCP connections ...... 245
## CONTENTS

8.9  Conclusions and discussion ............................................. 247  
8.10 Exercises ................................................................. 248  

### Appendices

<table>
<thead>
<tr>
<th>Appendix</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Proofs of Chapter 2</td>
<td>251</td>
</tr>
<tr>
<td>B</td>
<td>Proofs of Chapter 4</td>
<td>253</td>
</tr>
<tr>
<td>C</td>
<td>Proofs of Chapter 5</td>
<td>267</td>
</tr>
<tr>
<td>D</td>
<td>Proofs of Chapter 6</td>
<td>271</td>
</tr>
<tr>
<td>E</td>
<td>Proofs of Chapter 7</td>
<td>275</td>
</tr>
<tr>
<td>F</td>
<td>Proofs of Chapter 8</td>
<td>285</td>
</tr>
<tr>
<td></td>
<td>List of Main Symbols and Abbreviations</td>
<td>291</td>
</tr>
<tr>
<td></td>
<td>References</td>
<td>295</td>
</tr>
<tr>
<td></td>
<td>Index</td>
<td>307</td>
</tr>
</tbody>
</table>
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) \( G(x) = \exp\left\{ - \left( 1 + \gamma(x - \mu)/\sigma \right)^{-1/\gamma} \right\} \). In the POT method the values which are larger than some thresholds are modelled by the generalized Pareto distribution (GPD) with DF \( \Psi_{\gamma, \sigma}(x) = 1 - (1 + \gamma x/\sigma)^{-1/\gamma} \). 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 Markovich 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 greatly 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 \(R\).
2 DEFINITIONS AND ROUGH DETECTION OF TAIL HEAVINESS

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

Definition 3 The function $F(x) = P[\omega : X(\omega) \leq x]$, $x \in R$, 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 $X_1, X_2, \ldots, X_n$ ($X_i \in B_i \subseteq R$, $B_i$ is a finite set) are called independent if, for any $x_1, x_2, \ldots, x_n \in R$,

$$P[X_1 = x_1, \ldots, X_n = x_n] = P[X_1 = x_1] \cdots P[X_n = x_n]$$

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

$$P[X_1 \in B_1, \ldots, X_n \in B_n] = P[X_1 \in B_1] \cdots P[X_n \in B_n].$$

In terms of DFs and PDFs, independence means that

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

and

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

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 $X^n = \{X_1, \ldots, X_n\}$ be a sample of independent and identically distributed (i.i.d.) r.v.s with DF $F(x) = P[X_1 \leq x]$ and $M_n = \max(X_1, X_2, \ldots, X_n)$. 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

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

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

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

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

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

---

1 This result remains true if $X_1, \ldots, X_n$ 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 $x > 0$, 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 $X \in \text{MDA}(H_\gamma)$ ($F \in \text{MDA}(H_\gamma)$).

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

$$P[S_n > x] \sim nP[X_1 > x] \sim P[M_n > x] \quad \text{as } x \to \infty,$$

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 $\lim_{x \to x_1} f(x)/g(x) = 1$. 
DEFINITIONS AND ROUGH DETECTION OF TAIL HEAVINESS

**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.

<table>
<thead>
<tr>
<th>Heavy-tailed distributions</th>
<th>Subexponential:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pareto, lognormal, Weibull with shape parameter less than 1</td>
</tr>
<tr>
<td>With regularly varying tails:</td>
<td>Pareto, Cauchy, Burr, Fréchet, Zipf–Mandelbrot law</td>
</tr>
</tbody>
</table>

| 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 $\alpha = 1/\gamma$, $\gamma > 0$ ($X \in \mathbb{R}_{-1/\gamma}$), if

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

where $\ell(x)$ is called a slowly varying function ($\ell(x) \in \mathbb{R}_0$).

**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 summarized 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 $X' \in R_{1-\alpha}$ and $P\{X' > x\} \sim \ell(x)x^{1-\alpha}/((\alpha - 1)E[X])$ 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

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

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

$$F(x+y)/F(x) = \exp(-\lambda y), \quad 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 $\bar{F}(x)$ 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\{X_i - u > x|X_i > u\} \to (1 + \gamma x)^{-1/\gamma}, \quad 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
6 DEFINITIONS AND ROUGH DETECTION OF TAIL HEAVINESS

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

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

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

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

\[k \to \infty, \quad k/n \to 0 \quad \text{as} \quad n \to \infty \tag{1.6}\]

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

\[\sqrt{k} (\hat{\gamma}^H (n, k) - \gamma) \to^d N(0, \gamma^2)\]

(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
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

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

(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 Embrechts 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

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

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

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

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
DEFINITIONS AND ROUGH DETECTION OF TAIL HEAVINESS

procedure (Drees and Kaufmann, 1998), based on the fact that the maximal deviation of the statistic $\sqrt{i} (\hat{\gamma}^H(n, i) - \gamma)$, $2 \leq i \leq k$, is of order $(\log \log n)^{1/2}$, that is,

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

in probability, for all intermediate sequences $k_n$, where $b_{n,i} \in \mathbb{R}$ 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),

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

Here the bias is given by

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

and the variance is determined by

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

We assume that Hill’s estimate $\hat{\gamma}^H(n, k)$ 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

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

and the bootstrap estimate of the variance is determined by

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

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

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

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 $X^{*}_{(1)} \leq \ldots \leq X^{*}_{(n)}$ 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 \( n_1 \leq n \) 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).

Example 2 (Hall, 1990) Suppose \( \hat{\theta} \) is a linear function \( \hat{\theta} = \sum_{i=1}^{n} \varphi(X_i) \) of data \( X_1, \ldots, X_n \), and \( \hat{\theta}^* = \sum_{i=1}^{n} \varphi(X^*_i) \) is the same function constructed from the resample \( X^*_1, \ldots, X^*_n \). Then \( E\{\hat{\theta}^*|X^n\} = nE\{\varphi(X^*_i)|X^n\} = n \sum_{i=1}^{n} n^{-1} \varphi(X_i) = \hat{\theta} \), since \( X^*_i \) may be selected in \( n \) ways from \( X^n \). This implies that the bias of the bootstrap estimate is \( \text{bias}^* = E\{\hat{\theta}^*|X^n\} - \hat{\theta} = 0 \), but the bias of \( \hat{\theta} \) is \( E\{\hat{\theta}\} - \hat{\theta} \neq 0 \). Note that \( \text{bias}^* \) is random. Hence, it is not a bias in the usual sense.

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 normalized 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, \( n_1 \rightarrow \infty, n_1/n \rightarrow 0 \) and von Mises’ condition

\[
\frac{f(x)}{1 - F(x)} \rightarrow_{x \rightarrow \infty} \frac{1}{\gamma}
\]

is satisfied, then

\[
a_{n_1}(F_n)(\max(X^*_1, \ldots, X^*_{n_1}) - b_{n_1}(F_n)) \rightarrow H_\gamma(x).
\]
It was proved in Hall (1990) that if the tail satisfies
\[ 1 - F(x) = C_0 x^{-1/\gamma} + C_1 x^{-2/\gamma} + o(x^{-2/\gamma}) \]  
(1.9)
as \( x \to \infty, \gamma > 0, \) and \( k_1 \sim n_1^{2/3} \) then
\[ \text{MSE}(n_1, k_1) = E[(\hat{\gamma}^H(n_1, k_1) - \gamma)^2] \]
and its bootstrap estimate
\[ \widehat{\text{MSE}}(n_1, k_1) = E\{(\hat{\gamma}^H(n_1, k_1) - \hat{\gamma}^H(n, k))^2 | X^n\} \]
\[ = E(\hat{\gamma}^{sH}(n_1, k_1)^2 | X^n) - 2\hat{\gamma}^H(n, k)E[\hat{\gamma}^{sH}(n_1, k_1) | X^n] + (\hat{\gamma}^H(n, k))^2 \]
are close. The values of \( k_1 \) and \( k \) are related by:
\[ k = k_1 \left( \frac{n}{n_1} \right)^\alpha, \quad 0 < \alpha < 1. \]  
(1.10)
The value \( n_1 \) is chosen as
\[ n_1 = n^\beta. \]  
(1.11)
The optimal value of \( k_1 \) and, by (1.10), the optimal value of \( k \) are found by choosing that value which minimizes the estimated mean squared error \( \widehat{\text{MSE}}(n_1, k_1). \) Since the DF \( F(x) \) is unknown, one can estimate instead of the bias \( b^*(n_1, k_1) \) and variance \( \text{var}^*(n_1, k_1) \) their empirical bootstrap estimates
\[ \hat{b}^*(n_1, k_1) = \frac{1}{B} \sum_{b=1}^{B} \hat{\gamma}^{sH}_b(n_1, k_1) - \hat{\gamma}^H(n, k) \]
and
\[ \hat{\text{var}}^*(n_1, k_1) = \frac{1}{B-1} \sum_{b=1}^{B} \left( \hat{\gamma}^{sH}_b(n_1, k_1) - \frac{1}{B} \sum_{b=1}^{B} \hat{\gamma}^{sH}_b(n_1, k_1) \right)^2, \]
respectively (Efron and Tibshirani, 1993). Here \( B \) denotes the total number of \( n_1 \)-sized bootstrapped resamples, and \( \hat{\gamma}^{sH}_b(n_1, k_1) \) is Hill’s estimate derived from one of the resamples.

Caers and Van Dyck (1999) recommended finding \( k_1 \) by minimizing
\[ \text{MSE}^*(n_1, k_1) = (\hat{b}^*(n_1, k_1))^2 + \hat{\text{var}}^*(n_1, k_1). \]  
(1.12)
Here all possible values of \( k_1 \), where \( k_1 \) is an integer in the interval \([1, n_1]\), are examined. Hall (1990) recommended taking \( \alpha = \frac{2}{5}, \beta = \frac{1}{2} \) for Pareto-type distributions. Hence, the values of \( k \) satisfy (1.6). Caers and Van Dyck (1999) investigated these values of \( \alpha \) and \( \beta \) using Monte Carlo simulations for a variety of distributions and found that they lead to the best results for the MSE. From our experience they provide good estimates of \( \gamma \) if the tails are sufficiently heavy, that is, the EVI should be large enough.
DEFINITIONS AND ROUGH DETECTION OF TAIL HEAVINESS

Double bootstrap method for the selection of \( k \)

The double bootstrap, proposed in Danielsson et al. (1997), improves the bootstrap method (Hall, 1990), since it requires fewer parameters: one has to select \( n_1 \) and \( B \), but \( \alpha \) in (1.10) is not required.

Instead of estimation of the MSE we use the auxiliary statistic

\[
MSE(z_{n,k}) = E \left( z_{n,k} - z_{n,k}^* \right)^2,
\]

where

\[
z_{n,k} = M_{n,k} - 2(\hat{\gamma}^H(n, k))^2, \quad M_{n,k} = \frac{1}{k} \sum_{j=1}^{k} \left( \log X_{n-j+1} - \log X_{n-k} \right)^2,
\]

and \( z_{n,k}^* \) is a bootstrap estimate of \( z_{n,k} \). Since \( M_{n,k}/2\hat{\gamma}^H(n, k) \) and \( \hat{\gamma}^H(n, k) \) are consistent estimates of \( \gamma \), \( z_{n,k} \to 0 \) as \( n \to \infty \). Therefore, the asymptotic MSE of \( z_{n,k} \) is defined by

\[
AMSE(z_{n,k}) = \frac{E(z_{n,k}^2)}{\hat{\gamma}^H(n, k)} \to \min_k.
\]

The double bootstrap algorithm is as follows:

- Draw \( B \) bootstrap subsamples of size \( n_1 \in (\sqrt{n}, n) \) (e.g., \( n_1 \sim n^{3/4} \)) from the original sample and determine the value \( \hat{k}^*_{n_1} \) that minimizes the MSE of \( z_{n_1,k} \).
- Repeat this for \( B \) subsamples of size \( n_2 = [n_1^2/n] \) (\( [x] \) is the integer part of \( x \)) and determine the value \( \hat{k}^*_{n_2} \) that minimizes the MSE of \( z_{n_2,k} \).
- Calculate \( \hat{k}^{opt}_n \) by the formula

\[
\hat{k}^{opt}_n = \left[ \frac{(\hat{k}^*_{n_1})^2}{\hat{k}^*_{n_2}} \left( 1 - \frac{1}{\hat{\rho}_1} \right)^{\hat{\rho}_1^{-1}} \right], \quad \hat{\rho}_1 = \frac{\log \hat{k}^*_{n_1}}{2 \log(\hat{k}^*_{n_1}/n_1)},
\]

and estimate \( \gamma \) by Hill’s estimate with \( \hat{k}^{opt}_n \).

The method is robust with respect to the choice of \( n_1 \) (Gomes and Oliveira, 2000).

Sequential procedure for the selection of \( k \)

Drees and Kaufmann (1998) provide the following algorithm for this procedure:

- Obtain an initial estimate \( \hat{\gamma}_0 = \hat{\gamma}^H(n, 2\sqrt{n}) \) for the parameter \( \gamma \) by Hill’s estimate.
- For \( r_n = 2.5\hat{\gamma}_0 n^{0.25} \), compute

\[
\hat{k}(r_n) = \min\{k \in \{2, \ldots, n-1\} | \max_{2 \leq i \leq k} \sqrt{i}(\hat{\gamma}^H(n, i) - \hat{\gamma}^H(n, k)) > r_n \}.
\]
If \( r_n \) is too large and \( \max_{2 \leq i \leq k} \sqrt{i(\hat{\gamma}^H(n, i) - \hat{\gamma}^H(n, k))} > r_n \) is not satisfied, it is recommended to repeatedly replace \( r_n \) by \( 0.9r_n \) until \( \hat{k}(r_n) \) is well defined.

- Similarly, compute \( \hat{k}(r_n^e) \) for \( \varepsilon = 0.7 \).
- Calculate

\[
\hat{k}_{\text{opt}} = \frac{1}{3} \left( \frac{\hat{k}(r_n^e)}{(\hat{k}(r_n))^e} \right)^{1/(1-\varepsilon)} (2\hat{\gamma}_0)^{1/3}
\]

and estimate \( \gamma \) by \( \hat{\gamma}^H(n, \hat{k}_{\text{opt}}) \).

The method is sensitive to the choice of \( r_n \).

1.2.3 Estimators of a real-valued tail index

Among nonparametric estimators, we mention the moment estimators (Dekkers et al., 1989) and UH estimators (Berliinet et al., 1998) that are not restricted to positive \( \gamma \). They can be rewritten in terms of Hill’s estimator. The moment estimator is given by

\[
\hat{\gamma}^M(n, k) = \hat{\gamma}^H(n, k) + 1 - 0.5 \left( 1 - (\hat{\gamma}^H(n, k))^2 / S_{n,k} \right)^{-1},
\]

where \( S_{n,k} = (1/k) \sum_{i=1}^{k} \left( \log X_{(n-i+1)} - \log X_{(n-k)} \right)^2 \). The UH estimator is determined by

\[
\hat{\gamma}^{\text{UH}}(n, k) = \frac{1}{k} \sum_{i=1}^{k} \log UH_i - \log UH_{k+1},
\]

where \( UH_i = X_{(n-i)} \hat{\gamma}^H(n, i) \).

The kernel estimator of \( \gamma \),

\[
\hat{\gamma}^K(n, \lambda) = \frac{\sum_{i=1}^{n} (i/(n\lambda))K(i/(n\lambda)) \left( \log^+ X_{(n-i+1)} - \log^+ X_{(n-i)} \right)}{1/(n\lambda) \sum_{i=1}^{n} K(i/(n\lambda))},
\]

where \( \lambda \) is a bandwidth parameter, \( \log^+ x = \log(x \vee 1) \) and \( K \) is some nonnegative, nonincreasing kernel defined on \((0, \infty)\) and integrating to 1, was proposed in Csörgő et al. (1985). It is assumed that \( X_{(0)} = 1 \). This estimator generalizes Hill’s estimator because the latter estimator may be obtained from the kernel estimator by the selection of the indicator function on the interval \((0, 1)\) as a kernel, that is, \( K(u) = 1(0 < u < 1) \), and \( \lambda = k/n \).

The Pickands estimator

\[
\hat{\gamma}^P(n, k) = \frac{1}{\log 2} \log \frac{X_{(n-k+1)} - X_{(n-2k+1)}}{X_{(n-2k+1)} - X_{(n-4k+1)}}, \quad \text{for } k \leq n/4,
\]

has the same asymptotic properties (weak and strong consistency, asymptotic normality) as Hill’s estimator (Embrechts et al., 1997).
The estimator based on an exponential regression model essentially uses the approximate representation (in distribution, denoted by \( \approx \))

\[
j \log \frac{X_{(a-j+1)} - X_{(a-k)}}{X_{(n-j)} - X_{(n-k)}} \approx D \frac{\gamma}{1 - (j/(k+1))^\gamma} E_j, \quad j = 1, \ldots, k - 1,
\]

where \( E_j, j = 1, \ldots, n \), denote standard exponential r.v.s (Beirlant et al., 2004), from which \( \gamma \) can be estimated by the ML method (Section 2.1). The resulting estimator is invariant with respect to a shift and a rescaling of the data.

The POT method

In the peaks-over-threshold (POT) method a generalized Pareto distribution (GPD) \( \Psi_{\sigma, \gamma}(x) = \{ 
\begin{align*}
1 - (1 + \gamma x/\sigma)^{-1/\gamma}, & \quad \gamma \neq 0, \\
1 - \exp(-x/\sigma), & \quad \gamma = 0,
\end{align*}
\)

where \( \sigma > 0 \) and \( x \geq 0 \), as \( \gamma \geq 0; 0 \leq x \leq -\sigma/\gamma \), as \( \gamma < 0 \), is fitted to excesses over a high threshold. The method is based on the limit law for excess distributions (Balkema and de Haan, 1974; Pickands, 1975). Denote by

\[
F_u(x) = P[X - u \leq x | X > u]
\]

the conditional distribution of the excess of \( X \) over the threshold \( u \), given that \( u \) is exceeded. Pickands’ (1975) result states that condition (1.1) holds \((F \in \text{MDA}(H_\gamma))\) if and only if

\[
\lim_{u \to x_+} \sup_{0 < x < x_+ - u} |F_u(x) - \Psi_{\sigma, \gamma}(x)| = 0
\]

for some positive scaling function \( \sigma(u) \) depending on \( u \). Here \( x_+ \in (0, \infty) \) is the right endpoint of the distribution \( F \).

Thus, if one fixes \( u \) and selects from a sample \( X_1, \ldots, X_n \) only those observations \( X_{i_1}, \ldots, X_{i_{N_u}} \) that exceed \( u \), a GPD with parameters \( \gamma \) and \( \sigma = \sigma(u) \) is likely to be a good approximation for the distribution \( F_u \) of the \( N_u \) excesses \( Y_j = X_{i_j} - u \).

This approach allows one to estimate the EVI \( \gamma \), the excess distribution, and the unconditional tail \( \hat{F}(x) = \hat{F}(u) \hat{F}_u(x-u) \) by

\[
\hat{F}(x) = \frac{N_u}{n} \left( 1 + \hat{\gamma}_u \frac{x-u}{\hat{\sigma}_u} \right)^{-1/\hat{\gamma}_u}, \quad u < x < x_+,
\]

where \( \hat{F}(u) \) is estimated by the empirical exceedance probability \( N_u/n \). Inverting (1.18) then yields the POT estimator (6.4) for high quantiles above the threshold \( u \).

\footnote{It follows from the Rényi representation (D.6); see Appendix D.}
The parameters $\gamma$ and $\sigma$ may be computed in different ways, namely by the ML method (Smith, 1987), the method of moments (MOM), the method of probability-weighted moments (PWM) (Hosking and Wallis, 1987), the elemental percentile method (EPM) (Castillo et al., 2006), or Bayesian methods (Coles, 2001). For details, see Beirlant et al. (2004).

Smith (1987) describes the ML techniques and shows that the ML estimators for $\gamma$ and $\sigma$ are asymptotically normal if $\gamma > -1/2$.

Hosking and Wallis (1987) derive a simple moments-based method to estimate $\gamma$ and $\sigma$, but this only works if $\gamma < 1/2$. They also apply the PWM and find that the corresponding EVI estimator is a good alternative to the ML estimator for $\gamma < 1$.

The EPM does not impose any restrictions on the EVI $\gamma$. A simulation study shows that the ML method mostly provides the best estimators if $\gamma$ is estimated to be positive, whereas the EPM is to be preferred if $\gamma$ is estimated to be lower than 0 (Matthys and Beirlant, 2001).

The choice of the threshold $u$ resembles the choice of the value of $k$ for Hill’s estimator. Similarly, one can use the exceedance plot (see Section 1.2.2; see also Beirlant et al., 2004). Since the mean excess function of the GPD distribution is linear, i.e.

$$e(u) = \frac{\sigma + \gamma u}{1 - \gamma}, \quad \text{if } \gamma < 1,$$

one can choose $u = X_{n-k}$ as the point to the right of which a linear pattern appears in the plot $(u, e(u))$, $k = 1, \ldots, n-1$.

**Comparison of methods**

It is difficult to compare the estimators of $\gamma$. One can only look at the asymptotic variances and biases of estimates for known distributions with known parameters.

For Pareto tails the moment estimator is unbiased for any $\gamma$, since $S_{n,k} \approx \gamma^2$, but the variance of this estimate is larger than the variance of Hill’s estimate. Besides, it is known that

$$\sqrt{k} \left( \hat{\gamma}^M(n, k) - \gamma \right) \to_d \begin{cases} N(0, 1 + \gamma^2), & \gamma \geq 0, \\ N\left(0, (1 - \gamma)^2(1 - 2\gamma) \left(4 - 8\frac{1-2\gamma}{1-3\gamma} + \frac{5-11\gamma(1-2\gamma)}{(1-3\gamma)(1-4\gamma)} \right)\right), & \gamma < 0. \end{cases}$$

The UH estimator has larger asymptotical variance for $\gamma > 0$ than the Hill and moment estimators. For $\gamma < 0$ the UH estimator is more efficient than the moment estimator:

$$\sqrt{k} \left( \hat{\gamma}^{UH}(n, k) - \gamma \right) \to_d \begin{cases} N(0, (1 + \gamma)^2), & \gamma \geq 0, \\ N\left(0, \frac{(1 - \gamma)(1 + \gamma + 2\gamma^2)}{1 - 2\gamma} \right), & \gamma < 0, \end{cases}$$

(Caers and Van Dyck, 1999).
Under the conditions on $F(x)$ ($F \in \text{MDA}(H_\gamma)$) and $U(x) = F^{-1}(1 - x^{-1})$, the Pickands estimator has the following property (see Dekkers and de Haan, 1989, p. 1799): 

$$\sqrt{k} \left( \hat{\gamma}^P(n, k) - \gamma \right) \rightarrow^d N \left( 0, \frac{\gamma^2 (2^{2\gamma+1} + 1)}{(2(2\gamma-1) \log 2)^2} \right), \quad n \rightarrow \infty.$$ 

For the estimator based on an exponential regression model, we have that 

$$\sqrt{k} \left( \hat{\gamma}^{RMA}(n, k) - \gamma \right) \rightarrow^d N \left( 0, \frac{\sigma_\gamma^2}{a_\gamma^2} \right)$$ 

under some conditions on $U(x)$ and if we suppose that $k, n \rightarrow \infty$ with $k/n \rightarrow 0$ (see Matthys and Beirlant, 2003), where 

$$a_\gamma = \frac{1}{\gamma^2} \int_0^1 \left( \frac{1-u^\gamma + u^\gamma \log u^\gamma}{1-u^\gamma} \right)^2 du$$ 

and $\sigma_\gamma^2$ equals the variance of $K_{\gamma}(U)$ with $U$ uniformly distributed on $(0,1)$ and 

$$K_{\gamma}(U) = \frac{1}{\gamma} \log u + \frac{1+\gamma}{\gamma^2} \text{dilog}(u^\gamma),$$ 

where 

$$\text{dilog}(u) = \int_1^u \frac{\log t}{1-t} dt, \quad u \geq 0,$$ 

denotes the dilogarithm function.

For the POT ML estimator, 

$$\sqrt{N_u} \left( \hat{\gamma}^{MLP}_u - \gamma \right) \rightarrow^d N(0, (1+\gamma)^2)$$ 

for $N_u \rightarrow \infty$ provided $\gamma > -1/2$, under the assumption that the excesses exactly follow a GPD (Smith, 1987). The asymptotic variances of the POT, PWM, EPM, and MOM estimators are given in Beirlant et al. (2004). Figure 1.4 compares the asymptotic variance for different estimators.

The POT, Pickands, moment, and $\gamma^{RMA}(n, k)$ estimators may suffer from substantial bias for some ill-behaving distributions like the loggamma, the lognormal and the inverse Burr distributions (Matthys and Beirlant, 2001). This bias is caused by violation of the necessary conditions on the tail required for the properties of normality considered.

Sometimes it is difficult to compare a given estimator with others, since this estimator estimates some function of $\gamma$, but not $\gamma$ itself; see, for instance, Davydov et al. (2000) or Section 1.2.4. In Csörgő et al. (1985) the asymptotic normality is

\footnote{\(F^{-1}(x)\) denotes an inverse function.}
Figure 1.4 Asymptotic variances of $\sqrt{k}(\hat{\gamma} - \gamma)$ for the Hill (solid line), moment (solid line with + marks), UH (solid line with circles), Pickands (dot-dashed line) and POT ML (dotted line) estimators. The variance of $\hat{\gamma}^{RMA}$ is not represented, since it depends on the generator of uniform r.v.s. According to Matthys and Beirlant (2001), it nearly coincides with the asymptotic variance of the POT ML estimator for positive $\gamma$ and it is lower than that of the moment estimator for negative $\gamma$. The UH and POT ML estimators coincide for $\gamma > 0$.

proved for a more complicated statistic than $\sqrt{k}(\hat{\gamma} - \gamma)$ and does not allow one to represent the asymptotic variance and bias in such simple forms as before.

It is mentioned in Polzehl and Spokoiny (2002) and Grama and Spokoiny (2003) that Hill’s estimator estimates not the tail index, but another parameter (see the trends in Figure 1.2) called the fitted Pareto index. This parameter is interpreted as the parameter of the Pareto distribution. In order to estimate this parameter, the authors propose a method based on successive testing of the hypothesis that the first $k$ normed log-spacings follow exponential distributions with homogeneous parameters.\(^6\) Then the number $k$ is chosen as the change-point detected.

1.2.4 On-line estimation of the tail index

In practice, on-line estimates are an important tool. At present several estimators of the tail index are known (see Sections 1.2.1 and 1.2.3). All these estimators are based on the order statistics and cannot be organized recursively. Here, the estimator that was proposed in Davydov et al. (2000) and investigated in Markovich (2005a) is considered. It is based on independent ratios of the second largest values to the

\[^6\] It is known that $X = \exp T + \varepsilon$ is Pareto distributed,

$$F(x) = 1 - (x - \varepsilon)^{-a}, \quad x \geq \varepsilon,$$

with $a = 1/\theta$, $a \geq 1$, $x > 1 + \varepsilon$, $\varepsilon \geq 0$ if $T$ is exponentially distributed with parameter $\theta$. 
largest values of subsets of observations. The recursive behavior of this estimate is discussed. The bootstrap method for the estimation of its parameters is presented.

**Group estimator ($\gamma > 0$)**

Let us consider a sample $X^n = \{X_1, \ldots, X_n\}$ of size $n$ taken from a heavy-tailed DF $F(x)$. We assume that $X_1, \ldots, X_n$ are i.i.d. r.v.s. The tail index estimator considered in Davydov et al. (2000) has the essential advantage that it can be calculated recursively. According to this estimator the sample is divided into $l$ groups $V_1, \ldots, V_l$, each group containing $m$ random variables, that is, $n = l \cdot m$. In practice, $m$ is chosen and then $l = \lceil n/m \rceil$, where $[a]$ denotes the integer part of a number $a > 0$. Let

$$M_{l1}^{(1)} = \max\{X_j : X_j \in V_i\}$$

and let $M_{li}^{(2)}$ denote the second largest element in the same group $V_i$. Let us denote

$$k_{li} = M_{li}^{(2)}/M_{li}^{(1)}, \quad z_i = (1/l) \sum_{i=1}^{l} k_{li}. \quad (1.19)$$

Let a DF $F(x)$ satisfy the following relation as $x \to \infty$:

$$1 - F(x) = x^{-\alpha} \ell(x), \quad (1.20)$$

where $\alpha > 0$, and $\ell$ is slowly varying, $\lim_{x \to \infty} \ell(tx)/\ell(x) = 1$. The distribution may satisfy the second-order asymptotic relation (as $x \to \infty$)

$$1 - F(x) = C_1 x^{-\alpha} + C_2 x^{-\beta} + o(x^{-\beta}), \quad (1.21)$$

with some parameters $0 < \alpha < \beta \leq \infty$, where $C_1, C_2$ are positive constants. Then Davydov et al. (2000) proved that, for a distribution satisfying (1.20) and $l = m = \lceil \sqrt{n} \rceil$, $z_l \to^{a.s.} \frac{\alpha}{\alpha + 1} = \frac{1}{1 + \gamma}, \quad (1.22)$

and if $F(x)$ satisfies (1.21) with $\beta = 2\alpha$ then

$$l \left( l^{-1} \sum_{i=1}^{l} k_{li} - \alpha (1+\alpha)^{-1} \right) \left( \sum_{i=1}^{l} k_{li}^2 - l^{-1} \sum_{i=1}^{l} k_{li} \right)^{-1/2} \to^p N(0, 1), \quad (1.23)$$

$$\sqrt{l} \left( l^{-1} \sum_{i=1}^{l} k_{li} - \alpha (1+\alpha)^{-1} \right) \to^p N(0, \sigma^2) \quad (1.24)$$

with $\sigma^2 = \alpha(\alpha + 1)^{-2}(\alpha + 2)^{-1}$ as $n \to \infty$. One can construct confidence intervals using (1.23). Paulauskas (2003) also proved that (1.24) is valid for $F$ satisfying

\[7\] The case $\beta = \infty$ corresponds to a Pareto distribution, and $\beta = 2\alpha$ to a stable distribution with $0 < \alpha < 2$.\]
DEFINITIONS AND ROUGH DETECTION OF TAIL HEAVINESS

(1.21) not just for equal \( l \) and \( m \), but if \( l = \varepsilon_n n^{2\zeta/(1+2\zeta)} \), \( m = \varepsilon^{-1}_n n^{1/(1+2\zeta)} \), where \( \varepsilon_n \to 0 \) as \( n \to \infty \) and \( \zeta = (\beta - \alpha)/\alpha \). All these results are asymptotic, and any implementation for moderate sample size would in practice require additional research. In particular, since the parameters \( \alpha \) and \( \beta \) are unknown, it is impossible to find \( l \) and \( m \) exactly.

By (1.22) it follows that \( 1/z_l - 1 \) can be used as an estimator of \( \gamma \). We will call this estimator the group estimator.

It is easy to see that (formula (7) in Paulauskas, 2003)

\[
\sqrt{l} \left( \frac{1}{l} \sum_{i=1}^{l} k_{li} - \frac{\alpha}{1+\alpha} \right) = \sqrt{l} \left( \frac{1}{l} \sum_{i=1}^{l} k_{li} - Ek_{li} + b_m \right) = \frac{1}{\sqrt{l}} \sum_{i=1}^{l} (k_{li} - Ek_{li}) + \sqrt{l}b_m,
\]

where \( b_m \) is the bias of the estimator \( z_l \):

\[
Ez_l = E \frac{1}{l} \sum_{i=1}^{l} k_{li} = Ek_{li} = \frac{\alpha}{1+\alpha} + b_m.
\]

From Paulauskas (2003) we have that the MSE is given by

\[
E \left( \frac{1}{l} \sum_{i=1}^{l} k_{li} - \frac{\alpha}{1+\alpha} \right)^2 = \text{bias}^2 \left( \frac{1}{l} \sum_{i=1}^{l} k_{li} \right) + \text{var} \left( \frac{1}{l} \sum_{i=1}^{l} k_{li} \right) = \frac{\sigma_i^2}{l} + b_m^2,
\]

where \( \sigma_i^2 = E(k_{li} - Ek_{li})^2 \) is the variance of \( k_{li} \). Then \( l \) and \( m \) are chosen in such a way that for the first term on the right-hand side of (1.25), the central limit theorem holds and the bias must stay bounded. In Paulauskas (2003) the upper bound for \( b_m \) is obtained and the optimal \( l \) is given up to positive constants as the minimum of the upper bound of the MSE (1.26), \( l = O(n^{2\zeta/(1+2\zeta)}) \). The latter result is proved only for the distribution class (1.21). Subsequently, we do not assume that the shape of the distribution is known. The parameters \( l \) and \( m \) are selected by the bootstrap method, which is a nonparametric tool.

It is difficult to compare the estimator \( (1/l) \sum_{i=1}^{l} k_{li} \) of \( \alpha/(1+\alpha) \) with other estimators, since this estimator estimates a different function of the tail index \( \alpha \). One can only look at the asymptotic MSE of the estimates for known distributions with known parameters.

We consider, for example, the ratio estimator (1.7). For Pareto distribution (1.21) we have

\[
\text{MSE}(a_n) = E \left( a_n / \gamma - 1 \right)^2 \\
\sim (2\beta_\gamma - 1) \left( \left( 1 - \frac{1}{\beta_\gamma} \right) \frac{c_2}{c_1} \right)^{2/(2\beta_\gamma - 1)} \cdot (2(\beta_\gamma - 1)c_1 n)^{-2(\beta_\gamma - 1)/(2\beta_\gamma - 1)}
\]

(1.27)

(Novak, 2002) and
DEFINITIONS AND ROUGH DETECTION OF TAIL HEAVINESS

\[ \text{MSE}(z_i) = E(z_i - (1 + \gamma)^{-1})^2 \sim \sigma^2/l \sim \sigma^2 n^{-2\gamma/(1+2\gamma)} = \sigma^2 n^{-2(\beta\gamma-1)/(2\beta\gamma-1)}, \]

(Markovich, 2005a). For the standard Cauchy distribution \( \alpha = \gamma = 1, \beta = 3 \), we have

\[ \text{MSE}(a_n) \sim (5/4)(16/81)^{1/5}(\pi/n)^{4/5} \]

(Novak, 2002) and

\[ \text{MSE}(z_i) \sim (1/12)n^{-4/5}, \]

(Markovich, 2005a).

On-line estimation

An on-line estimator may be defined to be one where each update, following the arrival of a new data value, requires only \( O(1) \) (i.e., a fixed number) of calculations. In this respect, the recursiveness of the estimator is not necessary, although it is convenient.

It is important for us to use the recursive property of the new tail index estimate. Suppose we get the next group of updates \( V_{l+1} \) (this group should contain at least two but not more than \( m \) points). Denote the estimate of the EVI \( \gamma \) obtained by groups \( V_1, \ldots, V_l \) as \( \gamma_l \). Specifically, \( \gamma_l \) obeys the equation

\[ \frac{1}{l} \sum_{i=1}^{l} k_{li} = \frac{1}{1 + \gamma_l} \quad \text{(since} \ \alpha_l = 1/\gamma_l \text{).} \tag{1.28} \]

Furthermore, we have (Markovich, 2005a)

\[ \gamma_{l+i} = \left( \frac{1}{l+1} \sum_{i=1}^{l+1} k_{l+i} \right)^{-1} - 1 = \left( \frac{l}{l+1} \cdot \frac{1}{1 + \gamma_l + k_{l+i}} \right)^{-1} - 1, \]

and after getting \( i \) additional groups \( V_{l+1}, \ldots, V_{l+i} \) with \( m \) elements each,

\[ \gamma_{l+i} = (l+i) \left( \frac{l}{1 + \gamma_i} + k_{l+i} \right)^{-1} - 1, \]

that is, \( \gamma_{l+i} \) is obtained using \( \gamma_l \) after \( O(1) \) calculations. One can rewrite it as

\[ z_{l+i} = \left( l z_i + \sum_{j=1}^{i} k_{l+j} \right)/(l+i). \]

Obviously, the bias of \( z_{l+i} \) is the same as for \( z_l \), but the variance is less. In fact, since \( Ez_i = Ek_{l+i}, Ek_{l+i}^2 = Ek_{l+i}^2, Ez_i^2 = Ez_i^2 = (1/l - 1/l)\sigma_i^2 \), and \( \text{var}(z_i) = (1/l)\sigma_i^2 \), \( (\sigma_i^2 = \text{var}(k_{l+i})) \), we have

\[ \text{var}(z_{l+i}) = \text{var}(z_l)l/(l+i). \]

Evidently, \( \text{var}(z_{l+i}) < \text{var}(z_l) \) for all \( i > 0 \).
The number of operations to select two first maxima in the group with \( m \) elements is \( 2\log m \) (Knuth 1973, Section 5.2.3). We assume here that \( m \) is fixed. If real-time calculation is necessary, then it may not be practicable to recompute \( m \) whenever we believe that a very different value might be required, since the amount of recalculation would be too great. It would be more appropriate to update \( m \) with the new portion of data, relying on a long series of small changes to produce the large changes in \( m \). Indeed, the accuracy of the tail index estimate is expected to be worse than it would be if \( m \) changed with each new portion of observations or if the number of observations in each group were not constant.

As recommended in Paulauskas (2003), in practice one can plot \( \{(m, z_m), m_0 < m < M_0\}, \) \( m_0 > 2, M_0 < n/2, \) where \( z_m = (m/n) \sum_{i=1}^{[n/m]} k_{(n/m)_i} \) (similarly to a Hill plot \( \{(k, \hat{\gamma}^H(n,k)), 1 \leq k \leq n-1\} \) and then choose the estimate of \( z_m \) from an interval in which the function \( z_m \) demonstrates stability. The background of this approach is provided by the consistency result (1.22) as \( n, m \to \infty \), \( m < n \). Hence, there must be an interval \( [m_-, m_+] \) such that \( z_m \approx \alpha/(1+\alpha) = (1+\gamma)^{-1} \) for all \( m \in [m_-, m_+] \). We suggest choosing the average value

\[
\bar{z} = \text{mean}\{1/z_m - 1 : m \in [m_-, m_+]\}
\]

and \( m^* \in [m_-, m_+] \) as a point such that \( z_{m^*} = \bar{z} \). In Figure 1.5 the plot \( \{(m, 1/z_m - 1)\} \) is depicted for Pareto distributed samples with \( \gamma = 1 \); the true \( \gamma \) is shown by a dotted line. For \( n = 1000 \) we suggest taking \( \bar{z} \) over the interval \( [m_-, m_+] = [10, 40] \). Then \( m^* = 10 \) corresponds to the average \( \bar{z} = 1.087 \) and the estimate \( \hat{\gamma} = 0.999 \).

![Figure 1.5 Plot of \( \{(m, 1/z_m - 1)\} \) for the Pareto distribution with \( \gamma = 1 \) and sample sizes \( n = 150 \) (dot-dashed line), \( n = 500 \) (dotted line), \( n = 1000 \) (solid line). The true \( \gamma \) is shown by horizontal solid line.]
The automatic choice of \( m = n/l \) from empirical data could be done by minimizing the empirical bootstrap estimate of the mean squared error of \((1 + \gamma)^{-1}\), that is,

\[
\text{MSE}(\gamma) = E \left( \frac{1}{l} \sum_{i=1}^{l} k_i - \frac{1}{1 + \gamma} \right)^2 \rightarrow \min_m.
\]

The bootstrap estimate is obtained by drawing \( B \) samples with replacement from the original data set \( X^n \). Some observations from \( X^n \) may appear more than once, while others do not appear at all.

One can use smaller resamples \( \{X^*_1, \ldots, X^*_n\} \) of the size \( n_1 < n \) from \( X^n \) to avoid the situation where the bootstrap estimate of the bias (or its asymptotic form) is equal to 0 regardless of the true nonzero bias of the estimator (Hall, 1990). The values \( n_1 \) and \( n \) may be related by

\[ n_1 = n^d, \quad 0 < d < 1. \]

The resample is divided into \( l_1 \) subgroups and \( l_1 = [n_1/m_1] \) holds. The size of subgroups \( m_1 \) and \( m \) are related by:

\[ m = m_1(n/n_1)^c, \quad 0 < c < 1. \tag{1.30} \]

Since the DF \( F(x) \) is unknown, one can find \( m_1 \) by minimizing the empirical bootstrap estimate of the MSE

\[
\text{MSE}^*(l_1, m_1) = \left( \hat{b}^*(l_1, m_1) \right)^2 + \text{var}^*(l_1, m_1) \tag{1.31}
\]

and use this value of \( m_1 \) to calculate an optimal \( m \) using (1.30). All possible values of \( m_1 \), where \( m_1 \) is an integer in the interval \([2, n_1]\), are examined. Here,

\[
\hat{b}^*(l_1, m_1) = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{l_1} \sum_{i=1}^{l_1} k_i \left( z_i^b - z_i \right),
\]

\[
\text{var}^*(l_1, m_1) = \frac{1}{B - 1} \sum_{b=1}^{B} \left( \frac{1}{B} \sum_{b=1}^{B} z_i^b \right)^2
\]

are the empirical bootstrap estimates of the bias and the variance, respectively. We use \( z_i^b = (1/l_1) \sum_{i=1}^{l_1} k_i \), constructed from some resample. Hence, an optimal \( l = n/m \) may be calculated and further used in (1.28) to estimate \( \gamma \). The choice of suitable values of \( c \) and \( d \) is a problem. Based on asymptotic theory, Hall (1990) concludes that \( d = 1/2 \) and \( c = 2/3 \) lead to the most accurate results when the bootstrap estimation of the parameter \( \gamma \) by Hill’s estimate is considered. In what follows, we will tackle the problem by means of simulation. Taking into account the complicated proof technique related to the bootstrap approach we skip here the theoretical investigation of \( c \) and \( d \) for the group estimator.

The number of iterations to find the minimum of (1.31) is defined by the required accuracy and an optimization method (Knuth, 1973).
Application to simulated data

We investigate the influence of $c$ for a fixed $d$ by means of a Monte Carlo study. For this purpose, samples of the Pareto distribution with DF (4.8), Fréchet distribution with DF

$$F(x) = \exp(-(\gamma x)^{-1/\gamma}) 1\{x > 0\},$$

(1.32)

and Weibull distribution with PDF

$$f(x) = \begin{cases} sx^{s-1} \exp(-x^s), & x > 0, \\ 0, & x \leq 0, \end{cases}$$

(1.33)

were generated. The latter distribution does not belong to class (1.20) or (1.21). The use of the statistics $z_I$ as estimators of the tail index for distributions other than (1.20) or (1.21) has yet to be theoretically investigated. Here, the application of the new estimator to the Weibull distribution is investigated by a simulation study. The values $c \in \{0.05, 0.1 (0.1); 0.5\}$ are used for a fixed $d = 0.5$. In Figures 1.6–1.8 one can see that

$$\text{Bias } \gamma = \frac{1}{\gamma} \left( \frac{1}{N_R} \sum_{i=1}^{N_R} \hat{\gamma}_i - \gamma \right),$$

$$\text{RMSE } \gamma = \frac{1}{\gamma} \left( \frac{1}{N_R} \sum_{i=1}^{N_R} (\hat{\gamma}_i - \gamma)^2 \right),$$

which are the relative bias and the square root of the mean squared error of the EVI estimator (1.28) calculated over $N_R = 500$ repeated samples. The mean and standard deviation of the parameter $m$ are also presented. $B = 50$ bootstrap resamples were taken. Samples of sizes $n \in \{150, 500, 1000\}$ are considered.

From the simulation study one can see that the best values of $c$ for fixed $d = 0.5$ are $0.3, 0.4, 0.5$. It is important to mention that the mean and the standard deviation of $m$ behave rather similarly for $c < 0.3$ independent of the sample size. However, for $c > 0.3$ these values increase rapidly for larger $n$.

![Figure 1.6](image-url) Simulation results of $\gamma$ estimation for a Pareto PDF with $\gamma = 1$ and different $c$: 500 samples of $n$ observations; $n = 150$ (solid line), $n = 500$ (dotted line) and $n = 1000$ (dashed line). Relative bias and mean squared error of the EVI estimator $\gamma$ (first two plots on the left). Mean and standard deviation of the parameter $m$ (last two plots on the right).
DEFINITIONS AND ROUGH DETECTION OF TAIL HEAVINESS

Figure 1.7  Simulation results of $\gamma$ estimation for Fréchet PDF with $\gamma = 0.3$ and different $c$: 500 samples of $n$ observations; $n = 150$ (solid line), $n = 500$ (dotted line) and $n = 1000$ (dashed line). Relative bias and mean squared error of the EVI estimator $\gamma$ (first two plots on the left). Mean and standard deviation of the parameter $m$ (last two plots on the right).

Figure 1.8  Simulation results of $\gamma$ estimation for Weibull PDF with $s = 0.5$ and different $c$: 500 samples of $n$ observations; $n = 150$ (solid line), $n = 500$ (dotted line) and $n = 1000$ (dashed line). Relative bias and mean squared error of the EVI estimator $\gamma$ (first two plots on the left). Mean and standard deviation of the parameter $m$ (last two plots on the right).

The confidence intervals for the estimate $\gamma_i$ with bootstrap estimated $m$ are given in Table 1.2 for different heavy-tailed distributions, the levels $p \in \{0.025, 0.05, 0.1\}$, and the best values of $c \in \{0.3, 0.4, 0.5\}$. It is assumed that the bootstrap estimates are normally distributed with mean and variance constructed from the set of bootstrap estimates $\gamma_1^*, \ldots, \gamma_{N_R}^*$, where $N_R$ is the number of bootstrap resamples. Then one can calculate the tolerant bounds of confidence intervals by the well-known formula (Smirnov and Dunin-Barkovsky, 1965)

$$(u_1, u_2) = (\text{mean } \gamma - \rho \cdot \text{StDev } \gamma; \text{mean } \gamma + \rho \cdot \text{StDev } \gamma),$$

where mean $\gamma$ and StDev $\gamma$ are the mean and the standard deviation of $N_R$ bootstrap estimates. The interval is constructed in such a way that $100(1 - p)$% of the distribution falls in this interval with probability $P$. The value $\rho$ depends on $N_R$, $P$, $p$ and may be approximately represented by

$$\rho = \rho_{\infty} \left( 1 + \frac{t_p}{\sqrt{2N_R}} + \frac{5t_p^2 + 10}{12N_R} \right).$$

(1.34)
The equation is used to calculate the confidence interval of the bootstrap γ estimates. The interval is given by:

\[ \text{Confidence interval} = (\gamma - 2.5 \times \text{StDev}_\gamma, \gamma + 2.5 \times \text{StDev}_\gamma) \]

where StDev_\gamma is the standard deviation of the bootstrap γ estimates. The table below provides the confidence intervals for different distributions and values of γ:

<table>
<thead>
<tr>
<th>PDF</th>
<th>( c )</th>
<th>( \text{mean}<em>\gamma ) (StDev</em>\gamma)</th>
<th>( p \cdot 100% )</th>
<th>Confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pareto</td>
<td>0.3</td>
<td>1.191 (0.606)</td>
<td>2.5</td>
<td>((-0.278, 2.66))</td>
</tr>
<tr>
<td>( \gamma = 1 )</td>
<td></td>
<td></td>
<td>5</td>
<td>((-0.098, 2.48))</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>((0.115, 2.267))</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>1.141 (0.575)</td>
<td>2.5</td>
<td>((-0.253, 2.535))</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>((-0.082, 2.364))</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>((0.12, 2.162))</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.964 (0.508)</td>
<td>2.5</td>
<td>((-0.268, 2.196))</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>((-0.117, 2.045))</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>((0.062, 1.866))</td>
</tr>
<tr>
<td>Fréchet</td>
<td>0.3</td>
<td>0.3072 (0.155)</td>
<td>2.5</td>
<td>((-0.069, 0.683))</td>
</tr>
<tr>
<td>( \gamma = 0.3 )</td>
<td></td>
<td></td>
<td>5</td>
<td>((-0.023, 0.637))</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>((0.032, 0.583))</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.3501 (0.175)</td>
<td>2.5</td>
<td>((-0.074, 0.774))</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>((-0.022, 0.722))</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>((0.039, 0.661))</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.3417 (0.172)</td>
<td>2.5</td>
<td>((-0.075, 0.759))</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>((-0.024, 0.708))</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>((0.036, 0.647))</td>
</tr>
<tr>
<td>Weibull</td>
<td>0.3</td>
<td>0.738 (0.371)</td>
<td>2.5</td>
<td>((-0.161, 1.637))</td>
</tr>
<tr>
<td>( \gamma = 0.5 )</td>
<td></td>
<td></td>
<td>5</td>
<td>((-0.051, 1.527))</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>((0.079, 1.397))</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.5665 (0.291)</td>
<td>2.5</td>
<td>((-0.139, 1.272))</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>((-0.053, 1.186))</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>((0.05, 1.083))</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.562 (0.296)</td>
<td>2.5</td>
<td>((-0.156, 1.28))</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>((-0.068, 1.192))</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>((0.036, 1.088))</td>
</tr>
</tbody>
</table>

Reprinted from *Proceedings of 1st Conference on Next Generation Internet Design and Engineering*, On-line estimation of the tail index for heavy-tailed distributions with application to WWW-traffic, Markovich NM, Table 1, © 2005 IEEE. With permission from IEEE.

Here, \( \rho_\infty \) is defined by the equation

\[
\frac{1}{\sqrt{2\pi}} \int_{-\rho_\infty}^{\rho_\infty} e^{-t^2/2} dt = 2\Phi_0(\rho_\infty) = 1 - p,
\]

where \( \Phi_0(z) = \frac{1}{\sqrt{2\pi}} \int_{-\rho_\infty}^{z} e^{-t^2/2} dt \) is Laplace’s function. The DF of the standard normal distribution \( N(z; 0, 1) \) can be expressed as \( N(z; 0, 1) = 0.5 + \Phi_0(z) \) for positive \( z \). Furthermore, \( \Phi_0(-z) = -\Phi_0(z), \Phi_0(0) = 0, \Phi_0(-\infty) = -1/2, \Phi_0(\infty) = 1/2 \). The value \( t_p \) is calculated by the equation

\[
\frac{1}{\sqrt{2\pi}} \int_{-\rho_\infty}^{\rho_\infty} e^{-t^2/2} dt = 2\Phi_0(\rho_\infty) = 1 - p,
\]

where \( \Phi_0(z) = \frac{1}{\sqrt{2\pi}} \int_{-\rho_\infty}^{z} e^{-t^2/2} dt \) is Laplace’s function. The DF of the standard normal distribution \( N(z; 0, 1) \) can be expressed as \( N(z; 0, 1) = 0.5 + \Phi_0(z) \) for positive \( z \). Furthermore, \( \Phi_0(-z) = -\Phi_0(z), \Phi_0(0) = 0, \Phi_0(-\infty) = -1/2, \Phi_0(\infty) = 1/2 \). The value \( t_p \) is calculated by the equation

\[
\frac{1}{\sqrt{2\pi}} \int_{-\rho_\infty}^{\rho_\infty} e^{-t^2/2} dt = 2\Phi_0(\rho_\infty) = 1 - p,
\]
26 DEFINITIONS AND ROUGH DETECTION OF TAIL HEAVINESS

\[
\frac{1}{\sqrt{2\pi}} \int_{t_p}^{\infty} e^{-t^2/2} dt = 0.5 - \Phi_0(t_p) = 1 - P. \tag{1.36}
\]

For \( P = 0.99 \) we have \( \Phi_0(x_p) = P - 0.5 = 0.49 \) and \( t_p = 2.33 \). Furthermore, we have \( \rho_\infty \in \{2.245, 1.97, 1.645\} \) for \( p \in \{0.025, 0.05, 0.1\} \), respectively.

In Table 1.2 mean \( \gamma \) and StDev \( \gamma \) are shown. Samples of size \( n = 1000 \) and \( N_R = 500 \) were used. As before, \( B \) was taken equal to 50. From Table 1.2 it follows that the ratio \( r_1 = \text{mean} \gamma / \gamma \) is closer to 1 for the Pareto and Fréchet distributions than for the Weibull distribution. This may imply the larger bias of the new estimator for the latter distribution.

The ratios \( r_2 = u_2 / \gamma - 1 \) and \( r_3 = u_1 / \gamma - 1 \) corresponding to the upper and lower bounds \( u_1 \) and \( u_2 \) of the confidence interval are considered. The larger \( r_2 \) and \( r_3 \) are in absolute value, the wider is the confidence interval. From Table 1.2, \( r_2 \in M, M \in \{(0.866, 1.66); (1.58, 1.581); (1.166, 2.274)\} \), \( r_3 \in L, L \in \{(-1.278, -0.885); (-1.258, -0.88); (-1.322, -0.842)\} \) hold for Pareto, Fréchet

![Figure 1.9](image-url)

Figure 1.9 Absolute values of ratios \( r_2 \) and \( r_3 \) for Pareto (solid line), Fréchet (dotted line), and Weibull distributions (dashed line), for confidence levels \( p \in \{0.025, 0.05, 0.1\} \) and parameter \( c \in \{0.3, 0.4, 0.5\} \). Reprinted from Proceedings of 1st Conference on Next Generation Internet Design and Engineering, On-line estimation of the tail index for heavy-tailed distributions with application to WWW-traffic, Markovich NM, Figure 5, © 2005 IEEE. With permission from IEEE.
and Weibull, respectively. This implies that the confidence interval is worse for the Weibull distribution, at least with regard to the upper bound $u_2$.

Figure 1.9 shows the comparison of $r_2$ and $r_3$ for different $c$ and $p$. From this figure one may conclude that in most cases the values $c = 0.4$ and $c = 0.5$ correspond to the best values of $u_1$ and $u_2$, respectively. The value $c = 0.4$ gives a cautious decision in the sense that $|r_2|$ is the same irrespective of distribution and $|r_3|$ is not maximized. In all cases, the Weibull distribution has the largest $r_2$ and $r_3$. Together with the previous conclusion, this implies that the confidence intervals for this distribution are worse than for Pareto and Fréchet distributions.

1.3 Detection of tail heaviness and dependence

Before a serious analysis of the data is carried out, it is necessary to detect heavy tails in the data. For this purpose, nonparametric test procedures (e.g., Jurečková and Picek, 2001) or a set of rough statistical methods for heavy-tailed features (Embrechts et al., 1997; Markovich and Krieger, 2006b) can be applied. Here, we consider several simple procedures that may help us to detect heavy tails and the dependence structure of the data. We illustrate by means of real data how these methods are applied to analyze traffic measurements.

1.3.1 Rough tests of tail heaviness

Here, we consider several methods in order to check whether measurements $X^n = \{X_1, X_2, \ldots, X_n\}$ are derived from a heavy-tailed DF $F(x) = P[X_1 \leq x]$ or not. We may also give rough estimates of the number of finite moments of the DF $F(x)$.

**Ratio of the maximum to the sum**

Let $X_1, X_2, \ldots, X_n$ be i.i.d. r.v.s. We define the statistic (Embrechts et al., 1997, p. 308)

$$R_n(p) = \frac{M_n(p)}{S_n(p)}, \quad n \geq 1, p > 0, \quad (1.37)$$

where

$$S_n(p) = |X_1|^p + \ldots + |X_n|^p, \quad M_n(p) = \max(|X_1|^p, \ldots, |X_n|^p), \quad n \geq 1, \quad (1.38)$$

to check the moment conditions of the data. Then the following equivalent assertions

$$R_n(p) \xrightarrow{a.s.} 0 \iff E|X|^p < \infty,$$
$$R_n(p) \xrightarrow{p} 0 \iff E|X|^p 1\{|X| \leq x\} \in R_0,$$
$$R_n(p) \xrightarrow{p} 1 \iff P\{|X| > x\} \in R_0,$$
can be exploited. The class $R_{-\alpha}$ of distributions with regularly varying tails and the tail index $\alpha = 1/\gamma$, $\gamma > 0$, are defined in Definition 10.

For different values of $p$ the plot of $n \rightarrow R_n(p)$ gives some preliminary information about the distribution $P[|X| > x]$. Then $E|X|^p < \infty$ follows if $R_n(p)$ is small for large $n$. For large $n$, a significant difference between $R_n(p)$ and zero indicates that the moment $E|X|^p$ is infinite.

**Quantile–quantile plot**

The idea of the quantile–quantile (QQ) plot is to show the relationship

$$\left\{(X_{(k)}, F^{-1}\left(\frac{n-k+1}{n+1}\right)) : k = 1, \ldots, n\right\},$$

where $X_{(1)} \geq \ldots \geq X_{(n)}$ are the order statistics of the sample. A QQ plot is based on the following fact. It is well known that for an i.i.d. sample with a continuous DF $F(x)$ the r.v.s

$$U_i = F(X_i), \quad i = 1, \ldots, n,$$

are independent and uniformly distributed on $[0,1]$. Then $X_i = F^{-1}(U_i)$. For example, if the exponential DF $F(x)$ is believed to be the distribution of $X$, then we plot exponential quantiles $F^{-1}\left(\frac{n-k+1}{n+1}\right)$ against the order statistics $X_{(k)}$ of the underlying sample. Then $F^{-1}(x)$ is an inverse function of the exponential DF and a linear QQ plot corresponds to the exponential distribution. Generally, one can investigate the quantiles of any distribution, not just an exponential. The linearity of a QQ plot shows that the parametric model of the distribution is selected correctly.

**Plot of the mean excess function**

To study the tail behavior in more detail, this simple test allows us to detect visually whether a tail is light or heavy. Let $X$ be an r.v. with the finite right endpoint $X_F = \sup\{x \in R : F(x) < 1\}$ of its support. Then

$$e(u) = E(X - u|X > u), \quad 0 \leq u < X_F \leq \infty$$

(1.40)

is the mean excess function of the r.v. $X$ over the threshold $u$. The sample mean excess function is defined by

$$e_n(u) = \frac{\sum_{i=1}^{n}(X_i - u)\mathbb{1}\{X_i > u\}}{\sum_{i=1}^{n}\mathbb{1}\{X_i > u\}}.$$  

(1.41)

For heavy-tailed distributions the function $e(u)$ tends to infinity. A linear plot $u \rightarrow e(u)$ corresponds to a Pareto distribution, the constant $1/\lambda$ corresponds to an exponential distribution, and $e(u)$ tends to zero for light-tailed distributions.\(^8\)

\(^8\) The mean excess function is calculated by formula $e(u) = (1/F(u)) \int_u^{X_F} F(x)dx$, which follows from (1.40), (Embrechts et al., 1997, p. 162).
Figure 1.10  Left: Mean excess functions for some distributions: exponential (horizontal solid line), lognormal with parameters (0,1) (dotted line), Pareto with shape parameter equal to 1 (upper solid line), Weibull with shape parameter equal to 0.5 (solid line) and 2 (dot-dashed line). Right: Ten empirical mean excess functions $e_n(u)$, each based on simulated data of size $n = 1000$ from the Pareto distribution $F(x) = 1 - (1 + x/2)^{-2}$, $x \geq 0$. A very unstable behavior, especially towards the higher values of $u$, can be seen.

For different samples the curves $e_n(u)$ may differ strongly towards the higher values of $u$ since only sparse observations may exceed the threshold $u$ for large $u$, as shown in Figure 1.10, based on Embrechts et al. (1997). This makes the precise interpretation of $e_n(u)$ difficult.

**Hill’s estimator of the tail index**

Hill’s estimator (1.5) is valid for a positive EVI $\gamma$ of a heavy-tailed r.v. $X$ and can also be constructed for dependent data. Hill’s estimator may be considered as the empirical mean excess function of the r.v. $\ln X$ for the level $u = \ln X_{(n-k)}$.

Hill’s estimator is inadequate if the underlying DF does not have a regularly varying tail, that is, $F \subseteq R^{-\alpha}$ does not hold, $\alpha$ is not positive, the sample size is not large enough, and the tail is not heavy enough ($\gamma$ is not big). If $F \subseteq R^{-\alpha}$, then estimation by (1.5) strongly depends on the type of slowly varying function, which is usually unknown. The disadvantages of Hill’s estimate show that one has to apply several estimates of the tail index (see, Section 1.2) to deal with the complex analysis of data.

Hill’s and other estimators are very sensitive to the choice of smoothing parameter. In the case of Hill’s estimator (1.5) this is the number of largest order statistics $k$, while for the group estimator (1.19), (1.28) it is the number $m$ of observations in each group. The use of plots (e.g., a Hill plot or the respective plot of the group estimator) of the estimator against the smoothing parameter provides the easiest way to select such parameters.
In particular, the tail index estimates are applied to investigate the amount of finite moments of the r.v. It is well known that the $p$th moment exists, that is, $E|X_1|^p < \infty$ holds, if the tail index $\alpha = 1/\gamma$ satisfies $0 < p < \alpha$ and when the distribution has regularly varying tails, that is, belongs to class $\mathcal{R}_{-\alpha}$ (Lemma 1, p. 5). A positive $\gamma$ indicates the presence of a heavy tail. The simple tests sketched above can be successfully applied to analyze visually the features of a data attribute arising from Internet traffic measurements.

1.3.2 Analysis of Web traffic and TCP flow data

To illustrate the efficiency of the methods sketched above, two sets of real data are investigated. Data on Web traffic were gathered in the Ethernet segment of the Department of Computer Science at the University of Würzburg (Vicari, 1997). Data on transmission control protocol (TCP) flow sizes and transmission durations were measured from a mobile network.

Description of the Web traffic data

The measured traffic is described by a conventional hierarchical Web traffic model distinguishing a session and a page level, where the former is characterized by sub-sessions (see Figure 1.11, based on Figure 1.2 in Krieger et al., 2001). Responses to web requests are identified as the main part of the transferred data and the time between these responses is used to model the relationship between the responses.

![Figure 1.11](image-url) Hierarchical modeling of Web sessions.
Consequently, the data are described at the related coarse time scales by two basic characteristics and four related r.v.s; two of these r.v.s are characteristics of sub-sessions, that is, the size of a sub-session (s.s.s.) in bytes and its duration (d.s.s.) in seconds, and two are characteristics of the transferred Web pages, that is, the size of the response (s.r.) in bytes and the inter-response time (i.r.t.) in seconds; see Table 1.3, based on Table 1 in Krieger et al. (2001).

The page size data analyzed contain information on about 7480 Web pages downloaded in several TCP connections over a period of 14 days. The size of a response is defined as the sum of the sizes of all IP packets which are downloaded from a Web server to the client upon a request. To perform the analysis, we have used samples with reduced sample sizes, which were observed in a shorter period within these two weeks. The description of all these r.v.s is presented in Table 1.4, based on Krieger et al. (2001). For simplicity of the calculations, the data were scaled, that is, divided by a scale parameter \( s \). The value of \( s \) is indicated in Table 1.4.

### Description of TCP flow data and research motivation

We analyze real data on TCP flow traffic in an access network (Markovich and Kilpi, 2006). The data that we use are derived from a trace measured at a gateway.

<table>
<thead>
<tr>
<th>Level</th>
<th>Characteristic</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>sub-session</td>
<td>duration (d.s.s. [sec])</td>
<td>time between beginning and end of browsing a series of Web pages</td>
</tr>
<tr>
<td>page</td>
<td>size (s.s.s. [byte])</td>
<td>data volume of Web pages visited</td>
</tr>
<tr>
<td></td>
<td>inter-response time (i.r.t. [sec])</td>
<td>time between beginning of the old and of the new transfer of pages within a sub-session</td>
</tr>
<tr>
<td></td>
<td>page size (s.r. [byte])</td>
<td>total amount of transferred data (HTML, images, sound, . . .)</td>
</tr>
</tbody>
</table>

### Table 1.3 Characteristics of Web sessions.

<table>
<thead>
<tr>
<th>r.v.</th>
<th>Sample size</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>StDev</th>
<th>( s )</th>
</tr>
</thead>
<tbody>
<tr>
<td>s.s.s.(B)</td>
<td>373</td>
<td>128</td>
<td>5.884 \cdot 10^7</td>
<td>1.283 \cdot 10^6</td>
<td>4.079 \cdot 10^6</td>
<td>10^7</td>
</tr>
<tr>
<td>d.s.s.(sec)</td>
<td>373</td>
<td>2</td>
<td>9.058 \cdot 10^4</td>
<td>1.728 \cdot 10^3</td>
<td>5.206 \cdot 10^3</td>
<td>10^3</td>
</tr>
<tr>
<td>s.r.(B)</td>
<td>7107</td>
<td>0</td>
<td>2.052 \cdot 10^7</td>
<td>5.395 \cdot 10^4</td>
<td>4.931 \cdot 10^5</td>
<td>10^6</td>
</tr>
<tr>
<td>i.r.t.(sec)</td>
<td>7107</td>
<td>6.543 \cdot 10^{-3}</td>
<td>5.676 \cdot 10^4</td>
<td>80.908</td>
<td>728.266</td>
<td>10^3</td>
</tr>
</tbody>
</table>

Reprinted from *Proceedings of 1st Conference on Next Generation Internet Design and Engineering*, On-line estimation of the tail index for heavy-tailed distributions with application to WWW-traffic, Markovich NM, Table 2, © 2005 IEEE. With permission from IEEE.
between a mobile network and the Internet. The important thing is that all flows have passed through a mobile access device, hence the rather limited access rates.

TCP flow sizes and durations gathered from one source–destination pair both contain information about the performance of the TCP protocol at this individual level and of the network in question. A bivariate view of the flow size and flow duration provides even more information than a separate analysis of these two quantities.

We assume that a user selects some random Web content, that is, the user chooses the size $S$ according to the file size distribution, and then downloads this content using TCP. The action of a user initiates the TCP connection and its arrival time $AT$. In addition, the TCP protocol generates the flow departure time $DT$ when the download is finished and, thus, determines the flow duration $D = DT - AT$. We aim to analyze the dependence between the r.v.s $S$ and $D$.

The data analyzed consist of mobile TCP connections from periods of low, average, and high network load conditions. To obtain samples as homogeneous as possible only downstream TCP flows on port 80 are considered. This means that such flows are in principle running a WWW (HTTP) application. The total number of such analyzed flows is over 610,000 and, for practical reasons, we consider here 61 disjoint bivariate samples, each of size $n = 10,000$. Table 1.5 (Markovich and Kilpi, 2006) states the observed ranges [min, max] of sample means, variances, and maxima over these 61 samples.

‘Content’ in Table 1.5 refers to the size of the downloaded Web content and ‘Transmitted’ means Content plus segments retransmitted by TCP. Both are measures of the size of a flow. ‘SYN-FIN’ means from the three-way handshaking (synchronization) to finish. In other words, the flow duration is defined by the time difference between the SYN packet in the three-way handshake and the FIN packet at the end of the flow.

Regarding the analysis of TCP flow data, the distribution of the maximal rate (or throughput) $R = S/D$ and the expected throughput $ER$ (or $ES/ED$) that the transport system provides are the objects of interest.

A form of asymptotic independence for the pair $(D, R)$ is obtained in Resnick (2006, p. 239) as a result of the examination the tail of the product $DR$.

The distributions of both $S$ and $D$ are heavy-tailed and their expectations may not be finite (see Table 1.9). Thus, $ES/ED$ may be not computable. The heaviness

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Unit</th>
<th>Definition</th>
<th>Sample mean</th>
<th>Sample variance</th>
<th>Sample maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>kB</td>
<td>Content</td>
<td>9.0</td>
<td>20.3</td>
<td>1303</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Transmitted</td>
<td>9.5</td>
<td>20.3</td>
<td>1357</td>
</tr>
<tr>
<td>Duration</td>
<td>sec</td>
<td>SYN-FIN</td>
<td>18.2</td>
<td>30.4</td>
<td>2219</td>
</tr>
</tbody>
</table>
of tails means that outliers, that is, not typical observations, play a significant role in the distribution. Sometimes, this gives the wrong impression that outliers are not identically distributed with the rest of the data.

The appearance of outliers in the data is quite natural. If \( D \) is very large, then \( S \) will also be quite large. But the problem is that streaming applications affect the normal behavior of TCP by prohibiting it from using the full transfer capacity available and, hence, \( S \) is much smaller than it should be. It is shown in Kilpi and Lassila (2006) that this specific data trace contains such streaming applications.

Since \( S \) and \( D \) are dependent (see the results of the empirical study in Section 1.3.5) and positive, the DF of the ratio \( R = S/D \) is defined by

\[
F_R(x) = P\{S/D \leq x\} = \int_0^\infty \int_0^{\frac{x}{y}} f(y, z) dy dz = \int_0^\infty \int_0^{\frac{x}{y}} dF(y, z)
\]

where \( f(y, z) \) is a joint PDF of \( S \) and \( D \), and its expectation by

\[
ER = \int_0^\infty x dF_R(x),
\]

if the latter integral converges.

There are several alternatives to estimating \( F_R(x) \). The required joint bivariate distribution could be estimated by copulas and the bivariate PDF by the copula density (Nelsen, 1998). Specifically, Sklar’s theorem gives us a unique representation \( F(y, z) = C(F_S(y), F_D(z)) \) by means of copula \( C(u, v) \) if the marginal DFs \( F_S(y) \) and \( F_D(z) \) of two r.v.s \( S \) and \( D \) are continuous. However, how to select \( C(u, v) \) using statistical tools when marginal DFs are unknown remains a problem (Mikosch, 2006).

One can estimate \( f(y, z) \) by some multivariate nonparametric method, for example, the product kernel method (Scott, 1992). The problem is to find the appropriate values of bandwidths from samples of moderate size.

Another alternative is to estimate the marginal DF \( F_D(z) \) and the conditional DF \( F_{S|D}(z|x|z) \) from empirical data. \( F_{S|D}(z|x|z) \) requires a sufficiently large data set, that is, we need observations of the size for a fixed value of the duration, which may be not available. In the context of heavy-tailed distributions and statistical estimation of characteristics of heavy-tailed r.v.s, it is common practice to use the asymptotic distribution (in the sense that the sample size increases without bound) of the sample maxima as a model of the tail. We estimate a bivariate extreme value distribution (EVD) (1.48) of the pair \((S, D)\) instead of \( F(y, z) \) itself and use the EVD as its approximation to estimate \( f(y, z) = \frac{\partial^2}{(\partial y \partial z)} F(y, z) \). The estimation of the EVD requires the preliminary evaluation of the marginal distributions of both \( S \) and \( D \).
DEFINITIONS AND ROUGH DETECTION OF TAIL HEAVINESS

Results of the Web traffic analysis

For the d.s.s., s.s.s., i.r.t., and s.r. data sets, Figures 1.12 and 1.13 show plots of $R_n(p)$ against $n$ for various $p$. In all cases, the values $R_n(p)$ are dramatically large for large $n$ and $p \geq 2$. Hence, one may conclude that all moments of the r.v.s considered, apart from the first, are not finite. Furthermore, the plots $u \rightarrow e_n(u)$ tend to infinity for large $u$, implying heavy tails. These plots are close to a linear shape for all sets of data (Figures 1.14 and 1.15). The latter implies that the distributions considered can be modeled by a DF of Pareto type.

The QQ plots of d.s.s., s.s.s., i.r.t. and s.r. are shown in Figures 1.16–1.19. The left-hand plots show that the exponential distribution cannot be accepted as an appropriate model for these r.v.s. The right-hand plots show that the distributions of the d.s.s., s.s.s., i.r.t. and s.r. samples are close to a generalized Pareto distribution $\text{GPD}(\sigma, \gamma)$ (1.16) with different values of the parameters $\gamma$ and $\sigma$ (see also Table 1.7). Figure 1.17 shows that both $\text{GPD}(0.015, 1)$ and $\text{GPD}(0.05, 0.3)$ could

Figure 1.12  Plots of $R_n(p)$ against $n$ for the duration of sub-sessions (left) and the size of sub-sessions (right) for a variety of $p$-values: curves corresponding to $p = 0.5, 1, 2, 3, 4, 5$ are located from bottom to top, respectively.

Figure 1.13  Plots of $R_n(p)$ against $n$ for the inter-response times (left) and the size of responses (right) for a variety of $p$-values: curves corresponding to $p = 0.5, 1, 2, 3, 4, 5$ are located from bottom to top, respectively.
DEFINITIONS AND ROUGH DETECTION OF TAIL HEAVINESS

Figure 1.14 Exceedance $e_n(u)$ against the threshold $u$ for the duration of sub-sessions (left) and the size of sub-sessions (right).

Figure 1.15 Exceedance $e_n(u)$ against the threshold $u$ for the inter-response times (left) and the size of responses (right).

Figure 1.16 QQ plots for the duration of sub-sessions (d.s.s./s) against exponential quantiles (left) and quantiles of the GPD(1,0.3) distribution (right).
be appropriate models of s.s.s. This implies that the QQ plot does not give a unique model to fit the underlying distribution.

Table 1.6 shows the estimation of the EVI by means of the group estimator $\gamma_i$ with the bootstrap-selected parameter $m$ (these are denoted by $\gamma_i^b$ and $m_b$, respectively) and the selection of $m$ from a plot (see formula (1.29)); the latter is denoted by $\gamma_i^p$. The Hill estimates with the plot- and bootstrap-selected $k$ ($\hat{\gamma}_i^H(n, k)$ and $\hat{\gamma}_b^H(n, k)$, respectively) are also presented. In order to calculate $\gamma_i^p$ the averaging over $m = 10, 11, \ldots, 35, m = 10, 11, \ldots, 40, m = 100, 101, \ldots, 200,$
DEFINITIONS AND ROUGH DETECTION OF TAIL HEAVINESS

Figure 1.19  QQ plots for the size of responses (s.r./s) against exponential quantiles (left) and quantiles of the GPD(0.015, 1) distribution (right).

Table 1.6  Estimation of the EVI for Web traffic characteristics.

<table>
<thead>
<tr>
<th>r.v.</th>
<th>c</th>
<th>( m_b )</th>
<th>( \gamma^H_1 )</th>
<th>( \gamma^H_0 )</th>
<th>( \hat{\gamma}^H_p(n, k) )</th>
<th>( \hat{\gamma}^H_b(n, k) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>s.s.s.</td>
<td>0.3</td>
<td>8</td>
<td>1.179</td>
<td>0.877</td>
<td>0.96</td>
<td>0.949</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>10</td>
<td>0.856</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>22</td>
<td>0.902</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s.r.</td>
<td>0.3</td>
<td>72</td>
<td>0.75</td>
<td>0.8</td>
<td>0.84</td>
<td>0.898</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>71</td>
<td>0.87</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>92</td>
<td>0.85</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i.r.t.</td>
<td>0.3</td>
<td>42</td>
<td>0.69</td>
<td>0.495</td>
<td>0.48</td>
<td>0.712</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>65</td>
<td>0.625</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>156</td>
<td>0.611</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d.s.s.</td>
<td>0.3</td>
<td>10</td>
<td>0.658</td>
<td>0.739</td>
<td>0.6</td>
<td>0.601</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>13</td>
<td>0.539</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>18</td>
<td>0.683</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reprinted from Proceedings of 1st Conference on Next Generation Internet Design and Engineering, On-line estimation of the tail index for heavy-tailed distributions with application to WWW-traffic, Markovich NM, Table 3, © 2005 IEEE. With permission from IEEE.

\( m = 10, 11, \ldots, 70 \) in the cases of s.s.s., d.s.s., i.r.t., s.r. is done. According to our investigation regarding the bootstrap method, the values \( c \in \{0.3, 0.4, 0.5\} \) and \( d = 0.5 \) were considered. For Hill’s estimate the parameters \( \alpha = 2/3 \) and \( \beta = 1/2 \) were selected for the bootstrap scheme. As one can see, the values of \( \gamma^H_1 \) are sufficiently close to the values of \( \gamma^H_0 \) as well as \( \hat{\gamma}^H_p(n, k) \) and \( \hat{\gamma}^H_b(n, k) \), apart from the case of i.r.t.

Figures 1.20 and 1.21 illustrate the estimation of the EVI by Hill’s estimator and the group estimator \( \gamma_1 \) for the s.s.s., d.s.s., i.r.t., and s.r. data sets. One observes that the values of \( \gamma \) recommended by both estimates are similar. In the case of
Figure 1.20  EVI estimation by Hill’s estimator (dotted line) and the group estimator \( \gamma_i \) (solid line) for the s.s.s. (left) and d.s.s. (right) data sets. The two horizontal solid lines correspond to the levels of stability for the group and Hill’s estimators, 0.877 and 0.96 (left), 0.739 and 0.6 (right), respectively. Reprinted from *Proceedings of 1st Conference on Next Generation Internet Design and Engineering*, On-line estimation of the tail index for heavy-tailed distributions with application to WWW-traffic, Markovich NM, Figure 6, © 2005 IEEE. With permission from IEEE.

Figure 1.21  EVI estimation by Hill’s estimator (dotted line) and the group estimator \( \gamma_i \) (solid line) for the i.r.t. (left) and s.r. (right) data sets. The two horizontal solid lines correspond to the levels of stability for the group and Hill’s estimators, 0.495 and 0.48 (left), 0.8 and 0.84 (right), respectively. Reprinted from *Proceedings of 1st Conference on Next Generation Internet Design and Engineering*, On-line estimation of the tail index for heavy-tailed distributions with application to WWW-traffic, Markovich NM, Figure 7, © 2005 IEEE. With permission from IEEE.
d.s.s. the difference between the values (0.739 and 0.6) arises as a result of the selection of \( k \) in Hill’s estimate corresponding to one of the stability intervals of the Hill plot; see Figure 1.20 (right). Indeed, one may select another stability interval corresponding to a value closer to 0.739. The latter example demonstrates the obvious disadvantage of the Hill plot approach, namely, the selection of \( k \) and the opportunity to use different tail index estimates.

Observing the estimates of \( \gamma \), one may conclude that the estimates of the tail index \( \alpha = 1/\gamma \) are always less than 2 for all data sets considered, apart from i.r.t. with \( 1 < \alpha < 3 \). It follows from the extreme value theory (Embrechts et al., 1997), that at least the \( \beta \)th moments, \( \beta \geq 2 \), of the distribution of the s.s.s., s.r., and d.s.s. are not finite if one believes that these distributions are regularly varying. The distribution of i.r.t. may have two finite moments. It might be possible for s.s.s. (e.g., when \( 1 < \hat{\gamma} < 2 \)) that \( \alpha < 1 \) and the expectation may also be infinite. According to the previous investigation regarding the confidence interval of the bootstrap estimator, one may trust \( c = 0.4 \) as a most cautious choice. Then \( \hat{\gamma} < 1 \) holds with regard to s.s.s. and the first moment exists. The distributions of the Web traffic characteristics considered are heavy-tailed. The tail of the s.s.s. distribution is the heaviest since \( \gamma \) is the largest. With regard to the data analysis, this means that the Web data requires specific methods. More detailed information on the form of the distribution can be obtained from a sample mean excess plot or a QQ plot.

Table 1.7 summarizes the result of this preliminary analysis of the samples with our simple set of exploratory methods.

### Results of the TCP flow data analysis

The Hill’s \( \hat{\gamma}^H(n,k) \), moment \( \hat{\gamma}^M(n,k) \), UH \( \hat{\gamma}^{UH}(n,k) \) and group \( \gamma_i \) estimators of the tail index were applied to observed flow sizes and durations. The results, again [min,max] ranges over all 61 samples, are given in Table 1.8 (Markovich and Kilpi, 2006). For each sample the parameter \( k \) was estimated by a bootstrapping method. For the group estimator we used \( m=l=100 = \sqrt{n} \). The positive sign of all estimates allows us to suppose that both flow sizes and durations are heavy-tailed

<table>
<thead>
<tr>
<th>r.v.</th>
<th>Number of first finite moments</th>
<th>Type of distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( R_n(p) ) Hill and group estimator</td>
<td>QQ plot ( e_n(u) )</td>
</tr>
<tr>
<td>s.s.s.(B)</td>
<td>1 1</td>
<td>GPD(( \sigma = 0.015 ), ( \gamma = 1 )) Pareto-like</td>
</tr>
<tr>
<td>d.s.s.(sec)</td>
<td>1 1</td>
<td>GPD(( \sigma = 0.05 ), ( \gamma = 0.3 )) Pareto-like</td>
</tr>
<tr>
<td>s.r.(B)</td>
<td>1 1</td>
<td>GPD(( \sigma = 0.015 ), ( \gamma = 1 )) Pareto-like</td>
</tr>
<tr>
<td>i.r.t.(sec)</td>
<td>1 2</td>
<td>GPD(( \sigma = 0.015 ), ( \gamma = 0.8 )) Pareto-like</td>
</tr>
</tbody>
</table>
Table 1.8  Estimation of the EVI for flow sizes (‘Content’ and ‘Transmitted’) and durations (‘SYN-FIN’).

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\gamma}^H(n, k)$</th>
<th>$\gamma_I$</th>
<th>$\hat{\gamma}^M(n, k)$</th>
<th>$\hat{\gamma}^{UH}(n, k)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td>0.5923</td>
<td>0.8747</td>
<td>0.4483</td>
<td>0.9794</td>
</tr>
<tr>
<td>Transmitted</td>
<td>0.5760</td>
<td>1.1508</td>
<td>0.4476</td>
<td>0.9437</td>
</tr>
<tr>
<td>SYN-FIN</td>
<td>0.5213</td>
<td>1.0034</td>
<td>0.3770</td>
<td>0.7748</td>
</tr>
</tbody>
</table>

Figure 1.22  Plot of $R_n(p)$ against $n$ for the size of flows (left) and the duration of transmissions (right) for a variety of $p$-values: curves corresponding to $p = 0.5, 1, 1.5, 2$ are located from the bottom to top, respectively.

Figure 1.23  $\gamma$ estimation by Hill’s (solid line) and moment estimator (dotted line) for the flow size (left) and the duration of transmission (right). Horizontal lines show the bootstrap-selected values: $\hat{\gamma}^H(n, k) = 0.718445$ and $\hat{\gamma}^M(n, k) = 0.840629$ (left), $\hat{\gamma}^H(n, k) = 0.608669$ and $\hat{\gamma}^M(n, k) = 0.683828$ (right).

distributed. All estimators, apart from the group estimator $\gamma_I$, indicate that the flow size samples (both content and transmitted) may have infinite variance under the assumption that their distributions are regularly varying. Some samples of flow durations may have two finite first moments.
Figure 1.24  QQ plots for the size of flows (left) and the duration of transmission (right). Quantiles of GPD(1.3, 1) against the flow size (left) and of GPD(0.85, 1) against the duration of transmission (right). The linear curves correspond to the dependencies quantile against the quantile of the same distributions.

Figure 1.25  Exceedance $e_n(u)$ against the threshold $u$ for the flow size (left) and the duration of transmission (right).

The number of finite moments was investigated by the statistic $R_n(p)$ (Figure 1.22) and by the EVI estimators (Figure 1.23). The type of the distribution was investigated by QQ plots (Figure 1.24) and the mean excess function (Figure 1.25). An example of the results of this analysis for one sample of 10 000 observations is summarized in Table 1.9 (Markovich and Kilpi, 2006). The column ‘Estimators of $\gamma$’ summarizes Table 1.8. The indication ‘$< 1$’ implies that the fractional $p$th moments with $p < 1$ may exist.

---

9 Figures 1.22–1.25 and 1.30–1.35 were compiled from Markovich and Kilpi (2006).
Table 1.9 Comparison of the ‘rough’ methods for TCP flow data.

<table>
<thead>
<tr>
<th>Number of first finite moments</th>
<th>Type of distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_n(p)$</td>
<td>Estimators of $\gamma$</td>
</tr>
<tr>
<td>Content</td>
<td>1 or 2 or 1</td>
</tr>
<tr>
<td>Transmitted</td>
<td>1 or 2 or &lt;1</td>
</tr>
<tr>
<td>SYN-FIN</td>
<td>&lt;1 or 2 or &lt;1</td>
</tr>
</tbody>
</table>

### 1.3.3 Dependence detection from univariate data

The covariance and correlation are the simplest characteristics of the strength of dependence.

**Correlation**

The correlation $\rho(X_1, X_2)$ between two r.v.s $X_1$ and $X_2$ is defined by the formula

$$\rho(X_1, X_2) = \frac{\text{cov}(X_1, X_2)}{\sqrt{\text{var}(X_1)\text{var}(X_2)}},$$

where

$$\text{cov}(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2)$$

is the covariance, and $\text{var}(x)$ is the variance.

If $X_1$ and $X_2$ are independent then $\rho(X_1, X_2) = 0$, but it is well known that the converse is false. If Gaussian r.v.s are not correlated then they are independent. However, for non-Gaussian r.v.s this may be not true. This implies that in general the covariance and correlation cannot indicate dependence. They describe the degree of the linear dependence of two r.v.s. The value $\rho(X_1, X_2)$ lies between $-1$ and $1$. In particular, $\rho(X_1, X_2) = \pm 1$ if and only if $X_1$ and $X_2$ are perfectly linearly dependent, meaning that $X_2 = \alpha + \beta X_1$ almost surely, for some $\alpha \in \mathbb{R}$ and $\beta \neq 0$.

Among the tools that may be used for dependence detection, the mixing conditions of a stationary sequence should be considered.

**Definition 12 (Rosenblatt, 1956b)** The strictly stationary ergodic sequence of random vectors $X_t$ is strongly mixing with rate function $\phi_k$ for $\sigma$-fields generated by the random variables $\{X_t, t \leq 0\}$ and $\{X_t, t > k\}$, if

$$\sup_{A \in \sigma(X_t, t \leq 0), B \in \sigma(X_t, t > k)} |P(A \cap B) - P(A)P(B)| = \phi_k \to 0 \quad \text{as} \quad k \to \infty.$$ 

The rate $\phi_k$ shows how fast the dependence between the past and the future decreases.

The measure of dependence can be provided by the dependence index sequence $\beta_n$ (Section 2.3.1). However, it is difficult to estimate these dependence measures...
using statistical tools. Therefore, the sample autocorrelation function is widely used in statistical analysis.

**Sample autocorrelation function**

The autocorrelation function (ACF) at lag $h$ is defined by the formula

$$\rho_X(h) = \rho(X_t, X_{t+h}) = E \left( (X_t - E(X_t))(X_{t+h} - E(X_{t+h})) \right) / \text{var}(X_t).$$

Given the stationary sample series $\{X_t, t = 0, \pm 1, \pm 2, \ldots \}$, the standard sample autocorrelation function at lag $h \in \mathbb{Z}$ is determined by

$$\rho_{n,X}(h) = \frac{\sum_{t=1}^{n-h} (X_t - \overline{X}_n)(X_{t+h} - \overline{X}_n)}{\sum_{t=1}^{n} (X_t - \overline{X}_n)^2}. \quad (1.43)$$

Here $\overline{X}_n = \frac{1}{n} \sum_{t=1}^{n} X_t$ represents the sample mean. The accuracy of $\rho_{n,X}(h)$ may be poor if the sample size $n$ is small or if $h$ is large with respect to $n$. The relevance of this estimate is determined by its rate of convergence to $\rho_X(h)$. The slower the rate is, the wider the confidence interval. When the distribution of the $X_t$ is very heavy-tailed (in the sense that $EX_t^4 = \infty$), this rate can be extremely slow (Mikosch, 2004).

Moreover, if the variance is infinite the ACF does not exist. What does the sample ACF estimate in this case and what might the confidence intervals be for this estimate? For heavy-tailed data it is better to use the modified estimate without the usual centering by the sample mean (Resnick, 2006):

$$\tilde{\rho}_{n,X}(h) = \frac{\sum_{t=1}^{n-h} X_t X_{t+h}}{\sum_{t=1}^{n} X_t^2}. \quad (1.44)$$

However, this estimate may behave in a very unpredictable way and not estimate anything reasonable if one uses the class of nonlinear processes, in the sense that this sample ACF may converge in distribution to a nondegenerate r.v. depending on $h$. For linear processes it converges in distribution to a constant depending on $h$ (Davis and Resnick, 1985).

**Confidence intervals of the sample ACF**

The simplest case is provided by the linear processes.

The causal autoregressive moving average (ARMA) process\(^{10}\) $X_t$ has the representation

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}, \quad t \in \mathbb{Z}. \quad (1.45)$$

\(^{10}\) The ARMA($p,q$) model has the form $X_t = \sum_{j=0}^{p} \theta_j X_{t-j} + \sum_{j=0}^{q} \psi_j Z_{t-j}, \ t = 1, \ldots, n$, with real coefficients $\psi_j$, $\theta_j$. The MA($q$) model has the form $X_t = \sum_{j=0}^{q} \psi_j Z_{t-j}$.\]
DEFINITIONS AND ROUGH DETECTION OF TAIL HEAVINESS

where $Z_t$ is an i.i.d. noise sequence and $\{\psi_j\}$ is a sequence of real numbers depending on the tails of the distribution of $Z_t$ and providing the convergence of random series in (1.45); see Brockwell and Davis (1991).

It is known that under certain conditions (namely, linearity of the underlying process and vanishing of its fourth-order cumulants) $\rho_{n,X}(i)$ has asymptotic joint normal distribution with mean $\rho_X(i)$ and variance $\text{var}(\rho_{n,X}(i)) = c_{ii}/n$, where

$$c_{ii} = \sum_{k=-\infty}^{\infty} \left( \rho_X^2(k+i) + \rho_X(k-i)\rho_X(k+i) + 2\rho_X^2(i)\rho_X^2(k) - 4\rho_X(i)\rho_X(k)\rho_X(k+i) \right),$$

(1.46)
as $n \to \infty$ (Brockwell and Davis, 1991). Bartlett’s formula (1.46) allows us to check the hypothesis $\rho_{n,X}(i) = 0$. Rejection of this hypothesis implies a significant correlation between the underlying quantities. All stationary ARMA models driven by i.i.d. $\{Z_t\}$ having zero mean and finite variance satisfy the conditions of Bartlett’s formula.

**Example 3** For i.i.d. white noise $Z_t$ we have $\rho_Z(0) = 1$ and $\rho_Z(i) = 0$ for $i \neq 0$ (since $Z_t$ and $Z_{t+i}$ are independent) and $\text{var}(\rho_{n,Z}(i)) = 1/n$ by (1.46).

This implies that for ARMA process driven by such a noise $Z$, the sample ACF is approximately normally distributed with mean zero and variance $1/n$ for sufficiently large $n$. The latter provides 95% confidence interval with the bounds $\pm 1.96/\sqrt{n}$ for the sample ACF. The hypothesis $\rho_{n,X}(i) = 0$ is accepted if $\rho_{n,X}(i)$ falls within this interval.

The limit behavior of $\tilde{\rho}_{n,X}(h)$ for ARMA processes with i.i.d. regularly varying noise and tail index $0 < \alpha < 2$ was studied in Davis and Resnick (1985) and Resnick (2006). It was found that $\tilde{\rho}_{n,X}(h)$ estimates the quantity $\sum_j \psi_j \psi_{j+h}/\sum_j \psi_j^2$ in the case of infinite variance of the process when the ACF does not exist. What is remarkable is that the latter quantity represents the autocorrelation $\text{cov}(X_0, X_h)$ in the case of a finite variance. This result leads to the illusion that the heavy-tailed sample ACF $\tilde{\rho}_{n,X}(h)$ can be applied to heavy-tailed processes without a problem. For practical purposes it is recommended to use $\tilde{\rho}_{n,X}(h)$ if $\alpha < 1$ and the classical sample ACF $\rho_{n,X}(h)$ if $1 < \alpha < 2$ (Resnick, 2006, p. 349). Unfortunately, the calculation of confidence intervals in both cases is not easy.

In Mikosch (2004) the nonlinear GARCH($p, q$) process (generalized autoregressive conditionally heteroscedastic) $X_t$ was investigated. It was concluded that if the marginal distribution of the time series is very heavy-tailed, that is, the fourth moment is infinite, then the central limit theorem with Gaussian limit breaks down and the asymptotic normal confidence bounds for the sample ACF are not applicable anymore.

In practice, we do not know the model of the underlying process $X_t$. To draw further conclusions, we can estimate its tail index and, hence, the number of finite moments.
We refer to Mikosch (2004) and Resnick (1997, 2006) for an extended survey on the relation between the tail behavior and the dependence structure.

Testing of long-range dependence

Long-range dependence means that there is dependence in a time series over an unusually long period of time. There are various definitions of long-range dependence for (a second-order) stationary process\(^\text{11}\) \((X_t)\): \((X_t)\) is long-range dependent if

\[
\sum_{h=0}^{\infty} |\rho_X(h)| = \infty,
\]

where \(\rho_X(h)\) is the ACF at lag \(h \in \mathbb{Z}\), and short-range dependent otherwise. Property (1.47) implies that even though the \(\rho_X(h)\) are individually small for large lags, their cumulative effect is important. In particular, one can assume that for some constant \(c_\rho > 0\),

\[
\rho_X(h) \sim c_\rho h^{2(H-1)} \quad \text{for large } h \text{ and some } H \in (0.5, 1),
\]

(in this case (1.47) holds). The constant \(H \in (0.5, 1)\) is called the Hurst parameter – for its estimation see Beran (1994), Willinger et al. (1995) and Kettani and Gubner (2002). The closer \(H\) is to 1, the slower is the convergence of \(\rho_X(h)\) to zero as \(h \to \infty\), that is, the longer is the range of dependence in the time series.

To detect long-range dependence using statistical procedures one replaces \(\rho_X(h)\) by the sample ACF \(\rho_{n,X}(h)\). The long-range dependence effect is typical of long time series, with several thousand points. Then one can look at lags 250, 300, 350, etc.

If on graphing the sample heavy-tailed ACF \(\tilde{\rho}_{n,X}(h)\) one finds only small values, then it may be possible to model the data as i.i.d. If the sample ACF is small beyond lag \(q\), then there is some evidence that the moving average process MA\((q)\) may be an appropriate model (Resnick, 1997). Standard short-range dependent data sets would show a sample ACF dying after only a few lags and then remaining within the 95% Gaussian confidence window \(\pm 1.96/\sqrt{n}\).

The analysis above (see Tables 1.7, 1.9) shows that the Web and TCP flow data considered are heavy-tailed with possibly infinite variance. Therefore, the application of formula (1.44) is relevant. For a comparison we represent both estimates of the ACF for our Web data in Figures 1.26 and 1.27 and for our TCP flow data in Figure 1.28.

The sample ACFs of the Web data are small in absolute value at all lags (possible exceptions are several first lags that strag outside the 95% confidence interval). The ACFs of the i.r.t set visibly decrease after some lag. The latter tendencies are not quite so strong for the s.r. The s.s.s. and d.s.s. may be independent.

---

\(^{11}\) The process \(X_t\) is called second-order stationary if its mean \(E(X_t)\) does not depend on \(t\) and if the auto-covariance function depends on \(t\) and \(t + k\) only through their difference \(k\).
Both the ACFs of the TCP flow sizes are negligible at all lags apart from one. One may suppose that the TCP flow sizes are independent.

The ACFs of the TCP flow durations have small values except for three lags. One can recognize at least three clusters in the ACF plot that may indicate dependence (Mikosch, 2004).

Estimates of the Hurst parameter for Web traffic and a subsample \((n = 1000)\) of the TCP flow data are presented in Table 1.10. The method proposed in Kettani
DEFINITIONS AND ROUGH DETECTION OF TAIL HEAVINESS

Figure 1.27  Estimates of sample heavy-tailed ACF (1.44) and sample ACF (1.43) for the i.r.t. (first two plots left) and s.r. data sets (last two plots right). The dotted horizontal lines indicate 95% asymptotic confidence bounds \( \pm 1.96/\sqrt{n} \) corresponding to the ACF of i.i.d. Gaussian r.v.s.

and Gubner (2002) under the assumption that the process is exactly second-order self-similar\(^{12}\) is used. It has a simple formula:

\[
\hat{H}_n = 0.5 \left( 1 + \log_2 (1 + \rho_{n,X}(1)) \right).
\]

The general investigation allows to suppose that all data sets are heavy-tailed and not LRD.

\(^{12}\) The process \( X_t \) is called exactly second-order self-similar with Hurst parameter \( 0 < H < 1 \) if \( \rho_X(h) = \frac{1}{2} \left( |h+1|^{2H} - 2|h|^{2H} + |h-1|^{2H} \right) \).
Figure 1.28 Estimates of sample heavy-tailed ACF (1.44) and sample ACF (1.43) for one subsample ($n = 1000$) of the TCP flow sizes (first two plots left), and durations (last two plots right). The dotted horizontal lines indicate 95% asymptotic confidence bounds $±1.96/\sqrt{n}$.

Table 1.10 Hurst parameter estimation for Web traffic and TCP flow data.

<table>
<thead>
<tr>
<th>Data</th>
<th>s.s.s.</th>
<th>d.s.s.</th>
<th>i.r.t.</th>
<th>s.r.</th>
<th>TCP flow size</th>
<th>TCP flow duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{H}_n$</td>
<td>0.493</td>
<td>0.488</td>
<td>0.508</td>
<td>0.507</td>
<td>0.498</td>
<td>0.506</td>
</tr>
</tbody>
</table>
1.3.4 Dependence detection from bivariate data

There are several ways to detect and measure the dependence among TCP flow sizes (r.v. X) and duration of transmissions (r.v. Y). Two distribution-free measures, Kendall’s τ and Spearman’s ρ, are available. Other coefficients are given in Beirlant et al. (2004) and Weissman (2005). Importantly, all these measures can be represented by means of the so-called Pickands dependence function (Capéraà et al., 1997; Beirlant et al., 2004; Weissman, 2005). We shall briefly determine this function and apply some of its estimators to TCP flow data.

Let \((X_1, Y_1), \ldots, (X_n, Y_n)\) be a bivariate i.i.d. random sample with bivariate EVD \(G(x, y)\). Similarly to the univariate case (Section 1.1), there exist normalizing constants \(a_{j,n} > 0\) and \(b_{j,n} \in R, j = 1, 2\), such that as \(n \to \infty\),

\[
P\left(\frac{M_{1,n} - b_{1,n}}{a_{1,n}} \leq x, \frac{M_{2,n} - b_{2,n}}{a_{2,n}} \leq y\right) = F^n (a_{1,n}x + b_{1,n}, a_{2,n}y + b_{2,n}) \to G(x, y),
\]

where \(M_{1,n} = \max (X_1, \ldots, X_n)\), \(M_{2,n} = \max (Y_1, \ldots, Y_n)\) are the componentwise maxima (Fougères, 2004). Note that the vector \((M_{1,n}, M_{2,n})\) will in general not be present in the original data.

Let \(F_1(x)\) and \(F_2(y)\) be the DFs of \(X\) and \(Y\), respectively. One can easily find the link between \(G(x, y)\), a copula, and the dependence function: \(G(x, y)\), with continuous univariate margins \(G_1(x)\) and \(G_2(y)\),\(^{13}\) may be uniquely represented as a copula \(C\), \(G(x, y) = C(G_1(x), G_2(y))\), by Sklar’s theorem (Nelsen, 1998). A copula is determined in terms of the Pickands dependence function \(A(t)\), \(t \in [0, 1]\) (Pickands, 1981), by

\[
C(u, v) = P[G_1(X) \leq u, G_2(Y) \leq v] = \exp \left( \log (uv) A \left( \frac{\log (v)}{\log (uv)} \right) \right).
\]

A remarkable feature of this representation is that \(C\) depends only on \(A(x)\), but not on the margins. Hence, \(G(x, y)\) may be determined by margins \(G_1(x)\) and \(G_2(y)\) by the representation

\[
G(x, y) = \exp \left( \log (G_1(x)G_2(y)) A \left( \frac{\log (G_2(y))}{\log (G_1(x)G_2(y))} \right) \right);
\]

see Beirlant et al. (2004).

In the bivariate case the function \(A(t)\) satisfies two properties:

1. \((1 - t) \vee t \leq A(t) \leq 1, t \in [0, 1]\), that is, \(A(0) = A(1) = 1\) and lies inside the triangle determined by points \((0, 1), (1, 1)\) and \((0.5, 0.5)\);

2. \(A(t)\) is convex.

\(^{13}\) \(G_j(x), j = 1, 2\) is in itself a univariate extreme value DF and \(F_j(x)\) is in its domain of attraction (Beirlant et al., 2004, p. 254).
Cases \( A(t) \equiv 1 \) and \( A(t) = (1 - t) \vee t \) correspond to total independence and total dependence, respectively.

Let the random pair \((X^*, Y^*)\) have DF \( G(x, y) \). In practice, \((X^*, Y^*)\) are componentwise maxima over large blocks of data. It is convenient to transform initial random pairs \((X_i^*, Y_i^*)\) to new pairs \((\xi_i, \eta_i)\) in such a way the margins are all the same.

**Examples**

1. The transformation \( \xi_i = -1/\log(G_1(X_i^*)) \), \( \eta_i = -1/\log(G_2(Y_i^*)) \), leads to Fréchet distributed r.v.s \( \xi \) and \( \eta \).

2. The transformation \( \xi_i = -\log(G_1(X_i^*)) \), \( \eta_i = -\log(G_2(Y_i^*)) \), leads to exponential distributed r.v.s \( \xi \) and \( \eta \).

Pickands (1981) has shown that a bivariate DF \( G(x, y) \) is an extreme value DF with unit Fréchet margins if and only if

\[
G(x, y) = P\{\xi \leq x, \eta \leq y\} = \exp\left( - \left( \frac{1}{x} + \frac{1}{y} \right) A\left( \frac{y}{x+y} \right) \right). 
\]

In the case of exponential margins the joint survival function of the pair \((\xi, \eta)\) is given by

\[
P\{\xi > x, \eta > y\} = \exp\left( - (x+y) A\left( \frac{y}{x+y} \right) \right), \quad x \geq 0, y \geq 0. \tag{1.49}
\]

The latter equation is helpful in constructing the estimators of \( A(t) \). Many of these are presented in Beirlant et al. (2004).

**Problems with estimators of \( A(t) \)**

1. The estimators are not convex. They may be improved by taking a convex hull.

2. The margins \( G_1(x) \) and \( G_2(x) \) are unknown. One has to replace them by their estimates \( \hat{G}_1(x) \) and \( \hat{G}_2(x) \), for example by empirical DFs constructed from componentwise maxima over blocks of data. The number of these maxima may be very moderate, which may lead to poor accuracy of an empirical DF.

3. The componentwise maxima may be not observable together, that is, some of them are not present in the sample. Under certain conditions one can estimate a bivariate EVD from an initial random sample \((X, Y)\); see Beirlant et al. (2004, Section 9.4).
In order to approximate the $F_R(x)$ we can replace $F(y, z)$ in (1.42) by the estimated DF of componentwise maxima

$$\hat{F}(x, y) \approx \exp \left( \log \left( \hat{G}_1(x) \hat{G}_2(y) \right) \hat{A} \left( \frac{\log \left( \hat{G}_2(y) \right)}{\log \left( \hat{G}_1(x) \hat{G}_2(y) \right)} \right) \right), \quad (1.50)$$

(Beirlant et al., 2004, p. 326). We obtain

$$F_R(x) \approx \int_0^\infty \int_0^{z_{x'}} d \left( \exp \left( \log \left( \hat{G}_1(y) \hat{G}_2(z) \right) \hat{A} \left( \frac{\log \left( \hat{G}_2(z) \right)}{\log \left( \hat{G}_1(y) \hat{G}_2(z) \right)} \right) \right) \right).$$

The simulation study provided in Hall and Tajvidi (2000) indicates that the best estimators of $A(t)$ are $A_n^C(t)$ from Capéraà et al. (1997) and $A_n^{HT}(t)$ from Hall and Tajvidi (2000):

$$\hat{A}_n^{HT}(t) = \left( \frac{1}{n} \sum_{i=1}^n \min \left( \frac{\hat{\xi}_i / \bar{\xi}_n}{1 - t}, \frac{\hat{\eta}_i / \bar{\eta}_n}{t} \right) \right)^{-1},$$

$$\log \hat{A}_n^C(t) = \frac{1}{n} \sum_{i=1}^n \log \max \left( t \hat{\xi}_i, (1 - t) \hat{\eta}_i \right) - t \frac{1}{n} \sum_{i=1}^n \log \hat{\xi}_i - (1 - t) \frac{1}{n} \sum_{i=1}^n \log \hat{\eta}_i.$$

Here $\hat{\xi}_i = - \log \hat{G}_1(X_i)$ and $\hat{\eta}_i = - \log \hat{G}_2(Y_i), i = 1, \ldots, n$, $\bar{\xi}_n = n^{-1} \sum_{i=1}^n \hat{\xi}_i$, $\bar{\eta}_n = n^{-1} \sum_{i=1}^n \hat{\eta}_i$.

### 1.3.5 Bivariate analysis of TCP flow data

First, we have to check the dependence between the pairs $(S_1, D_1), \ldots, (S_n, D_n)$ to apply (1.48) and hence (1.50). For this purpose, we can calculate the ACF of the r.v.s $r_i = \sqrt{S_i^2 + D_i^2}, i = 1, \ldots, n$ (Figure 1.29). The sample ACF of $\{r_i\}$ is small in absolute value at all lags (possible exceptions are two lags that stray outside the 95% confidence interval). One may suppose that the size–duration pairs are independent.

Both $A$-estimators were applied to TCP flow size and duration data. The 61 pairs $(M_{S,m}^l, M_{D,m}^l)$ of block maxima of TCP flow size $S$ and duration $D$ (for block $j$) were used to estimate the unknown DFs $G_1(x)$ and $G_2(x)$. These maxima were selected from the groups of data with similar statistical properties which correspond to the daily pattern. We also considered a larger sample of block maxima corresponding to $m = 1000$ points in each group. The size of the latter sample is $n = 610$ (Figure 1.30).

Some pairs $(M_{S,m}^l, M_{D,m}^l)$ are not present in the initial data. These artificial pairs influence the estimation of the $A$-function because $\xi_i$ and $\eta_i$ are included in (1.51) with the same number $i$. The exclusion influences the trade-off between the bias and variance of the estimation. Here, we do not exclude these pairs from consideration.
Figure 1.29 Estimate of sample ACF (1.43) for \( r_i, i = 1, \ldots, n, \ n = 1000 \) with 95% confidence bounds \( \pm 1.96/\sqrt{n} \).

Figure 1.30 Scatter plot of pairs of block maxima \( (M_{s,m}^j, M_{D,m}^j), \ j = 1, \ldots, 610, \) when the block size is \( m = 1000 \). The pairs that are presented in the initial sample are marked by dots, whereas the pairs that are not presented (e.g., the maximal size in the group does not necessarily correspond to the maximal duration in this group) are marked by circles. Lines \( D = S/384 \) and \( D = S/42 \) indicate 384 kB/s (EDGE) and 42 kB/s (GPRS) access rates, respectively.
in order to retain a larger sample. The accuracy of the estimation is more sensitive to the number of blocks than to any other parameter.

Before estimating the DFs it is important to be sure that the block maxima are independent. Then the estimation is easier. The ACFs of both maxima samples of size 61 and size 610 allow us to suppose their independence (Figures 1.31 and 1.32).

One can then find parametric models of the unknown margins $G_1(x)$ and $G_2(x)$ and evaluate their parameters by the ML method. The ‘blocks’ method is sensitive to the size of the block. The larger the number of blocks (the smaller the size of blocks), the lower the variance and the greater the bias of the parameter estimates; some trade-off between the two is thus required. The number of blocks is connected to the dependence of the data. Roughly speaking, the size of blocks should be

![Figure 1.31](image1.png)  
**Figure 1.31** Estimates of sample ACF (1.43) for maxima samples of size 61 corresponding to TCP flow sizes (left) and durations (right). The dotted horizontal lines indicate 95% asymptotic confidence bounds $\pm 1.96/\sqrt{n}$.

![Figure 1.32](image2.png)  
**Figure 1.32** Estimates of sample ACF (1.43) and sample heavy-tailed ACF (1.44) for maxima samples of size 610 corresponding to TCP flow sizes (left) and durations (right). The dotted horizontal lines indicate 95% asymptotic confidence bounds $\pm 1.96/\sqrt{n}$. 
Table 1.11  ML estimates of GEV parameters of TCP flow data

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Definition</th>
<th>( \gamma )</th>
<th>( \mu )</th>
<th>( \sigma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>Content</td>
<td>0.332259</td>
<td>7075.92</td>
<td>4605.53</td>
</tr>
<tr>
<td>Duration</td>
<td>SYN-FIN</td>
<td>0.10263</td>
<td>3775.8</td>
<td>2433.27</td>
</tr>
</tbody>
</table>

Figure 1.33  QQ plots of block maxima samples corresponding to TCP flow sizes (left) and durations (right).

chosen so as to give an approximately independent sample of maxima. The more dependent the data are, the larger should be the size of blocks (Leadbetter, 1983).\(^\text{14}\)

Here, the GEV (1.3) is applied as such a model in a more general form:

\[
H_\gamma(x) = \begin{cases} 
\exp\left(-\left(1 + \gamma \left(\frac{x-\mu}{\sigma}\right)^{\gamma^{-1}}\right)^{\frac{1}{\gamma}}\right), & \gamma \neq 0, \\
\exp\left(-\left(e^{-\gamma(x-\mu)}\right)\right), & \gamma = 0.
\end{cases}
\]

The ML parameter estimates\(^\text{15}\) of the GEV calculated with block maxima of size 610 for both TCP samples are summarized in Table 1.11. We check our hypothesis regarding the GEV model with QQ plots (Figure 1.33). These evidently show not quite linear behavior. The GEV model does not accurately fit our data. Nevertheless, such a parametric model is convenient for calculating the inverse

---

\(^{14}\) The weakness of the ‘blocks’ method is its poor accuracy of approximation. The ‘runs’ method initiated by Newell (1964) seems to provide better approximation. The idea of this method is to construct blocks of unequal size by establishing a sequence of thresholds. An observation is assigned to a cluster that is bounded by two neighboring threshold values.

\(^{15}\) The parameters can be estimated by the method of probability-weighted moments (McNeil et al., 2005).
functions $\hat{G}_1^{-1}(x)$ and $\hat{G}_2^{-1}(x)$ in formula (1.54) for bivariate quantiles. We use this model to calculate the estimates $\hat{A}_{HT}^{n}(t)$ and $\hat{A}_{C}^{n}(t)$ of the $A$-function (Figure 1.34). The convex hull of these estimates is required to further apply the quantiles formula. Both estimators are situated under the upper boundary of the triangle that indicates the dependence of TCP flow size and duration (Markovich and Kilpi, 2006).

Since the size of the block maxima sample is moderate, a further improvement may be achieved by estimating the DFs $G_1(x)$ and $G_2(x)$ by means of the combined estimator. This estimator is a mixture of a smoothed empirical DF within the range of the sample and a parametric model (e.g., GEV) in the tail. This estimator fits the tail domain better than an empirical DF. At the same time, it may fit the DF within the range of the sample better than the GEV model, especially in the case of mixtures of distributions.

Both estimates of $A(t)$ allow us to get the estimate $\hat{G}(x, y)$ and construct bivariate quantile curves

$$Q(\hat{G}, p) = \{(x, y) : \hat{G}(x, y) = p\}, \quad 0 < p < 1,$$

for the TCP flow data. According to the method of Beirlant et al. (2004, p. 325), one can assume that $\hat{G}_1(x) = P^{(1-w)/\lambda(w)}$ and $\hat{G}_2(x) = P^{w/\lambda(w)}$ for some $w \in [0, 1]$ in order to get $\hat{G}(x, y) = p$. Then the quantile curve (Figure 1.35) consists of the points

$$Q(\hat{G}, p) = \left\{\left(\hat{G}_1^{-1}(p^{(1-w)/\lambda(w)}), \hat{G}_2^{-1}(p^{w/\lambda(w)})\right) : w \in [0, 1]\right\}.$$  

(1.54)
Figure 1.35 Estimated quantile curves of TCP flow data for $p \in \{0.75, 0.9, 0.95\}$ corresponding to estimator $\hat{A}_n^C(t)$: the maxima sample of size 61 (left), of size 610 (right).

Conclusions

One can draw the following conclusions from the investigation:

- The samples used in the analysis are of moderate size.
- Size $S$ and duration $D$ are heavy-tailed with probably infinite second moment.
- Their distributions are complicated in the sense that they do not belong to any known parametric models.
- Estimates of the Pickands dependence function show that $S$ and $D$ are dependent.
- Bivariate quantile curves show that the bivariate EVD of $(S, D)$ is ‘not quite heavy-tailed’ in the sense that not many observations can be considered outliers, that is fall beyond the 97.5% quantile curve. This may be a special property of these mobile TCP data.
- Bivariate quantile curves are sensitive to (at least) the following: violations of the independence assumption within a block; estimation of parameters of margins of $G(x, y)$ and estimates of $A(t)$; and the number of componentwise maxima, or the block size.

1.4 Notes and comments

When analyzing real data, one must undertake the preliminary detection of heavy tails, as well as investigating the dependence structure of univariate data and of multivariate data. For heavy-tailed distributions and, in particular, with infinite
variance, the classical statistical methods are not adequate and flexible enough. An example is given by the sample ACF that may not represent the ACF of a heavy-tailed distribution properly.

We should distinguish between methods that are valid for independent and dependent data. For example, an empirical DF cannot be applied to dependent data. The same is true regarding the ML method that is used to estimate parameters in a parametric model of the DF.

The evaluation of the distributions of univariate data is particularly important for estimating multivariate quantiles and distributions. When the data are independent or weakly dependent one can apply traditional methods such as kernel estimators to estimate the PDF. The problem arises when the data are long-range dependent (Section 2.3).

The exploratory techniques introduced in Section 1.3.1, apart from the ratio of the maximum to the sum, the QQ plot, and the group estimator of the tail index, do not require an i.i.d. assumption on the underlying data. The interpretation of the criteria mentioned may become hazardous when they are applied to the non-i.i.d. case. However, we have applied these tools to real data despite the sometimes evident non-i.i.d. structure. The reason is that the methods mentioned, apart from the QQ plot, have an asymptotic background and their application to samples of moderate size requires additional research. As one may conclude from Tables 1.7 and 1.9, all these methods are consistent in their conclusions.

To estimate the dependence structure of two r.v.s and corresponding bivariate quantiles, one needs to evaluate the marginal DFs of maxima of both samples and the Pickands function.

1.5 Exercises

1. Generators.

Generate 100 Fréchet distributed r.v.s with DF

$$F(x) = \exp \left( -\left( \gamma x \right)^{-1/\gamma} 1 \{ x > 0 \} \right), \quad \gamma = 1.5.$$ 

To do this, generate 100 r.v.s $U_i$ uniformly distributed on $[0,1]$ and apply the transformation $X_i = (-\ln U_i)^{-\gamma}/\gamma$.

2. The ratio of the maximum to the sum.

Calculate the statistic $R_n(p)$ by formulas (1.37) and (1.38) for the samples $X^n = \{ X_1, \ldots, X_n \}$, $n \to \infty$. For practical calculations, one can take parts of the same sample, i.e., $X^i, i = 1, \ldots, n$. The sample $X^n$ may be generated using some random generator or real data. Plot the dependence $R_n(p)$ against $n$ for different $p$. Investigate this plot for large $n$ and draw conclusions regarding the number of finite moments $E|X|^p$ of the distribution.
3. QQ plot

Using the empirical or generated data \( X^n = \{X_1, \ldots, X_n\} \), construct a QQ plot, i.e., draw the dependence

\[
\left\{ \left( X_{(k)}, F^{-1} \left( \frac{n-k+1}{n+1} \right) \right) : k = 1, \ldots, n \right\},
\]

where \( X_{(1)} \geq \ldots \geq X_{(n)} \) are the order statistics of the sample \( X^n = \{X_1, \ldots, X_n\} \), and \( F^{-1} \) is an inverse of the DF \( F \).

Check different choices of \( F(x) \), e.g. normal, lognormal, exponential, and the generalized Pareto distribution (1.16).

If the QQ plot is linear for some \( F(x) \) then the underlying sample is distributed according to this \( F(x) \). Depending on whether the QQ plot is below or above the diagonal line draw conclusions whether the empirical model has a heavier tail than a parametric model or not.

4. Mean excess function.

Using the empirical or generated data \( X^n = \{X_1, \ldots, X_n\} \), calculate the empirical mean excess function by formula (1.41). Investigate the behavior of \( e_n(u) \) for large \( u \).

For heavy-tailed distributions the function \( e(u) \) tends to infinity. A linear plot \( u \rightarrow e(u) \) corresponds to a Pareto distribution, the constant \( 1/\lambda \) corresponds to an exponential distribution, and \( e(u) \) tends to 0 for light-tailed distributions.

5. Estimation of the tail index.

Using the empirical or generated data \( X^n = \{X_1, \ldots, X_n\} \), reorder the data as \( X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)} \). Calculate and compare the following estimates of the tail index of your data: Hill’s estimator (1.5) for some \( k = 1, \ldots, n-1 \); the ratio estimator (1.7) for some \( X_{(1)} < x_n < X_{(n)} \); the moment estimator (1.13); the UH estimator (1.14) and the Pickands’ estimator (1.15) for some \( k = 1, \ldots, n-1 \). Investigate the sign of the estimate and draw conclusion regarding heavy tails.

6. Choice of parameter \( k \) of Hill’s estimator by a Hill plot.

Considering a sample \( X^n = \{X_1, \ldots, X_n\} \), calculate Hill’s estimate (1.5).

Plot the dependence \( \{ (k, \hat{\gamma}^H(n, k)) , 1 \leq k \leq n-1 \} \) and then choose the estimate of \( \hat{\gamma}^H(n, k) \) from an interval in which these functions demonstrate stability.

Draw conclusions regarding the number of finite moments\(^{16} \) of the underlying distribution and the existence of heavy tails. A positive estimate \( \hat{\gamma}^H(n, k) \) may indicate heavy tails.

\[^{16}\text{For regularly varying distributions (1.4), the moments } EX^\beta \text{ are finite only if } \beta < 1/\gamma \text{ (Lemma 1).}\]
7. Repeat Exercise 6 with the group estimator (1.19), (1.28). Find an appropriate value of \( m \) from the plot \( \{(m, z_m), m_0 < m < M_0\} \) \( m_0 > 2, M_0 < n/2 \); see Section 1.2.4.

8. Investigation of Hill’s estimator (1.5).

Generate several samples distributed with the regularly varying DFs (1.4), where \( \ell(x) = 1, \ell(x) = 2, \ell(x) = \ln \ln x \) and \( \ell(x) = \ln x \), and with Weibull DF

\[
1 - F(x) = \exp(-cx^{1/\gamma}), \quad c = 1, \gamma = 2; c = 2, \gamma = 3.
\]

Calculate Hill’s estimate and investigate the influence of a slowly varying function \( \ell(x) \) on the estimate. Compare the true values of the EVI \( \gamma \) with the results of the estimation for different distributions. The estimation should be worse in the case of the Weibull distribution.

9. Repeat Exercise 8 for the group estimator.

10. Dependence detection by bivariate data.

Generate 1000 r.v.s \( Y^n = \{Y_1, \ldots, Y_n\} \) such that \( Y_i = 2X_i + 1, i = 1, \ldots, n \). Calculate the Pickands dependence functions using the estimators \( \hat{A}^{HT}_n(t) \) and \( \hat{A}^C_n(t) \) (see (1.51)). In order to do this, separate both samples \( X^n \) and \( Y^n \) into ten equal-size blocks and select the block maxima samples. Estimate the marginal DFs \( \hat{G}_1(x) \) and \( \hat{G}_2(y) \) corresponding to the block maxima of r.v.s \( X_i \) and \( Y_i \) by empirical DF and by the GEV model (1.52) using the block maxima data. Estimate the parameters \( \gamma, \mu, \) and \( \sigma \) of the GEV model by the ML method.

Plot \( \hat{A}^{HT}_n(t) \) and \( \hat{A}^C_n(t) \) separately for two estimates of the DF. Draw conclusions regarding the dependence of \( X \) and \( Y \).

11. Draw the bivariate quantiles curves for \( p \in \{0.75, 0.9, 0.95, 0.975\} \) by formula (1.54). Compare the quantile curves for both A-estimators \( \hat{A}^{HT}_n(t) \) and \( \hat{A}^C_n(t) \).

---

17 The last two \( \ell(x) \) require special generators which do not allow inversion of the DF. For instance, one can use the acceptance–rejection method (Law and Kelton, 2000).
Classical methods of probability density estimation

In this chapter the main principles of density estimation – Lebesgue’s theorem, Fisher’s scheme, $L_1$, $L_2$, $\chi^2$ approaches, the exponent method and the estimation of the PDF as solution of an ill-posed problem – are considered. The links between these approaches and examples of their application are shown. Classical methods of PDF estimation – kernel estimators, projection estimators, histogram and polygram – and their smoothing tools, among them cross-validation and the discrepancy method, are presented. We shall estimate the unknown PDF $f(x)$ from a sample $X^n = \{X_1, \ldots, X_n\}$ of i.i.d. observations of $X$, where $n$ is the sample size.

2.1 Principles of density estimation

Lebesgue’s theorem

According to Lebesgue’s theorem on densities,

$$\lim_{h \to 0} \int_{S_{xh}} \frac{f(y)}{\lambda(S_{xh})} dy = f(x)$$

for almost all $x$, where $S_{xh}$ is a closed ball with radius $h$ and center $x$, and $\lambda$ is Lebesgue’s measure. The expression under the limit can be approximated by
This estimate was proposed in Rosenblatt (1956a) and developed in Parzen (1962).

**Fisher’s scheme**

Let \( f(x, \alpha), \alpha \in \Lambda \subseteq R, x \in R^d \), be a set of PDFs containing a true PDF \( f(x, \alpha_0) \) that we wish to find. Let

\[
H_{\alpha_0} (\alpha) = - \int \ln f(x, \alpha)f(x, \alpha_0)\,dx.
\]  

(2.1)

A sequence \( f(x, \alpha_n), \alpha_n = \alpha(X_1, \ldots, X_n) \), that leads to convergence can be found in Kullback’s metric

\[
\mathcal{I}(f, f_n) = H_{\alpha_0} (\alpha_n) - H_{\alpha_0} (\alpha_0) = - \int \ln \frac{f(x, \alpha_n)}{f(x, \alpha_0)} f(x, \alpha_0)\,dx \to_{n \to \infty}^p 0.
\]

Here, \( f(x, \alpha_0) \) is unknown. Therefore, instead of the functional (2.1) Fisher (1952) decided to minimize the empirical functional

\[
H_{\text{emp}}(\alpha) = - \frac{1}{n} \sum_{i=1}^{n} \ln f(X_i, \alpha),
\]

constructed from the sample \( X^n = \{X_1, \ldots, X_n\} \) on the set \( f(x, \alpha), \alpha \in \Lambda \). This is called the maximum likelihood (ML) method.

The consistency of the ML method is provided by the convergence

\[
\inf_{\alpha \in \Lambda} \left( - \frac{1}{n} \sum_{i=1}^{n} \ln f(X_i, \alpha) \right) \to_{n \to \infty}^p - \int \ln f(x, \alpha_0)f(x, \alpha_0)\,dx
\]

for any PDF \( f(x, \alpha_0), \alpha_0 \in \Lambda \).

**\( L_1 \) approach**

The \( L_1 \) approach is natural because of two remarkable effects. First of all, the estimation error in the metric space \( L_1 \) is invariant for any monotone increasing, continuously differentiable, one-to-one transformation \( T : R^1 \to [0, 1] \) (the derivative of the inverse function \( T^{-1} \) is assumed to be continuous)\(^1\) of the data, i.e.,

\[
\int_0^\infty |f_n(x) - f(x)|\,dx = \int_0^1 |g_n(x) - g(x)|\,dx.
\]

(2.2)

\(^{1}\) The transformation \( T(x) \) may be extended to \( T : R^1 \to R^1 \).
Here, $f(x)$ and $g(x)$ are the PDFs of the r.v.s $X$ and $Y = T(X)$ respectively, and $f_n(x)$ and $g_n(x)$ are estimates of $f(x)$ and $g(x)$. Hence, the accuracy of the estimate $g_n(x)$ defines the accuracy of the estimation of $f(x)$.

Furthermore, according to Scheffé’s theorem, the $L_1$-distance between two PDFs $f$ and $f_n$ is equal to twice the maximal deviation of the probabilities of any Borel sets (‘total variation’), calculated by PDFs $f$ and $f_n$:

$$L_1(f, f_n) = \int |f_n(x) - f(x)| \, dx = 2\sup_{B \in \mathcal{B}}\left| \int_B f(x) \, dx - \int_B f_n(x) \, dx \right| = 2T(f, f_n).$$

(2.3)

In practice, we are looking for different probability functions such as the DF $F(x)$ or the tail $1 - F(x)$, rather than for the PDF itself. Therefore, the necessity of the convergence in the metric space $L_1$ is obvious (Barron et al., 1992). The connection between the $L_1$-error and Kullback’s metric is

$$\frac{1}{2} \left( \int_{\mathbb{R}^d} |f(x, \alpha_0) - f(x, \alpha_n)| \, dx \right)^2 \leq \int_{\mathbb{R}^d} f(x, \alpha_0) \ln \frac{f(x, \alpha_0)}{f(x, \alpha_n)} \, dx = \mathcal{I}(f, f_n).$$

(2.4)

The connection between the total variation and Kullback’s metric $\mathcal{I}(f, f_n)$ follows from (2.3) and (2.4):

$$2T^2(f, f_n) \leq \mathcal{I}(f, f_n).$$

$L_2$ approach

The space $L_2$ is the most convenient for constructing the estimates of the PDF and finding their accuracy with respect to the $L_2$-norm.

Projection estimators (Čencov, 1982) are obtained by the approximation of a PDF in terms of the expansion by an orthogonal basis $\varphi_j(t)$, $j = 1, \ldots, n$:

$$\int \left( f(t) - \sum_{j=1}^{n} a_j \varphi_j(t) \right)^2 \, dt \to \min_{a_j}.\]

Unfortunately, there is no connection between the distances in the spaces $L_1$ and $L_2$. Therefore, the results of the $L_2$-theory cannot be extended to $L_1$. For example, the convergence in the sense of the mean integrated squared error\(^2\) (MISE), cannot be extended to $EL_1(f, f_n)$.

$\chi^2$ approach

The $\chi^2$-distance

$$\chi^2(f, f_n) = \int_{\mathbb{R}^d} \frac{(f(x) - f_n(x))^2}{f_n(x)} \, dx$$

(2.5)

\(^2\) See (4.12).
was introduced in Györfi et al. (1998). Since
\[
\frac{(L_1(f, f_n))^2}{2} \leq I(f, f_n) \leq \chi^2(f, f_n),
\]
the convergence in Kullback’s metric is stronger than the convergence in \( L_1 \), and \( \chi^2 \)-convergence is stronger than the convergence in Kullback’s metric.

The exponent method

The main idea of the exponent method, proposed independently by Stratonovich (1969) and Čencov (1982), is to estimate the logarithm of the PDF, \( \log f(x) \), in terms of its linear expansion in some basic functions \( \phi_k(x) \), e.g., polynomials, splines, trigonometric functions.

Let \( X_1, \ldots, X_n \) be i.i.d. r.v.s with positive PDF \( f(x) \) bounded on some limited interval, for example \([0, 1]\). The exponent estimator is given by
\[
f_\theta(x) = f_0(x) \exp \left( \sum_{k=1}^{m} \theta_k \phi_k(x) - \psi_m(\theta) \right),
\]
where
\[
\psi_m(\theta) = \log \int f_0(x) \exp \left\{ \sum_{k=1}^{m} \theta_k \phi_k(x) \right\} dx,
\]
is a normalizing multiplier to make \( \int f_\theta(x) dx = 1 \), and \( f_0(x) \) is an auxiliary PDF with the same smoothing features as \( f(x) \) and independent of \( \theta = (\theta_1, \ldots, \theta_m) \) (e.g., a uniform PDF).

The PDF \( f_\theta(x) \) may be estimated by the ML method. The uniqueness of the solution follows from the convexity of the maximum likelihood function.

The maximum of the function
\[
\Gamma(\theta) = \theta \hat{\alpha} - \psi_m(\theta)
\]
is taken as an estimate of \( \theta \), where
\[
\theta \hat{\alpha} = \sum_{k=1}^{m} \theta_k \hat{\alpha}_k, \quad \hat{\alpha}_k = \frac{1}{n} \sum_{i=1}^{n} \phi_k(X_i).
\]
The coefficients of the expansion \( \theta \) are defined by the following system of equations:
\[
\frac{\int f_0(x) \exp(\sum_{i=1}^{m} \theta_i \phi_i(x)) \phi_j(x) dx}{\int f_0(x) \exp(\sum_{i=1}^{m} \theta_i \phi_i(x)) dx} = \hat{\alpha}_j, \quad j = 1, \ldots, m.
\]
The advantages of the exponent method are as follows:

• It is simple.
• The final estimate \( f_{\hat{\theta}(\hat{\alpha})}(x) \) is positive and integrates to 1.
• The estimate \( \hat{f}_\theta(\hat{\alpha}) \) minimizes the Kullback–Leibler distance.

• The method can be applied to heavy-tailed PDFs.

However, one cannot apply the exponent method directly to a heavy-tailed PDF since the method requires the PDF \( f(x) \) to be bounded away from zero (\( f(x) \) is strongly positive) on a bounded interval. The log-density should be bounded and its derivative integrable for the convergence of a basis function expansion. If \( \log f(x) \) belongs to Sobolev’s space \( W^r_2[0,1], \ r \geq 1 \), that is, \( (\log f(x))^{(r-1)} \) is absolutely continuous and \( \int_0^1 ((\log f(x))^{(r)})^2 \, dx \) is finite, this condition holds.

There are a number of problems with the exponent method:

• The accuracy of the estimate depends on the choice of \( m \). It is proved in Barron and Sheu (1991) that the rate of convergence of \( \hat{f}_\theta(x) \) to \( f(x) \) in the Kullback–Leibler metric is of order \( O(n^{-2r/(2r+1)}) \) if \( m \sim n^{1/(2r+1)} \) and \( \log f(x) \) belongs to Sobolev’s space. This result is extended in Koo and Kim (1996) to a wavelet basis and PDFs such as \( \log f(x) \) belonging to the Hölder and Sobolev spaces.

• The method requires a preliminary transformation of the data to a bounded interval if the PDF has long tails.

• The transformation must be constructed in such a way as to avoid unbounded values of \( \log f(x) \) in the neighborhood of the boundaries of the interval.

The exponential families (2.6) arise in many problems of mathematical statistics and statistical physics. Čencov (1982) gives the connection between \( \psi_m(\theta) \) and the Kullback–Leibler metric \( \psi_m(\theta) = \int f_0(x) \log(f_0(x)/f_\theta(x)) \, dx \), under assumptions \( \int \phi_k(x)f_0(x) \, dx = 0 \) on the basic functions. Related works on estimates of PDFs based on exponential families include Koo and Chung (1998) and Azoury and Warmuth (2001).

Density estimation from observations of the log-density

Sometimes it is easier to estimate \( \ln f(x) \) rather than \( f(x) \), particularly if \( f(x) \) is long-tailed, for example, an exponential, Weibull or two-mode Gaussian distribution. In order to estimate \( \ln f(x) \) one can take a sample of observations of the logarithm of the PDF and apply a regression method (Dubov, 1998).

Let \( X^n = \{X_1, X_2, \ldots, X_n\} \) be a sample of observations of some r.v. with continuous DF \( F(x) \), and

\[
X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}
\]

be order statistics corresponding to \( X^n \).

It is known that the square \( S_i \) under the curve \( f(x) \) in the interval \( (X_{(i)}, X_{(i+1)}) \) is a beta distributed r.v. with PDF

\[
p_{S_i}(S) = n (1 - S)^{n-1},
\]
if \( 0 < S < 1 \) holds. Obviously,
\[
S_i = \int_{x(i)}^{x(i+1)} f(x) \, dx = (X_{i+1} - X_i) \, f(z_i), \quad z_i \in (X_i, X_{i+1}).
\]

Let us construct a new sample of observations of the function \( \ln f(x) \) at the points \( x^v_i = (X_i + X_{i+1})/2 \):
\[
R = \{x^v_i, \ln f^v_i\}_{i=1}^{n_R}, \quad n_R = n - 1,
\]
where
\[
f^v_i = \frac{\exp(\alpha)}{X_{i+1} - X_i}, \quad \alpha = E(\ln S_i) = -\sum_{k=1}^{n} \frac{1}{k}.
\]
Since
\[
E(\ln f^v_i) = \ln f(z_i)
\]
and the variance of \( \ln f^v_i \) is\(^3\)
\[
\sigma^2 = \text{var}(\ln f^v_i) = \sum_{k=1}^{n} \frac{1}{k^2} < \frac{\pi^2}{6}.
\]
the observations \( R \) satisfy the additive model
\[
y_i = \varphi(z_i) + \xi_i,
\]
where \( y_i = \ln f^v_i, \varphi(z_i) = \ln f(z_i), \quad E(\xi_i) = 0, \quad E(\xi_i^2) = \sigma^2, \quad E(\xi_i \xi_j)/\sigma^2 = r_0 \) for \( i \neq j, \quad r_0 = 1 - \pi^2/(6\sigma^2) \). Thus, one can approximate \( \ln f(x) \) based on a new sample by some method of dependency reconstruction (Vapnik, 1982). Finally, the function \( \hat{f}^*(x) = \exp \varphi^*_\text{opt}(x) \) is obtained, where \( \varphi^*_\text{opt}(x) \) is the estimate of \( \ln f(x) \).

**The estimation of the density as a solution of an ill-posed problem**

The estimation of the unknown PDF using i.i.d. observations \( X_1, \ldots, X_n \) may be considered as an ill-posed problem of the approximate solution of Fredholm’s integral equation of the first kind,
\[
\int_{-\infty}^{\infty} \theta(x - t) f(t) \, dt = F(x),
\]
\(^3\) This follows from
\[
\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}.
\]
where
\[
\theta(x) = \begin{cases} 
1, & x \geq 0, \\
0, & x < 0. 
\end{cases}
\]

The theoretical DF is unknown, but an empirical estimate of it is given, for example, in the form
\[
F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \theta(x - X_i). 
\] (2.9)

Equation (2.8) may be represented in operator form as
\[
Af = F, \quad f \in U, \quad F \in V, 
\] (2.10)
where \(U\) and \(V\) are normed spaces, and \(A\) is a linear one-to-one integral operator from \(U\) to \(V\).

**Definition 13** The problem of the solution of an operator equation is called correct by Hadamard if the solution exists, is unique, and is stable. The problem is ill-posed if the solution does not satisfy at least one of these three conditions (Tikhonov and Arsenin, 1977).

**Regularization method**

In order to find the solution of (2.10) by the regularization method, one minimizes the functional
\[
R_{\gamma}(f, F_n) = \|Af - F_n\|_V^2 + \gamma_n \Omega(f),
\]
in a set \(D\) of functions \(f\) from \(U\). Here, \(\gamma_n > 0\) is a regularization parameter, and \(\Omega(f)\) is a stabilizing functional that satisfies the following conditions:

1. \(\Omega(f)\) is defined on the set \(D\).
2. \(\Omega(f)\) assumes real nonnegative values and is lower semi-continuous on \(D\).
3. All sets \(M_c = \{f : \Omega(f) \leq c\}\) are compact in \(U\).

Vapnik and Stepanyuk (1979) proved that the application of the regularization method allows one to obtain regularized solutions \(f_n^\gamma(x)\) that converge to the true PDF \(f(x)\) with probability one as the sample size \(n\) goes to infinity and the regularization parameter \(\gamma_n\) satisfies the following conditions:
\[
\gamma_n \to 0 \quad \text{as } n \to \infty, \\
\sum_{n=1}^{\infty} \exp(-\mu n \gamma_n) < \infty
\]
for at least one \( \mu > 0 \). The rate of this convergence may be obtained under additional assumptions regarding the smoothness of \( f(x) \).

We consider PDFs defined on the interval \([a, b]\). One may suppose that \( f(x) \) has \( m \) continuous derivatives or that \( f(x) \) satisfies the Lipschitz condition

\[
|f(t) - f(\tau)| \leq K|t - \tau|^{\mu}, \quad 0 < K < \infty, \quad 0 < \mu \leq 1, \quad t, \tau \in [a, b].
\]

Under these assumptions the uniform convergence rate for regularized PDF estimates is proved.

**Theorem 2** (Stefanyuk, 1980). An asymptotic rate of convergence in the metric \( C[a, b] \) of the regularized estimates \( f_n^{\beta}(x) \) to the true PDF \( f(x) \) is determined by the expression

\[
P\left\{ \lim_{n \to \infty} \left( \frac{n}{\ln \ln n} \right)^{\beta/(2(1+\beta))} \sup_x |f_n^{\beta}(x) - f(x)| \leq c \right\} = 1,
\]

if \( \gamma_n = (\ln \ln n) / n \). Here, \( 0 < c < \infty \) is a constant, and \( \beta \) depends on the smoothness of \( f(x) \): \( \beta = m \) if \( f(x) \) has \( m \) continuous derivatives in the interval \([a, b]\), and \( \beta = \mu \) if \( f(x) \) satisfies the Lipschitz condition at \([a, b]\).

The latter rate of the convergence is better than the rate \( (\ln n/n)^{\beta/(1+2\beta)} \) proved under similar conditions in Reiss (1975) for kernel estimates (see Section 2.2.1) if the bandwidth \( h = (\ln n/n)^{1/(1+2\beta)} \) is used.

The assignment of different norms \( \| \cdot \|_\nu \) and stabilizers \( \Omega(f) \) allows different PDF estimators to be obtained, such as kernel, projection, histograms, and spline estimators. Theoretical background on the statistical regularization method and a numerical solution of ill-posed inverse problems can be found in Sections 7.3 and 7.5.

**Example 4** (Vapnik, 1982). The minimization of the regularization functional

\[
R_\gamma(f, F_n) = \int_{-\infty}^{+\infty} \left( \int_{-\infty}^{+\infty} \theta(x - \tau)f(\tau)d\tau - F_n(x) \right)^2 dx + \gamma_n \int_{-\infty}^{+\infty} f^2(\tau)d\tau
\]

with respect to \( f(x) \) for a fixed regularization parameter \( \gamma_n > 0 \) gives the Parzen–Rosenblatt kernel estimator (Section 2.2.1)

\[
\hat{f}_h(x) = (2n\gamma_n^{-1/2})^{-1} \sum_{i=1}^{n} K \left( \frac{x - X_i}{\gamma_n^{1/2}} \right) \tag{2.11}
\]

where \( K(x) = \exp(-|x|) \).

**Example 5** (Vapnik et al., 1992). Let \( X^n = \{X_1, \ldots, X_n\} \) be i.i.d. r.v.s with PDF \( f(x) \) that has compact support \([0, 1]\). In addition, it is assumed that the kth

---

4 In order that all integrals exist, the densities \( f(x) \in L_2(-\infty, +\infty) \) are considered such that the function \( r(x) = \int_{-\infty}^{+\infty} \theta(x - \tau)f(\tau)d\tau - F_n(x) \) belongs to \( L_2(-\infty, +\infty) \). These are, for instance, densities bounded away from zero just in the sets of limited measure.
derivative \((k \geq 1)\) of the PDF \(f(x)\) exists and has bounded variation on \([0, 1]\), and \(f(x)\) may be extended in an even manner to \([-1, 1]\) and then periodically to the entire real axis so that its \(k\)th derivative will be continuous. The set of PDFs satisfying these conditions, will be denoted by \(\varphi\). According to Fikhtengol’ts (1965), any function from \(\varphi\) has a Fourier series of the form

\[
f(x) = 1 + \sum_{j=1}^{\infty} \theta_j \varphi_j(x),
\]

uniformly convergent to it on \([0, 1]\). Here,

\[
\theta_j = 2 \int_0^1 f(x) \varphi_j(x) dx, \quad j = 1, 2, \ldots
\]

\(
\{\varphi_j(x) = \cos(\pi j x), j = 0, 1, 2, \ldots\}, is an orthogonal basis in \(L_2[0, 1]\).

The minimum of the regularization functional

\[
R_\gamma(f, F_n) = \int_0^1 \left( \int_0^x f(t) dt - F_n(x) \right)^2 dx + \gamma_n^{2k+2} \int_0^1 (f^{(k)}(x))^2 dx
\]

by \(f(x) \in \varphi\) with respect to \(\theta_j\) for a fixed regularization parameter \(\gamma_n > 0\) is given by the smoothed projection estimator

\[
\hat{f}^{pe}_{\gamma_n}(x, X^n) = 1 + \sum_{j=1}^{\infty} \lambda_j a_j \varphi_j(x),
\]

where \(\lambda_j = \lambda_j(n, \gamma) = (1 + (\pi j \gamma)^{2k+2})^{-1}\) and \(a_j = (2/n) \sum_{i=1}^{n} \varphi_j(X_i)\).

**Example 6** (Markovich, 1989). Let \(X^n = \{X_1, \ldots, X_n\}\) be i.i.d. r.v.s with PDF \(f(x)\) that has compact support \([A, B]\) \((-\infty < A, B < \infty)\) and is square integrable, i.e. \(f(x) \in L_2[A, B]\). According to Fikhtengol’ts (1965), the function \(f(x) \in L_2[A, B]\) has a Fourier series of the form

\[
f(x) = a_0/2 + \sum_{j=1}^{\infty} \left( a_j \cos \left( j(x - \Delta) \pi/d \right) + b_j \sin \left( j(x - \Delta) \pi/d \right) \right),
\]

uniformly convergent to \(f(x)\) in \(L_2[A, B]\). Here,

\[
a_j = (1/d) \int_A^B f(x) \cos \left( j(x - \Delta) \pi/d \right) dx, \quad j = 0, 1, 2, \ldots,
\]

\[
b_j = (1/d) \int_A^B f(x) \sin \left( j(x - \Delta) \pi/d \right) dx, \quad j = 1, 2, \ldots,
\]

\[
\Delta = (A + B) / 2, \quad d = (B - A) / 2.
\]

The minimum of the regularization functional

\[
R_\gamma(f, F_n) = \int_A^B \left( \int_A^x f(t) dt - F_n(x) \right)^2 dx + \gamma_n \int_A^B f^2(x) dx
\]
by \( f(x) \) with respect to \( a_j \) and \( b_j \) is given by the smoothed projection estimator

\[
\hat{f}_\gamma^{pr}(x) = f_1(x) + f_2(x),
\]

(2.17)

where

\[
f_1(x) = \frac{1}{d} \left[ 0.5 + \sum_{j=1}^{\infty} \left[ \frac{\sin j}{1 + \gamma_n(\pi j/d)^2} \sin \left( \frac{j(x - \Delta)\pi}{d} \right) \right. \right.
\]
\[
\left. + \left. \frac{\cos j}{1 + \gamma_n(\pi j/d)^2} \cos \left( \frac{j(x - \Delta)\pi}{d} \right) \right] \right],
\]

\[
f_2(x) = \frac{1}{2n\sqrt{\gamma_n}} \exp \left( (x - \Delta)/\sqrt{\gamma_n} \right) - \exp \left( -(x - \Delta)/\sqrt{\gamma_n} \right)
\]
\[
\cdot \sum_{i=1}^{n} \left( \exp \left( \frac{X_i - \Delta}{\sqrt{\gamma_n}} \right) - \exp \left( -\frac{X_i - \Delta}{\sqrt{\gamma_n}} \right) \right),
\]

\[
\sin j = \frac{1}{n} \sum_{i=1}^{n} \sin \left( j(X_i - \Delta)\pi/d \right), \quad \cos j = \frac{1}{n} \sum_{i=1}^{n} \cos \left( j(X_i - \Delta)\pi/d \right).
\]

2.2 Methods of density estimation

2.2.1 Kernel estimators

The Parzen–Rosenblatt kernel estimator is defined by

\[
\hat{f}_h(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K \left( \frac{x - X_i}{h} \right),
\]

(2.18)

where \( f(x) \) is defined on \( \mathbb{R}^d \), \( h > 0 \) is a smoothing parameter (window width or bandwidth), and \( K(x) \) is a kernel function that usually satisfies the conditions

\[
K(x) \geq 0, \quad \int K(x) \, dx = 1.
\]

Definition 14 A kernel \( K(x) \) has an order \( r \) when a kernel function is chosen such that

\[
\int u^k K(u) \, du = \begin{cases} 
1, & k = 0 \\
0, & 1 \leq k \leq r - 1, \\
0, & k = r.
\end{cases}
\]

(2.20)

For \( r > 2 \) a kernel \( K \) and consequently the estimate \( \hat{f}_h(x) \) may have negative values. Then the estimate has to be normalized to the positive estimate:

\[
\left( \hat{f}_h(x) \right)^+ / \int_R \left( \hat{f}_h(x) \right)^+ \, dx.
\]
Symmetric kernels of odd orders are not considered, since their ‘moments’ \( \int u^k K(u)du \) are always equal to 0.

Examples of kernels are given by the normal PDF \( \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \), \( \exp(-|x|) \), \( \sin^2|x|/|x|^2 \), and Epanechnikov’s kernel

\[
K(x) = \frac{3}{4}(1-x^2)1\{|x| \leq 1\}.
\] (2.21)

Roughly speaking, the idea of kernel estimators is that all measurement points ‘are covered’ by the bell-shaped curves. The form of the ‘bell’ is determined by a kernel function. Figure 2.1 shows the bells constructed over several data points \( \{X_1, X_2, \ldots \} \) and the kernel estimate with bandwidth \( h = 0.7 \) for the Pareto PDF with the tail index \( \alpha = 0.3 \).

The accuracy of kernel estimates depends more on \( h \) than on \( K(x) \). The variances of all kernel estimates decay at rate \( O((nh)^{-1}) \) as \( nh \to \infty \), but the bias has order \( O(h^2) \) for second-order kernels \( (r = 2) \) and order \( O(h^4) \) for fourth order kernels \( (r = 4) \); see Silverman (1986). The variance of kernel estimates increases and the bias decreases as \( h \) decreases. Therefore, to select \( h \) one should find a trade-off between these effects.

Figure 2.1  Estimation by a standard kernel estimate with Epanechnikov’s kernel: Pareto PDF with \( \alpha = 0.3 \) (dot-dashed line), kernel estimate with \( h = 0.7 \) (solid line), and kernel constructed over sample points (dotted line).
It is well known that the MSE for a nonvariable bandwidth kernel estimate (2.18) \((d = 1)\) with second-order kernel\(^5\) obeys
\[
\text{MSE}(\hat{f}_h) = E(\hat{f}_h(x) - f(x))^2
= h^4(f''(x))^2K_1^2/4 + (nh)^{-1}f(x)R(K) + o((nh)^{-1} + h^4),
\]
as \(h \to 0\), where \(R(K) = \int K^2(t)dt\), when the second derivative of the PDF is continuous. The right-hand side is minimal for
\[
h_{\text{opt}} = \left( \frac{f(x)R(K)}{(f''(x))^2K_1^2} \right)^{1/5} n^{-1/5}.
\]
(2.22)

For such \(h_{\text{opt}}\), we obtain
\[
\text{MSE}(\hat{f}_h) = \frac{5}{4} \left( K_1 R(K)^2 \right)^{2/5} \left( f^2(x)f''(x) \right)^{2/5} n^{-4/5}.
\]

Epanechnikov’s kernel is the optimal second-order kernel for the estimate \(\hat{f}_h\) in the sense of the least \(K_1 R(K)^2\). But an ideal estimate cannot be attained since \(h_{\text{opt}}\) depends on the unknown derivative \(f''(y)\). In practice, \(f(y)\) can be replaced by the preliminary kernel estimate
\[
\hat{f}_{h_1}(x) = \frac{1}{nh_1} \sum_{i=1}^{n} K \left( \frac{x - X_i}{h_1} \right).
\]

Then the derivative \(f''(x)\) is evaluated by the second derivative of \(\hat{f}_{h_1}(x)\). Let us choose a kernel \(K(x)\) having two bounded derivatives and satisfying (2.19) (the nonnegativity of the kernel is not required), and (2.20), for example, the Epanechnikov kernel. Hence, since \(f''(x)\) is assumed to be continuous, \(\hat{f}_{h_1}(x)\) is the estimate of \(f''(x)\); see Prakasa Rao (1983). The smoothing parameter \(h_1\) of \(\hat{f}_{h_1}(x)\) can be taken to be \(\sigma_X \left( 84 \sqrt{\pi/(5n^2)} \right)^{1/13} \) (\(\sigma_X\) is the standard deviation of the sample \(X^n\)), which is consistent with the parameter of the kernel estimate with Gaussian kernel guaranteeing optimal estimation of \(\int (f''(x))^2 dx\) (Wand et al., 1991). It can be chosen by data-dependent methods such as cross-validation (Chow et al., 1983) or by the discrepancy method (Markovich, 1989).

A kernel estimator has the following advantages:

- It can be easily applied to multivariate densities.
- Recursive behavior: it is easy to use it for on-line estimation.

It has the following disadvantages:

- It may be negative for kernels of order \(r > 2\).

\(^5\) The simplest example of such kernels is provided by a symmetric kernel with compact support.
• It may exhibit boundary effects (distortions) on compactly supported PDFs due to kernel truncation near the boundaries.

• Kernel estimation is not good for nonsmooth PDFs, such as a uniform PDF.

Consistency conditions
The asymptotic conditions for the convergence of a kernel estimate to the true uniform continuous PDF were obtained in Parzen (1962) and Nadaraya (1965). According to Parzen (1962) for a marginal distribution and Murthy (1966) for a multivariate distribution,

\[ h \to 0 \quad \text{as} \quad n \to \infty \]  

provides an asymptotically unbiased estimate when \( K(x) \) obeys (2.19) and \( \sup_{-\infty < x < \infty} K(x) < \infty, \lim_{|x| \to \infty} xK(x) = 0 \) hold. The additional assumption

\[ nh^d \to \infty \quad \text{as} \quad n \to \infty, \]  

where \( d \) is the dimension, provides asymptotic convergence, and

\[ nh^{2d} \to \infty \quad \text{as} \quad n \to \infty \]

provides uniform convergence in probability (weak consistency). In Nadaraya (1965) the assumptions that the function \( K(x) \) has bounded variation and the expansion \( \sum_{n=1}^{\infty} \exp(-\gamma nh^2(n)) \) converges for any \( \gamma > 0 \) are shown to be necessary and sufficient conditions for convergence with probability 1 (strong consistency) to the true PDF as \( n \to \infty \).

On-line kernel estimators
An on-line PDF estimator may be defined as one where each update following the arrival of a new data value requires only \( O(1) \) calculations. Kernel estimators are usually used as on-line PDF estimators due to their recursive natural form. A closely related estimator,

\[ f_n(x) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h_j} K \left( \frac{x - X_j}{h_j} \right), \]

was introduced by Wolverton and Wagner (1969) and independently by Yamato (1971). This estimator can be calculated recursively, i.e.

\[ f_n(x) = \frac{n-1}{n} f_{n-1}(x) + \frac{1}{nh_n} K \left( \frac{x - X_n}{h_n} \right). \]

This property is particularly useful for large sample sizes since \( f_n(x) \) can be updated easily with each additional observation after only \( O(1) \) computations.
Deheuvels (1973) introduced the estimator

\[ f_n^*(x) = \left( \sum_{i=1}^{n} h_i \right)^{-1} \sum_{j=1}^{n} K \left( \frac{x - X_j}{h_j} \right). \]  

(2.26)

It was proved that \( f_n^*(x) \) has smaller asymptotic variance than (2.25). However, asymptotically (2.25) has smaller mean square error than (2.26); see Wertz (1985, p. 286).

In Hall and Patil (1994) more general on-line estimators are introduced, namely

\[ \tilde{f}_n(x) = \sum_{i=1}^{k} p_i \hat{f}_{ni}(x), \]  

(2.27)

where

\[ \hat{f}_{ni}(x) = b_i(n) \sum_{j=N_{i-1}+1}^{N_i} a_j h_j^{-1} K \left( \frac{x - X_j}{h_j} \right). \]  

(2.28)

Here, the \( a_j, j \geq 1 \), denote positive constants; \( k \geq 1 \) is a fixed integer; \( 1 = N_0 < N_1 < \ldots < N_k = n \) are integers which may depend on \( n \); \( p_1, \ldots, p_k > 0 \) satisfy \( \sum p_i = 1 \) and \( b_i(n) = (\sum_{N_{i-1} < j \leq N_i} a_j)^{-1} \). Roughly speaking, \( \tilde{f}_n(x) \) is a sum of weighted estimators \( \hat{f}_{ni}(x) \) constructed over the parts of the sample. The choice \( k = 1 \) and \( a_j = 1 \) leads to the estimator (2.25), while choosing \( k = 1 \) and \( a_j = h_j^{1/2} \) yields an estimator

\[ f_n^+(x) = n^{-1} h_n^{-1/2} \sum_{j=1}^{n} h_j^{-1/2} K \left( \frac{x - X_j}{h_j} \right), \]

considered by Yamato (1971) and Wegman and Davies (1979). For \( k \geq 2 \), the estimator defined by (2.27), (2.28) is not recursive, but is on-line in the sense defined at the beginning of this section. It is important that in Hall and Patil (1994), in contrast to other papers, the problem of the on-line calculation of the bandwidth (which may be updated in \( O(1) \) operations with the arrival of each new data value) is considered. At the same time, the problem is that the classical methods of the empirical choice of the bandwidth, such as cross-validation, typically demand \( O(n) \) calculations per data value.

### 2.2.2 Projection estimators

Projection estimators have the form

\[ \hat{f}^{pr}(x) = \sum_{i=1}^{N} \alpha_i \varphi_i(x), \]

where \( \varphi_i(x), i = 1, 2, \ldots, \) is some orthogonal basis in \( L_2 = L_2(\Omega) \), and \( f(x) \in L_2 \) is a true PDF. The coefficients \( \alpha_i = \frac{1}{n} \sum_{j=1}^{n} \varphi_i(X_j) \) are empirical estimates of the
true coefficients \( a_i = \int \varphi_i(x)f(x)dx \) of an expansion by the basis \( \varphi_i(x) \), \( \alpha_i \) are unbiased estimates of \( a_i \).

The number of terms \( N \) is a smoothing parameter. The bias of the estimate decreases, but the variance increases as \( N \) increases.

The Laguerre and trigonometric polynomials or a wavelet basis\(^6\) provide examples of the basis \( \{\varphi_i(x)\} \).

Projection estimators have the following advantages:

- They are faster than kernel estimates. It is necessary to keep only the values of the coefficients \( \alpha_i, i = 1, 2, \ldots, N \). One calculates only the values of the basis functions \( \varphi_i(x) \) at each point \( x \).
- They are convenient for on-line calculations.

Their disadvantages are as follows:

- They are defined only on compact sets (bounded intervals). The necessity of the integrability of PDF estimates leads to the integrability of all basic functions \( \varphi_i(x), i = 1, 2, \ldots, N \). This may lead to \( \int_{-\infty}^{\infty} \hat{f}^{\varphi}(x)dx \neq 1 \), when the projection estimate is considered on an infinite set (Devroye and Györfi, 1985). Application to long-tailed PDFs requires a preliminary transformation of the data to a bounded interval.
- They can be negative, i.e. the estimate may not be a PDF. The estimate can be replaced by zero in domains of negativity and may further be normalized in such a way that its integral is equal to 1 (the error in the space \( L_1 \) will just decay after the latter procedure (Devroye and Györfi, 1985)). Nevertheless, the simplicity of the construction which may be important for on-line estimates, may be lost.
- The projection estimate may not be consistent. Then it is better to consider the smoothed projection estimators

\[
\hat{f}_{\gamma}^{\varphi}(x) = \sum_{i=1}^{N} \lambda_i(n, \gamma) \alpha_i \varphi_i(x),
\]

where \( \lambda_i(n, \gamma) \) are smoothing tools that are stronger than \( N \). Markovich (1989) uses simulated data to show that the smoothed projection estimate (2.17) obtained by the regularization method and based on the cosine and sine functions is more accurate for finite nonsmooth PDFs than kernel estimates with Gaussian kernel. Vapnik et al. (1992) prove that the smoothed projection estimate (2.14) obtained by the expansion by the basic functions \( \varphi_i(x) = \cos(\pi i x), i = 0, 1, 2, \ldots, \) has rate of convergence \( n^{-(k+0.5)/(2k+3)} \) in

\(^6\) Haar’s basis is an example of the simplest wavelet basis.
metric space $L_2$. It is close to the best rate $n^{-(k+0.5)/(2k+2)}$ if the $k$th derivative of the PDF has bounded variation and the smoothing parameter $\gamma$ is selected by the $\omega^2$ discrepancy method (see Sections 2.2.4 and 4.8).

### 2.2.3 Spline estimators

For spline estimates, the empirical DF of the data is approximated by some piecewise smooth function using some quality criterion. The derivative of this smoothed empirical DF will be the spline estimate of the PDF. A smoothed histogram can be considered as a very rough spline estimate. Let us connect the centers of the tops of adjacent histogram bars by straight lines. The centers of the tops of the outside bars should be connected with the extreme points of the support of the DF. We call the curve obtained the smoothed histogram or frequency polygon.

However, histograms with equiprobable cells generally achieve better results than those with equally-sized cells (Devroye and Györfi, 1985; Tarasenko, 1968). Let $X_1 \leq X_2 \leq \ldots \leq X_n$ be the order statistics of the sample $X^n$. We set $\Delta_{1L} = [X_1, X_{(L)}]$, $\Delta_{2L} = (X_{(L)}, X_{(2L)})$, $\Delta_{3L} = (X_{(2L)}, X_{(3L)})$, $\ldots$ using the order statistics of the sample. A polygram, i.e., a histogram with variable bin width, is defined by

$$f_{L,n}(t) = \frac{L}{(n + 1)\lambda(\Delta_{rL})}, \quad t \in \Delta_{rL};$$

see Tarasenko (1968). Here $\lambda(\Delta)$ is the length of $\Delta$, and one assumes that $\lambda(\Delta_{rL}) \to 0$ and $L = o(n)$. The number of observations inside each interval $\Delta_{rL}$ is less than or equal to $L$.

The advantages of the polygram are twofold: First, the asymptotic convergence rate of a polygram in the $L_1$ metric reaches $n^{-2/5}$ for some PDFs, the same rate as achieved for a kernel estimate. In contrast, a histogram with equally-sized cells achieves a limit rate of $n^{-1/3}$ in $L_1$. Second, a polygram dynamically adapts the bin width to the data and works better than a histogram. It has the disadvantage that it cannot be directly applied to heavy-tailed PDFs since it is defined on bounded intervals, and so requires a preliminary transformation of the data.

### 2.2.4 Smoothing methods

The application of nonparametric estimates requires the definition of a so-called smoothing parameter. This is the bandwidth $h$ in the kernel estimator (2.18), the number of observations $L$ inside each sub-interval $\Delta_{rL}$ or, generally, the regularization parameter $\gamma$, since nonparametric estimates can be obtained by the regularization method (see Section 2.1). The choice of the smoothing parameter is the most important part of the estimation.

Conditions on a smoothing parameter such as (2.23) and (2.24) are asymptotic. In practice, one needs to select a smoothing parameter using a sample of moderate size.

Generally, methods to select a smoothing parameter fall into two classes.
Minimization of quality criteria

The first class contains methods connecting the unknown parameters to criteria of quality. For example, for the estimate $\hat{f}_h(x)$ in (2.18), minimizing $E \left( \left( \hat{f}_h(x) - f(x) \right)^2 \right)$ with respect to $h$ gives $h = h_n = c(x)n^{-1/5}$, where $c(x)$ depends on the second derivative of an unknown PDF (Silverman, 1986). Such methods are not suitable in practice, since

- the unknown parameters depend on the unknown PDF and its derivatives;
- they assume the existence of derivatives of the PDF, which may not exist.

The over-smoothing bandwidth selection

This is used for the kernel estimator (2.18). It relies on the fact that there is an upper bound for the bandwidth that minimizes the AMISE for estimation of PDFs having standard deviation $\sigma$ (Wand and Jones, 1995). The over-smoothing bandwidth selector

$$\hat{h}_{OS} = \left( \frac{243R(K)}{35\mu_2(K)^2n} \right)^{1/5} \cdot \sigma$$

(2.30)

provides the value of $h$ at the minimum of this upper bound. Here,

$$\mu_2(K) = \int z^2 K(z) dz, \quad R(K) = \int K^2(x) dx.$$

While $\hat{h}_{OS}$ gives a too large bandwidth for an arbitrary PDF, it provides an excellent starting point for the choice of $h$. It is also reasonable to consider fractions of $\hat{h}_{OS}$ such as $\hat{h}_{OS}/2$ and $\hat{h}_{OS}/4$. This method does not require a preliminary estimation of the derivatives of the PDF.

Data-dependent methods

The second class contains methods which provide a data-dependent choice of the smoothing parameter by the minimization of some functional. Such methods are a more practical tool for PDF estimation than estimators derived from theory such as $h(n) \sim n^{-1/5}$.

Cross-validation

Cross-validation is a well-known data-dependent method that is very close to the ML method (see Stone, 1974; Wahba, 1981). The bandwidth, $h$ is selected by finding the maximum of the functional

$$\prod_{i=1}^{n} \hat{f}_{-i}(X_i; h) \rightarrow \max_h$$

(2.31)
where
\[ \hat{f}_{-i}(x; h) = \frac{1}{(n-1)h} \sum_{j=1, j\neq i}^{n} K\left(\frac{x-X_j}{h}\right) \]
is the kernel estimate constructed by the sample with excluded observation \(X_i\). The exclusion of one observation or cross-validation is needed because the likelihood function \(L(h) = \prod_{i=1}^{n} \hat{f}_h(X_i)\), where \(\hat{f}_h(x)\) is the estimate (2.18), tends to infinity as \(h \to 0\), since \(X_j - X_i = 0\) as \(i = j\).

The convergence of kernel estimates with the choice of \(h\) by cross-validation is proved for PDFs which have a compact support, for example, uniform or triangular PDFs (Chow et al., 1983). For heavy-tailed PDFs the estimates with \(h\) selected by cross-validation do not converge in the space \(L_1\), since \(h \to \infty\) as \(n \to \infty\) (Devroye and Györfi, 1985).

**Least-squares cross-validation**

The maximization of (2.31) is equivalent to the minimization of the functional
\[ \frac{1}{n} \sum_{i=1}^{n} H\left(\delta_{X_j}, \hat{f}_{-j}(X_j; h), f(x)\right), \tag{2.32} \]
where \(\delta_{X_j}\) is the delta-functional at the point \(X_j\), \(H(p, q, r) = \mathcal{I}(p, q) - \mathcal{I}(p, r)\) is a relative loss function, and \(\mathcal{I}(p, q) = \int p(x) \ln(p(x)/q(x)) dx\) is Kullback’s metric. Since
\[ H(p, q, r) = \int p(x) \ln(r(x)/q(x)) dx \]
in Kullback’s metric, (2.32) can be rewritten as
\[ \frac{1}{n} \sum_{i=1}^{n} \ln\left(\frac{f(X_j)}{\hat{f}_{-j}(X_j; h)}\right). \tag{2.33} \]
Expressions (2.31) and (2.33) are equivalent criteria for the selection of \(h\). To see this, it is enough to take the logarithm of (2.31).

One can use the different metric
\[ \mathcal{I}(p, q) = \int (p(x) - q(x))^2 dx \]
in the expression for \(H(p, q, r)\). Minimization of (2.32) then leads to selection of \(h\) by the minimization of the sum
\[ \text{LSCV}(h) = n^{-1} \sum_{i=1}^{n} \int \hat{f}_{-i}(x; h)^2 dx - 2n^{-1} \sum_{i=1}^{n} \hat{f}_{-i}(X_i; h). \tag{2.34} \]
This method is called least-squares cross-validation or integrated squared error cross-validation (Rudemo, 1982; Bowman, 1984). The integral in \(\text{LSCV}(h)\) can
be calculated analytically, when the kernel estimate $\hat{f}_{-i}(x; h)$ has a normal kernel function $N(x, h^2) = (1/(h\sqrt{2\pi})) \exp(-x^2/(2h^2))$, that is,

$$LSCV(h) = \frac{1}{n-1} N(0, 2h^2) + \frac{n-2}{n(n-1)^2} \sum_{i \neq j} N(X_i - X_j, 2h^2)$$

$$- \frac{2}{n(n-1)} \sum_{i \neq j} N(X_i - X_j, h^2).$$

(2.35)

Another version of cross-validation

Hall (1983, 1985), Rudemo (1982), and Bowman (1982) have proposed finding the value of $h$ that minimizes the criterion

$$CV(h) = \int \hat{f}^2_h(x) dx - 2n^{-1} \sum_{i=1}^n \hat{f}_{-i}(X_i; h),$$

(2.36)

where $\hat{f}_h$ is kernel estimator (2.18). In Stone (1984) it is proved that an $h^*$ obtained by this method is the best in the sense that

$$\int (\hat{f}_h(x) - f(x))^2 dx \min_h \int (\hat{f}_h(x) - f(x))^2 dx \to 1$$

almost surely for any bounded PDFs $f(x)$ at $R$, if the kernel $K(x)$ is symmetric, has compact support, and satisfies the Lipschitz condition. It is assumed that $\int K(x) dx = 1$ and $\int K(x)^2 dx < 2K(0)$.

Advantages of cross-validation

- It allows $h$ to be adapted to a concrete sample.
- In contrast to (2.22), it avoids estimation of the first two derivatives of the PDF, which is an awkward problem in itself.

Disadvantages of cross-validation

- Slow convergence rate and high sampling variability (see Park and Marron, 1990).
- The multiple minima of $LSCV(h)$ or generally $\prod_{i=1}^n \hat{f}_{-i}(X_i; h)$ create problems when finding $h$ in practice. This is inherited from the ML method.
The discrepancy method

One of the data-dependent smoothing tools is given by the so-called discrepancy method. It is an alternative to cross-validation. The idea is to select \( h \) as a solution of the discrepancy equation

\[
\rho(\hat{F}_h, F_n) = \delta. \tag{2.37}
\]

Here, \( \hat{F}_h(x) = \int_{-\infty}^{x} \hat{f}_h(t) \, dt \), where \( \hat{f}_h(t) \) is some estimate of the PDF, \( \delta \) is a known uncertainty of the estimation of the DF \( F(x) \) by the empirical DF \( F_n(t) \), i.e. \( \delta = \rho(F, F_n) \), and \( \rho(\cdot, \cdot) \) is a metric in the space of distribution functions. The discrepancy method was proposed and investigated in Markovich (1989) and Vapnik et al. (1992) for the smoothing of nonparametric PDF estimates. Since \( \delta \) is usually unknown, in these papers certain quantiles of the limit distribution of the von Mises–Smirnov statistic

\[
\omega_n^2 = n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \, dF(x)
\]

or, equivalently,

\[
\omega_n^2 = n \int_{0}^{1} (F_n(t) - t)^2 \, dt
\]

for the transformed sample \( t_i = F(X_i), i = 1, \ldots, n \), and the Kolmogorov–Smirnov statistic\(^8\)

\[
\sqrt{n}D_n = \sqrt{n} \sup_{-\infty < x < \infty} |F(x) - F_n(x)|
\]

were used as \( \delta \).

Let \( X_{(1)} < X_{(2)} < \ldots < X_{(n)} \) be order statistics of the sample \( X'' \). The probability of any two order statistics being equal is zero since \( F(x) \) is continuous (\( f(x) = F'(x) \) is assumed to exist). For calculations one can use the following simple

---

\(^7\) Here, \( F_n(t) \) is the empirical DF calculated by the sample \( t_1, t_2, \ldots, t_n \). \( t_i \) is uniformly distributed if \( F(x) \) is the DF of the r.v. \( X \).

\(^8\) The distributions of the two latter statistics do not depend on \( F(x) \).
expressions (Smirnov and Dunin-Barkovsky, 1965) for the statistics $\omega_n^2$ and $\sqrt{n}D_n$, respectively:

$$\hat{\omega}_n^2(h) = \sum_{i=1}^{n} \left( \hat{F}_n(X_{(i)}) - \frac{i - 0.5}{n} \right)^2 + \frac{1}{12n}, \quad (2.38)$$

$$\sqrt{n}D_n(h) = \sqrt{n} \max(\hat{D}_n^+, \hat{D}_n^-), \quad (2.39)$$

where

$$\hat{D}_n^+ = \max_{1 \leq i \leq n} \left( \frac{i}{n} - \hat{F}_n(X_{(i)}) \right), \quad \hat{D}_n^- = \max_{1 \leq i \leq n} \left( \hat{F}_n(X_{(i)}) - \frac{i - 1}{n} \right). \quad (2.40)$$

Kolmogorov proved that the DF of statistic $\sqrt{n}D_n$ has limit

$$K_f(x) = \sum_{i=-\infty}^{\infty} (-1)^i \exp(-2i^2x^2),$$

as $n \to \infty$, which does not depend on $F(x)$. For sufficiently large $n \geq 20$ and any $x > 0$ we have

$$P\{\sqrt{n}D_n < x\} \approx K_f(x),$$

(Bolshev and Smirnov, 1965). The limit distribution of $\omega_n^2$ is rather complicated, namely,

$$\lim_{n \to \infty} P\{\omega_n^2 < x\} = \frac{1}{\sqrt{2\pi}} \sum_{j=0}^{\infty} \frac{\Gamma(j + 1/2)}{\Gamma(1/2) \Gamma(j + 1)} \sqrt{4j + 1} \exp \left( -\frac{(4j + 1)^2}{16x} \right) \left\{ I_{-1/4} \left[ \frac{(4j + 1)^2}{16x} \right] - I_{1/4} \left[ \frac{(4j + 1)^2}{16x} \right] \right\},$$

where $I_i(z)$ is a modified Bessel function (Bolshev and Smirnov, 1965; Martynov, 1978). This limit distribution is safe to use for the distribution of $\omega_n^2$ for $n > 40$.

---

9 Since the empirical DF

$$F_n(x) = \begin{cases} 0, & x < X_{(1)}, \\ k/n, & X_{(k)} \leq x < X_{(k+1)}, k = 1, 2, \ldots, n-1, \\ 1, & x \geq X_{(n)}, \end{cases}$$

is used, we get

$$\frac{\omega_n^2}{n} = \int_{-\infty}^{X_{(1)}} (0 - F(x))^2 dF(x) + \sum_{k=1}^{n-1} \int_{X_{(k)}}^{X_{(k+1)}} \left( \frac{k}{n} - F(x) \right)^2 dF(x) + \int_{X_{(n)}}^{\infty} (1 - F(x))^2 dF(x)$$

$$= \frac{F^2(X_{(1)})}{3} + \sum_{k=1}^{n-1} \frac{(F(X_{(k+1)}) - k/n)^3}{3} - \sum_{k=1}^{n-1} \frac{(F(X_{(k)}) - k/n)^3}{3} + \frac{(1 - F(X_{(n)}))^3}{3}.$$
Figure 2.2  Approximate derivative $\sum_{i=-10000}^{10000} (-1)^i \exp(-2i^2x^2)(-4i^2x)$ of $K_f(x)$ against $x$. Its maximum occurs at $x = 0.7$.

According to the tables of the distributions of statistics $\omega^2_n$ and $D_n$ (Bolshev and Smirnov, 1965), 0.05 and 0.7 are the maximum likelihood values (i.e., the quantiles corresponding to a maximum of the PDF) of the $\omega^2_n$ and $D_n$ statistics, respectively (Figure 2.2).

The corrected value of $\delta$ for moderate samples is equal to 0.5 for $D_n$ (Markovich, 1989). Hence, the practical versions of the two discrepancy methods imply the choice of $h$ from the equations

$$\hat{\omega}^2_n(h) = 0.05 \quad (2.41)$$

for the $\omega^2$ method and

$$\sqrt{n}\hat{D}_n(h) = 0.5 \quad (2.42)$$

for the $D$ method.

Number of operations for the discrepancy method

The number $N^*$ of operations required to find the solution of (2.41) or (2.42) depends on the accuracy. For example, the accuracy of the method of dividing the interval on two halves is determined by $\varepsilon = 2^{-1-N^*/2}$ (Knuth, 1973). The number of operations for a standard kernel estimator is $O(n^2)$ for a fixed value of $h$. Project estimators require fewer operations, that is, $O(Nn)$, $N < n$.

Example 7  For $n = 10000$, $\varepsilon = 0.01$ and $N = 12$ we need $\approx NnN^* \approx 1.44 \times 10^6$ operations for a projection estimator and $\approx n^2N^* \approx 1.2 \times 10^9$ operations for a kernel estimator.

Advantages of the discrepancy methods

- The discrepancy methods are based on the observed (ungrouped) sample points.
• In contrast to the cross-validation method, calculating the maximum of some criterion is not required. Hence, one can avoid the problem of cross-validation falling into local extremes.

• According to a simulation study (Markovich, 1989), the discrepancy methods (2.41) and (2.42) provide better results than cross-validation for nonsmooth (e.g., triangular and uniform) distributions.

• The rate of convergence in $L_2$ for the $\omega^2$ method applied to a projection estimator is not worse than $n^{-(k+1/2)/(2k+3)}$, which is close to the best $n^{-(k+1/2)/(2k+2)}$ for PDFs with a bounded variation of the $k$th derivative (see Section 4.9).

Disadvantages of the discrepancy methods

• The number of operations may be high for large samples.

Open problems of the discrepancy methods

• For some heavy-tailed distributions the discrepancy equations (2.41) and (2.42) may have no solution. This implies that higher quantiles of the distributions of the statistics $\omega_n^2$ and $\sqrt{n}D_n$ may be required.

• The value of $h$ that corresponds to the largest local minimum of the statistics $\omega_n^2$ and $\sqrt{n}D_n$ may provide an accurate PDF estimate.

These problems have as yet not been investigated. Theoretical properties of the $\omega^2$ and $D$ discrepancy methods are considered in Chapter 4.

2.2.5 Illustrative examples

Figures 2.3–2.5 show the results of the application of kernel estimate (2.18) with $d = 1$ to a finite PDF (i.e., uniform), and heavy-tailed PDFs (i.e., Pareto $(1 + \gamma x)^{-1/(\gamma + 1)}$ with $\gamma = 0.5$ and Cauchy $(\pi(1 + x^2))^{-1}$). In order to calculate the bandwidth $h$ the over-smoothing method (2.30), the LSCV (2.35) and the $D$ discrepancy method (2.42) were applied. The sample size was $n = 100$ (see Table 2.1).

With regard to the Cauchy distribution (see Figure 2.5) one can observe the following phenomenon. The actual solution of the discrepancy equation (2.42) does not exist, since the statistic $\sqrt{n}D_n(h)$ is larger than the level 0.5 for any $h$. This implies that the quantile 0.5 of the distribution of the Kolmogorov–Smirnov statistic cannot be used as a value of $\delta$ in (2.37) for such a heavy-tailed distribution. The higher quantiles seems to be more appropriate values of $\delta$. Hence, the $D$ method is modified. The largest local minimizer of $\sqrt{n}D_n(h)$ is selected. Obviously, it gives a better estimate than the bandwidth that is provided by the over-smoothing and the LSCV methods.
Figure 2.3  Kernel estimate with different smoothing methods – discrepancy method (solid line), over-smoothing (dotted line), LSCV (line with circles) – for the uniform PDF (dotted-dashed line) (left) and the statistic $\sqrt{n \hat{D}_n(h)}$ against $h$ (solid line) (right). A normal kernel is used in the case of the LSCV, otherwise Epanechnikov’s kernel.

Figure 2.4  Kernel estimate with different smoothing methods – discrepancy method (solid line), over-smoothing (dotted line), LSCV (line with circles) – for the Pareto PDF (dotted-dashed line) (left) and the statistic $\sqrt{n \hat{D}_n(h)}$ against $h$ (solid line) (right). A normal kernel is used in the case of the LSCV, otherwise Epanechnikov’s kernel.

A similar effect is observed for the cross-validation method. The LSCV($h$) statistic may have several minima. The largest local minimizer provides smaller MISE than the actual minima of LSCV($h$); see Wand and Jones (1995, p. 64).
Figure 2.5  Kernel estimate with different smoothing methods – discrepancy method (solid line), over-smoothing (dotted line), LSCV (line with circles) – for the Cauchy PDF (dotted-dashed line) (left), and the statistic $\sqrt{n}\hat{D}_n(h)$ against $h$ (solid line) (right). The $D$ method is modified: $h$ corresponds to a minimum of $\sqrt{n}\hat{D}_n(h)$. A normal kernel is used in the case of the LSCV, otherwise Epanechnikov’s kernel.

Table 2.1  Bandwidth values for different smoothing methods.

<table>
<thead>
<tr>
<th>PDF</th>
<th>Over-smoothing</th>
<th>LSCV</th>
<th>$D$ method</th>
</tr>
</thead>
<tbody>
<tr>
<td>uniform</td>
<td>0.281</td>
<td>0.102</td>
<td>0.14</td>
</tr>
<tr>
<td>Pareto</td>
<td>1.915</td>
<td>1.123</td>
<td>0.23</td>
</tr>
<tr>
<td>Cauchy</td>
<td>17.878</td>
<td>1.38</td>
<td>0.19</td>
</tr>
</tbody>
</table>

The examples demonstrate that the over-smoothing method may be poor for heavy-tailed distributions like the Cauchy distribution having no finite second moment.

2.3  Kernel estimation from dependent data

This section is devoted to the nonparametric kernel estimation of a univariate PDF with dependent time series data.

It is known that the bias of a kernel estimate is the same for independent and dependent data. However, the variance is larger for the dependent case and depends on the correlation structure of the data.

The bandwidth $h$ is a smoothing parameter of the kernel estimator that drives its accuracy. The idea is to select such a bandwidth of the kernel estimate to reduce
the mean squared error in the case of dependent data. Then the MSE can converge for short- or long-range dependent data at the same rate as the data would under the assumption of independence.

2.3.1 Statement of the problem

Let \( \{X_j, j = 1, 2, \ldots \} \) be a stationary sequence with marginal PDF \( f(x) \). Suppose that the process is observed up to ‘time’ \( n + 1 \). It is assumed that the bivariate distributions of the sequence are absolutely continuous, writing \( f_j(x, y) \) for the joint PDF of \( X_1 \) and \( X_{1+j}, j = 1, 2, \ldots \). It is assumed that \( \text{cov}(X_j, X_{j+k}) \) depends only on \( k \). Consider the kernel estimator

\[
\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h (x - X_i)
\]  

(2.43)

as an estimate of \( f(x) \). Here, \( K_h (x) = (1/h)K(x/h) \) is a kernel function, \( h \) is a bandwidth.

The bias of \( \hat{f}_h(x) \) is unaffected by dependence since

\[
E\hat{f}_h(x) = EK_h (x - X_1) = (K_h * f) (x),
\]

similar to the case of independent data. Because of stationarity the variance of this kernel estimate

\[
\text{var}(\hat{f}_h(x)) = \frac{1}{n} \text{var}K_h (x - X_1) + \frac{2}{n} \sum_{j=1}^{n-1} \left( 1 - \frac{j}{n} \right) \text{cov}\{K_h (x - X_1), K_h (x - X_{j+1})\}
\]

(2.44)

consists of the variance of the kernel estimator based on independent observations \( \{X_j\} \) (the first term) and a term reflecting the dependence structure of the data (the second term); see Wand and Jones (1995). The variance of all kernel estimates constructed from independent data has order \( \sim 1/(nh) \). This well-known result cannot be improved. By Lemma 3.2 of Castellana and Leadbetter (1986), the second term has order \( O(\beta_n/n) \), where

\[
\beta_n = \sup_{x,y} \sum_{j=1}^{n} |f_j(x, y) - f(x)f(y)|
\]

is the dependence index sequence. Clearly, for i.i.d. sequences \( \beta_n = 0 \) for all \( n \), for sequences with strong long-range dependence \( \beta_n \) may tend to infinity, and in between \( \beta_n \) may converge to a finite limit at various rates.

Hence, one can see that for long-range dependent data that we may observe for Internet traffic the kernel estimator may provide an estimate with a large variance. Our aim is to find the value of \( h \) that minimizes the second term on the right-hand side of (2.44). The correlation structure of the data cannot be changed. However, the parameter \( h \) can make the second term less. At the same time, it would not be correct to reduce only this part of the variance. We shall select such an \( h \) that
reduces the mean squared error of the kernel estimate (Markovich, 2006b). Let us rephrase the results of Castellana and Leadbetter (1986) for estimator (2.43).

We suppose that the kernel function $K(x)$ of estimator (2.43) satisfies the following properties:

(i) $\int K(x)dx = 1$;
(ii) $\int xK(x)dx = 0$;
(iii) $K(x)$ is compactly supported on the interval $|x| \leq \lambda$, that is, $K(x) = 0$ for $|x| > \lambda > 0$.

Let the PDF $f(x)$ have a continuous, bounded second derivative $f''(x)$. Then the bias of $\hat{f}_h(x)$ satisfies

$$\text{bias}(\hat{f}_h(x)) = \frac{1}{2} h^2 K_1 f''(x) + o(h^2), \quad \text{as } n \to \infty,$$

where $K_1 = \int u^2 K(u)du$. By Theorem 3.5 of Castellana and Leadbetter (1986) it follows for the kernel estimator (2.43) that

$$nh\text{var}(\hat{f}_h(x)) = f(x)R(K) + \frac{1}{2} h^2 f''(x) K_1 (1 + o(1)) - hf^2(x)(1 + o(1)) + O(h\beta_n),$$

where $R(K) = \int K^2(u)du$ and $\beta_n$ is the dependence index sequence of the process $\{X_j; j \geq 1\}$, if the PDF $f(x)$ has a bounded second derivative $f''(x)$. In the i.i.d. case (when $\beta_n = 0$) the final term vanishes. However, for dependent cases $\beta_n$ may even go to infinity. The latter implies that the kernel estimate may have an infinite variance.

In Hall et al. (1995, Theorem 2.1), the covariance $\text{cov}\{K_h(x - X_1), K_h(x - X_{j+1})\}$ is represented by $g_j(x, x) + h^2 r(x)$, denoting $g_j(x, y) = f_j(x, y) - f(x)f(y)$. It is assumed that each $g_j$ has two derivatives of all types and the conditions that provide short-range dependence are satisfied. Here, $r(x)$ is a value which depends on the derivatives of $g_j(x, y)$ and the kernel $K$. Thus, it follows that

$$nh\text{var}(\hat{f}_h(x)) = f(x)R(K) + \frac{1}{2} h^2 f''(x) K_1 (1 + o(1)) - hf^2(x)(1 + o(1))$$

$$+ O(h(\beta_n + h^2)).$$

We assume additionally that the kernel function satisfies the condition

$$\int_{-h}^{h} |K(u)|du \sim h.$$
The Epanechnikov’s kernel \( K(x) = \frac{3}{4}(1 - x^2)1(|x| \leq 1) \) is an example of such a kernel. Then Lemma 3.2 in Castellana and Leadbetter (1986) can be rewritten in the following way.

**Lemma 2** If the stationary sequence \( \{X_i; i = 1, 2, \ldots \} \) has a dependence index sequence \( \{\beta_n; n \geq 1\} \) and the kernel function \( K(x) \) satisfies (2.48) then, for any fixed real \( x, y \),

\[
\sum_{i=1}^{n} \left| \text{cov} (K_h (x - X_i), K_h (x - X_{1+i})) \right| = O \left( (h + \eta)^2 \beta_n \right),
\]

where \( \eta > 0 \) is a constant.

Therefore, one can rewrite (2.46) using Lemma 2:

\[
nh \text{Var} \left( \hat{f}_h(x) \right) = f(x)R(K) + \frac{1}{2} h^2 f''(x) K_1 (1 + o(1)) - h f^2(x) (1 + o(1))
\]

\[
+ O \left( h (h + \eta)^2 \beta_n \right).
\]

(2.50)

The mean squared error of the estimator \( \hat{f}_h(x) \) can be represented in a standard way by

\[
\text{MSE} \left( \hat{f}_h(x) \right) = \text{bias} \left( \hat{f}_h(x) \right)^2 + \text{var} \left( \hat{f}_h(x) \right).
\]

(2.51)

Hence, from (2.45)–(2.47) and (2.50) it follows that

\[
\text{MSE} \left( \hat{f}_h(x) \right) = \frac{h^4(K_1 f''(x))^2}{4} + \frac{f(x)}{n h} R(K) + \frac{1}{2} \frac{h f''(x)}{n} \frac{K_1}{(1 + o(1))}
\]

\[
- \frac{f^2(x)}{n} (1 + o(1)) + T_n,
\]

(2.52)

where \( T_n = O \left( \frac{\beta_n}{n} \right) \), \( T_n = O \left( \frac{\beta_n + h^2}{n} \right) \), and \( T_n = O \left( \frac{(h + \eta)^2 \beta_n}{n} \right) \) correspond to (2.45), (2.46), and (2.47), respectively.

We assume that the bandwidth \( h \) satisfies the standard conditions of the reliability of kernel estimates, namely, \( h > 0, h \to 0, n \to \infty \). Then the term \( h f''(x) K_1/(2n) (1 + o(1)) \) should be omitted since it is no larger than the fourth term. We shall find the value of \( h \) that minimizes the remainder.

The derivative of MSE with respect to \( h \) gives \( h \sim n^{-1/5} \) and MSE \( \sim n^{-4/5} \) (the best possible rate for the class of PDFs considered) if \( \beta_n = 0 \).

Let us consider the dependent case where \( \beta_n \neq 0 \). Note that the term \( T_n = O \left( \frac{\beta_n}{n} \right) \) does not depend on \( h \) (Castellana and Leadbetter, 1986). Hence, the derivative of the corresponding MSE gives \( h \sim n^{-1/5} \) similar to the case of independent data. The value \( h \) is not affected by the dependence \( \beta_n \) and the rate of the MSE is determined by \( \max \{ n^{-4/5}, \beta_n/n \} \). This implies that one cannot reduce the critical part of the variance dependent on \( \beta_n \) by means of \( h \).
The same is true for the term $T_n = O\left(\frac{\beta_n + h^2}{n}\right)$ that is valid under the short-range dependent conditions ($\beta_n < \infty$), since $O\left(\frac{h^2}{n}\right)$ is less than the third term on the right-hand side of (2.52). In Hall et al. (1995) a stationary Gaussian sequence $\{X_i\}$ with zero mean, unit variance and covariance $\gamma(i) = E(X_{i+j}X_j) \sim ci^{-\alpha}$ (as $i \to \infty$) (the long-range dependence corresponds to $\alpha \leq 1$) for all integers $j \geq 0$ is considered.\footnote{$\sum_{i=1}^{\infty} |\gamma(i)| < \infty$ is the usual definition of ‘short-range’ dependence.} It is shown that the optimal convergence rate $\text{MISE} \sim \text{const.} n^{-1/5}$ for independent data is maintained for $\alpha > 4/5$, which includes many cases of long-range dependence.

In our approach, the value $h$ is linked with $\beta_n$ in order to influence on $\text{var} \left(\hat{f}_h(x)\right)$. The derivative of the upper bound of $\text{MSE} \left(\hat{f}_h(x)\right)$ with $T_n = \left(\frac{\beta_n + h^2}{n}\right)$ with respect to $h$ gives the following equation in $h$:

$$h^5 + 2\beta_n h^2 (h + \eta)/(n(K_1f''(x))^2) - f(x)R(K)/(n(K_1f''(x))^2) = 0. \quad (2.53)$$

However, the direct solution of (2.53) is complicated. To avoid this problem, one can select

$$h = \begin{cases} \min(\sqrt{2\beta_n/(nK_1^2c_2)}, (c_1R(K)/(4\beta_n))^{1/3}), & \beta_n \neq 0, \\ (c_1R(K)/(nK_1^2c_2))^{1/5}, & \beta_n = 0. \end{cases} \quad (2.54)$$

This follows from the equivalent form of (2.53), that is, $h^3 (h^2 + 2\beta_n/(n(K_1f''(x))^2)) = (\hat{f}(x) - 2\beta_n h^2 \eta)/(n(K_1f''(x))^2)$, and the assumptions that $h^2$ has the same rate of convergence as $2\beta_n/(n(K_1f''(x))^2)$ if $n \to \infty$ and $f(x) \leq c_1, f''(x)^2 > c_2, c_1, c_2 > 0$.

From (2.52) and (2.54) we have $h \sim n^{-1/5}$ and the optimal rate $\text{MSE} \sim n^{-4/5}$ if $\beta_n = 0$. Assume now that $\beta_n \sim n^\alpha, \alpha \in R$. Then we get

$$\text{MSE} \sim \begin{cases} n^{\alpha - 1}, & \alpha \geq 1/3, \\ n^{-(1+\alpha)/2}, & \alpha < 1/3, \end{cases} \quad (2.55)$$

as $n \to \infty$. The minimal value of $\text{MSE} \sim n^{-2/3}$ corresponds to $\alpha = 1/3$ (the latter result is worse than the optimal rate $\text{MSE} \sim n^{-4/5}$ for the independent case); see Figure 2.6. We remind the reader that $\alpha > 0$ and $\alpha \leq 0$ correspond to long- and short-range dependence, respectively. For $\alpha > 1$ and $\alpha < -1$ we have $\text{MSE} \to \infty$.

### 2.3.2 Numerical calculation of the bandwidth

Here, we consider the approach with $T_n = O\left(\frac{(h+\eta)^2\beta_n}{n}\right)$. The calculation of $h$ by formula (2.54) requires a preliminary analysis of the data dependence and a pilot
estimation of the unknown functions \( f(x) \), \( f''(x) \), and \( \beta_n \). Dependence may be detected via the estimation of Pickands dependence function (Beirlant et al., 2004; see also Section 1.3.4 above).

To estimate \( f(x) \), \( f''(x) \) and \( \beta_n \), a kernel estimator can be used. Suppose the \( \{X_j, j \geq 1\} \) are observed up to ‘time’ \( n + 1 \) within \( m \) ‘days’. We denote by \( X_{i,j} \) the \( i \)th observation of the process measured on ‘day’ \( j \). In other words, we need \( m \) realizations of the process, assuming that all observed r.v.s are identically distributed. Indeed, for better estimation \( m \) should be large enough, at least 50–100.

Then the product kernel estimator (Scott, 1992)

\[
\hat{f}_j^{h_1,h_2}(x, y) = \frac{1}{m h_1 h_2} \sum_{i=1}^{m} K\left(\frac{x - X_{1,i}}{h_1}\right) K\left(\frac{y - X_{1+j,i}}{h_2}\right)
\]  

(2.56)

may be used to estimate \( \beta_n \) by \( \hat{\beta}_n = \sup_{x,y} \sum_{j=1}^{n} |\hat{f}_j^{h_1,h_2}(x, y) - \hat{f}_{h_1}(x) \hat{f}_{h_2}(y)| \). For \( K(x) \) one can use the same kernels as for the univariate case, e.g., a normal kernel. One can take \( \hat{h}_i = \sigma_i n^{-1/6} \) as the estimate of \( h_i, i = \{1, 2\} \).\(^{13}\) Here, \( \sigma_i \) is the sample standard variation constructed by the observations \( \{X_{i,j}, j = 1, \ldots, m\} \).

\(^{12}\) The alternative is to separate the sequence into \( m \) blocks. Then \( X_{i,j} \) is the \( i \)th observation within the \( j \)th block.

\(^{13}\) For other kernels, the equivalent \( \hat{h} \) may be obtained as \( A(K) \sigma_i n^{-1/6} \), for example, \( A(K) = 1.77 \) for Epanechnikov’s kernel (Silverman, 1986).
2.3.3 Data-driven selection of the bandwidth

Hart and Vieu (1990) proposed a variant of the cross-validation method (2.36) that seems to be more appropriate for dependent data. They defined the so-called leave-out $l$ cross-validation function by

$$CV_l(h) = \int \hat{f}_h^2(x)dx - 2n^{-1} \sum_{i=1}^{n} \hat{f}_{-i}^l(X_i; h).$$

(2.57)

This is based on a different representation of $\hat{f}_{-i}(x; h)$ than in (2.31), namely

$$\hat{f}_{-i}^l(x; h) = \frac{1}{n_i h} \sum_{j=1, |i-j|>l} K\left(\frac{x-X_j}{h}\right),$$

(2.58)

for $1 \leq i \leq n$, where $l = l_n$ is a sequence of positive integers, called the leave-out sequence, and $n_i$ is such that

$$nn_l = \#\{(i, j) : |i-j| > l \text{ and } 1 \leq i, j \leq n\}.$$ 

The motivation of such an approach is rather natural. Deleting $l$ neighboring data points reduces the dependence between the r.v.s $X_i$ and $\{X_j : |i-j| > l\}$. The value $l = 0$ corresponds to the usual cross-validation method for independent data.

However, a simulation study provided in Hall et al. (1995) allows us to conclude that cross-validation can be difficult to implement with heavily dependent data and can produce bandwidths of very high variability, owing to the relative flatness of the function $CV_l(h)$. Varying the value of $l$ in the cross-validation algorithm of Hart and Vieu (1990) had relatively little effect either on the mean value of $\hat{h}$ or on MISE, although it did influence the variability of the algorithm. The discrepancy methods (2.41) and (2.42) require independent data like for cross-validation. They can be revised from the same perspective. In other words, $\hat{F}_h(X_j)$ may be calculated by $\int_{-\infty}^{X_j} \hat{f}_{-j}^l(t; h)dt$. Thus, $\hat{F}_h(X_{(i)})$ could be replaced in the latter formulas, where $X_j$ corresponds to $X_{(i)}$.

2.4 Applications

2.4.1 Finance: evaluation of market risk

In the analysis of financial data we are mainly interested in the probability of large losses and, hence, the upper tail of the loss distribution. These large losses are caused by the sudden volatility that may be observed in financial markets (Chen et al., 2005). Risk factors $Z_t$ are usually assumed to be observable, so that $Z_t$ is known at time $t$. It is convenient to consider the series of risk factor changes $X_t = Z_t - Z_{t-1}$. These are objects of interest in statistical studies of financial time series. The value of a portfolio at time $t$ or the asset price $P(t)$ are examples of such risk factors. The stationary log-return process

$$X_t = \log P(t + \Delta) - \log P(t)$$
that reflects the volatility of the prices at time point \( t = 1, 2, \ldots, T \) is often investigated.

Value at risk (VaR) and expected shortfall are accepted as the standard measures of market risk (McNeil et al., 2005). These are nothing more than a high quantile and the expectation of the loss distribution. The PDF of the loss distribution (e.g., the PDF of log-returns) is also an object of interest.

To be specific, the VaR at probability level \( p \in (0, 1) \) is the smallest number \( x_p \) such that the probability that the loss \( X \) exceeds \( x_p \) is no larger than \( 1 - p \). Formally,

\[
\text{VaR}_p = F_X^{-1}(p) = \inf \{ x_p \in \mathbb{R} : P(X > x_p) \leq 1 - p \}
\]

is a quantile of the DF \( F_X \) of the corresponding loss distribution within a fixed time horizon \( \Delta \), and \( F_X^{-1} \) is its inverse (Franke et al., 2004; McNeil et al., 2005). It measures the maximum loss which is not exceeded with a given high probability \( p \).

The expected shortfall at confidence level \( p \in (0, 1) \), is defined as

\[
ES_p = E[X | X \geq \text{VaR}_p].
\]

One averages over realizations of \( X \) which are bigger than \( \text{VaR}_p \).

It is important for us now that the estimation of VaR, expected shortfall and the PDF of log-returns depend heavily on the assumption of the underlying distribution. The latter distribution may be estimated by parametric and nonparametric methods.

The parametric approach assumes that the type of distribution is known. For example, one can expect the distribution of log-returns to be normal. Note that for the normal distribution the log-density is a parabola. However, investigation of financial data has shown that the Gaussian model is not reliable for small timescales (Eberlein and Keller, 1995), and that Pareto-like tails are more appropriate for returns (Mikosch, 2004). The deviation from normality is in contrast to the Black–Scholes model which is most frequently used for stock prices.

Eberlein and Keller (1995) and Eberlein et al. (2003) applied the generalized hyperbolic (GH) distribution (hyperbolic distributions are characterized by their log-density being a hyperbola) to the PDF of log-returns and VaR calculation. Assuming independence and identically distributed r.v.s \( X_t \), a maximum likelihood method can be applied to estimate GH parameters (Eberlein and Keller, 1995).

The empirical distribution of daily returns from many financial data is often skewed, having one heavy, and one semi-heavy or more Gaussian-like tail. Taking this into account, a subclass of the GH, the normal inverse Gaussian distribution (Barndorff-Nielsen, 1977), and skew Student’s \( \bar{t} \) distributions were applied in Aas and Haff (2006) to many financial data. The loss distribution can be estimated nonparametrically by integrating the corresponding estimate of the PDF. The PDF

\[14\] See also Chapter 6 on the estimation of high quantiles.
can be estimated by one of the nonparametric methods described in Sections 2.1, 2.2 and Chapters 3, 4 in the case of independent losses.

However, the losses may be dependent or even long-range dependent. Then the regression model based on the observations of the log-density (2.7) or the kernel estimator with a special bandwidth choice (see Section 2.3) can be applied.

Moreover, the losses may be heavy-tailed. The estimation of a heavy-tailed PDF requires a special methodology that is described in Chapters 3 and 4.

### 2.4.2 Telecommunications

Applications in telecommunications have many analogies to finance.

**Overload control**

Such phenomena as file lengths, call holding times, and inter-arrival times between packets may be used for system overload control. For this purpose the VaR and expected shortfall in other interpretations are again the objects of interest. VaR at probability level $p \in (0, 1)$ can be interpreted, for example, as the smallest number $x_p$ such that the probability that the file size $X$ exceeds $x_p$ is no larger than $1 - p$. Expected shortfall shows an average file size among file sizes exceeding the VaR level. Their treatment in an overloaded router can lead to different control procedures.

**Volatility control**

Control of volatility of characteristics of the system such as file lengths, number and duration of call attempts, may be useful both for overload control and intrusion detection. Then the PDF of ‘log-returns’, correctly interpreted, can be used.

**Simulation and generation of random numbers**

The generation of random numbers is widely used in simulation of complex multi-component systems such as the Internet. Moreover, random numbers with heavy-tailed distributions are often required. For this purpose, the correct parametric form of a distribution needs to be found, which is no mean feat.

Alternatively, only general information about the distribution is used in a nonparametric approach, but the form of the distribution is not required. Thus, the nonparametric estimation of the PDF can be useful to construct the generators of heavy-tailed random numbers.

The simulation of the traffic that is sometimes (e.g., during holidays) unobservable requires the estimation of the PDF using past heavy-tailed and dependent data. Then retransformed kernel estimators that use the preliminary data transformation (see Chapters 3 and 4) can be applied.
Multivariare analysis

Analysis of teletraffic data often requires measurement of the dependence between two time series, for example, TCP flow sizes $S \geq 0$ and durations $D \geq 0$. The evaluation of univariate marginal distributions is especially important for the detection of the dependence and the bivariate analysis of data (see Section 1.3.4). A joint PDF $f(y, z)$ of $S$ and $D$ is required to estimate the distribution of the throughput $R = S/D$ that a flow gets using (1.42).

The DF of $R$ is given by

$$F_R(x) = P\{S/D \leq x\} = \int_0^\infty F_S(xz)f_D(z)dz$$

if some r.v.s $S$ and $D$ are independent. Here, $F_S(x)$ is the DF of $S$ and $f_D(x)$ is the PDF of $D$. The estimation of the bivariate analogs of the VaR (bivariate quantiles) requires the preliminary evaluation of the marginal distributions of both $S$ and $D$.

The estimation of the dependence may be used in the intrusion detection, too. The question is what pairs of indices could better eliminate the intrusion. For the latter purpose the PDFs of the volatility of some characteristics of the Internet, such as the number and duration of session attempts, are useful.

2.4.3 Population analysis

In population analysis different characteristics of longevity are of interest. The survival function of the lifetime $T$ of an individual is an example of such a characteristic. It is defined by the formula

$$S(t) = P\{T > t\} = \int_t^\infty f(t)dt,$$

if the PDF $f(t)$ of $T$ exists, and determines the probability of living beyond the age $t$. It is nothing more than a tail probability.

Parametric models of the distribution of $T$ are usually assumed to estimate $S(t)$. The parameters are estimated by the observed lifetimes $T_1, \ldots, T_n$, for example, the lifetimes of $n$ individuals after heart surgery. However, parametric models are not easy to set up, especially when the influence of many risk factors, sometimes hidden, is investigated. In this case, simple models usually do not fit empirical data well, and when one tries to find the parameters for a complex model, in most cases the parameters have excessive uncertainties. Thus, the alternative is to use nonparametric estimators of $S(t)$ or $f(t)$.

Another characteristic is provided by the so-called mortality risk (or hazard rate in technical applications)

$$h(t) = f(t)/S(t)$$
which also requires the existence of the PDF \( f(t) \). The length of lifetime without disease before the age \( d \) can be determined via \( S(t) \) by

\[
\tau = \int_0^d S(t) \, dt
\]

if \( T \) is interpreted as the onset of a disease.

One can again find many analogs to finance. Thus, the quantile of the lifetime distribution (VaR) represents the smallest age \( t_p \) such that the probability that death occurs after \( t_p \) is no larger than \( 1 - p \). The quantiles can be useful for constructing so-called mortality tables. These tables show the number (or proportion) of individuals of a fixed population who died within given age ranges (e.g., 0–5, 5–10, 10–20, \ldots, 90–100).

The expected shortfall has the meaning of the average lifetime of individuals that live beyond age \( t_p \) for a fixed probability level \( p \).

The investigation of the dependence of several indices and risk factors that requires multivariate analysis of r.v.s is also of interest.

The specific uncertainty of the population data is reflected in censoring. For instance, the lifetimes of individuals who are under supervision after an operation cannot be observed completely. Censored data are thus widely used in estimation of the PDF and hazard rate.

## 2.5 Exercises

1. Derive formula (2.11).

2. Derive formula (2.14). \textit{Hint}: replace \( f(x) \) and \( f^{(k)}(x) \) in (2.13) by their Fourier series using (2.12).

3. Derive formula (2.17). To do this, replace \( f(x) \) and \( f^{(k)}(x) \) in (2.16) by their Fourier series using (2.15).

4. Generate \( n = 50 \) and \( n = 100 \) Fréchet distributed r.v.s with \( \gamma = 1.5 \) as shown in Exercise 1 of Chapter 1. Using the sample \( X^n \) generated, calculate a kernel estimate \( \hat{f}_h(x) \) by (2.18) \((d = 1)\). Select the normal PDF, the Epanechnikov kernel and \( \exp(-|x|) \) as kernel function \( K(x) \) (see Section 2.2.1). Calculate the bandwidth \( h \) by formula (2.30). Generate \( N = 50 \) Fréchet samples, each with sample size \( n \). For each sample calculate the following loss functions:

\[
\nu_1 = \frac{1}{n} \sum_{i=1}^n \left( \hat{f}_h(X_i) - f(X_i) \right)^2, \quad \nu_2 = \sup_{i=1,2\ldots,n} |\hat{f}_h(X_i) - f(X_i)|.
\]

Compare the accuracy of the estimates for different kernels by calculating the statistics

\[
\bar{p}_j = \frac{1}{N} \sum_{i=1}^N \nu^j_i, \quad \sigma_j^2 = \frac{1}{N-1} \sum_{i=1}^N \left( \nu^j_i - \bar{p}_j \right)^2, \quad j = 1, 2.
\]
For larger $n$ the values of the statistics should be smaller. This implies better accuracy.

5. Repeat Exercise 4 for the Epanechnikov kernel. Calculate $h$ by the discrepancy methods (2.41) and (2.42). Compare the accuracy of methods in terms of the values of statistics $\rho_j$ and $\sigma_j^2$, $j = 1, 2$.

6. Repeat Exercises 4 and 5 for the standard normal and exponential (with $\lambda = 1$) distributions and sample sizes $n \in \{50, 100\}$.

7. Generate $n = 100$ uniform distributed r.v.s on $[0, 1]$. Calculate the kernel estimate (2.18) ($d = 1$) with Epanechnikov’s kernel and the projection estimate (2.17). Calculate the bandwidth $h$ in the kernel estimator and the smoothing parameter $\gamma$ in the projection estimator by discrepancy methods (2.41) and (2.42). Draw conclusions from the results of a visual analysis.

8. Generate $n = 100$ Fréchet distributed r.v.s with $\gamma = 1$. Calculate the histogram with equal bin width and compare it with the polygram (2.29). Select $\sqrt{n}$ and $n^{1/4}$ as $L$.

9. Simulate an MA(2) process $X_t = Z_t - 0.4Z_{t-1} + 1.1Z_{t-2}$ of length 200, where $\{Z_t\}$ are normally distributed $N(0, 1)$. Estimate the PDF of this process by kernel estimator (2.43). Use Epanechnikov’s kernel and a normal kernel. Calculate the bandwidth by cross-validation method (2.57), (2.58) with $l = 0, 5, 10, 15, 20, 25$. Compare the results for different kernels and values $l$.

10. Repeat Exercise 9 with the discrepancy method for dependent data (see Section 2.3.3) instead of cross-validation.

11. (Dubov, 1998) Simulate $n = 200$ r.v.s $X_1, \ldots, X_n$ with PDF

$$f(x) = \frac{\exp \left(- (x - a_1)^2 / (2\sigma_1^2) \right)}{2\sqrt{2\pi}\sigma_1^2} + \frac{\exp \left(- (x - a_2)^2 / (2\sigma_2^2) \right)}{2\sqrt{2\pi}\sigma_2^2},$$

where $a_1 = 3$, $\sigma_1 = 0.5$, $a_2 = 7$, $\sigma_2 = 1.5$. For this purpose, mix two Gaussian samples of size $n = 100$ and corresponding parameters $a_i$ and $\sigma_i$, $i = \{1, 2\}$ with weights 0.5. Estimate the PDF using the regression model $\varphi^*(x) = b_1 + \sum_{i=2}^{m} b_ix^{i-1}$, based on the observations of the log-density (2.7). The coefficients $b_i$, $i = 1, \ldots, m$, may be selected by the least-squares method. Among possible polynomials with $m = 1, \ldots, N - 1$ select the one that minimizes the variance of the approximation

$$\rho^2 = E \left( \frac{1}{n_R} \sum_{i=1}^{n_R} (\varphi(z_i) - \varphi^*(z_i))^2 \right).$$
For this purpose, find the value of $m$ that minimizes the sample variance of
the approximation

$$\hat{\rho}^2 = \tilde{\sigma}^2 - (1 - 2\tilde{r}_0) \tilde{\sigma}^2 + 2 (1 - \tilde{r}_0) \tilde{\sigma}^2 n_R^{-1} \sum_{i=1}^{n_R} b_i.$$  

Here, $\tilde{\sigma} = n_R^{-1} \sum_{i=1}^{n_R} (y_i - \varphi(z_i))^2, \ z_i \in (X_{(i)}, X_{(i+1)}), \ y_i = \ln f_i, \ \tilde{r}_0 = 1 - \frac{\pi^2}{6(\tilde{\sigma}^2)}$.

12. Repeat Exercise 11 for the MA(2) process.
Heavy-tailed density estimation

In this chapter problems of heavy-tailed PDF estimation are discussed. Three approaches are considered.

1. Combined parametric–nonparametric methods. The ‘tail’ domain of the PDF\(^1\) is fitted by some parametric model, and the main part of the PDF (the ‘body’, i.e., that limited area of relatively small values of an underlying r.v.) is fitted by some nonparametric method such as a histogram. A similar approach involving Barron’s estimator is considered.

2. Variable bandwidth kernel estimators. The optimal accuracy of these estimates as well as their disadvantages for heavy-tailed PDF estimation are discussed.

3. Retransformed nonparametric estimators. These estimators require a preliminary transformation of the initial sample into a new one whose PDF is more convenient for estimation.

\(^1\) We use the quotation marks to indicate the area of relatively small values of the PDF and to distinguish it from the tail of the distribution \(1 - F(x)\).
3.1 Problems of the estimation of heavy-tailed densities

The main features of heavy-tailed distributions may be formulated as follows:

- The heavy tail goes to zero at infinity at a slower than exponential rate.
- Cramér’s condition is violated.
- Sparse observations occur in the tail domain of the distribution.

In many applications, like the evaluation of the risk to be ruined due to a huge amount of claims in insurance and queuing or huge file sizes transferred through a network, just the tail of the distribution (specifically, the behavior of $1 - F(x)$ as $x \to \infty$) is of interest. In order to evaluate a tail, some parametric models of the tail such as, for instance, Pareto or Weibull type models are used. Then the main focus is on the estimation of the tail index $\alpha = 1/\gamma$ (see Section 1.2), the main index of the heaviness of the tail, from a limited number of measurements.

Sometimes it is necessary to evaluate the heavy-tailed PDF as a whole. Experience of estimation with parametric models has shown that some models describe the tails quite well and other models are better for the small-values area of the PDF (Nabe et al., 1998). The estimation of the PDF may become complicated if the distributions of r.v.s are multimodal. Besides, it is difficult to propose the parametric form of a PDF (e.g., from a QQ plot) arising from a frequently changing random load entity as happens, for instance, in such a dynamic environment as the Internet. Figure 1.24 gives an example of a typical situation when the parametric model does fit outliers.

We are looking for the estimates that fit both the ‘tail’ and the ‘body’ of the heavy-tailed PDF well enough. For this purpose, we shall consider a natural joint parametric–nonparametric estimation approach in Section 3.2. This combines the advantages of parametric tail models to describe the ‘tail’ well enough and of nonparametric methods to describe the ‘body’ domain in a good way, (Markovitch and Krieger, 2002a).

Similar ideas were proposed by Barron et al. (1992) (see Section 3.3) and by Horváth and Telek (2000). However, in the latter paper the ‘boundary’ between the ‘tail’ and the ‘body’ of the PDF is assumed to be a fixed point that is independent of the sample. According to Markovitch and Krieger (2002a), this ‘boundary’ is a random variable, for example, some empirical quantile. In Barron’s estimator $\hat{f}_B(x)$ (see Section 3.3) this is the largest observation. The estimator $\hat{f}_B(x)$ combines a histogram with some parametric tail model. Although it is simple to calculate, this estimate is very sensitive to the choice of the parametric tail model and fits the ‘body’ of the PDF rather poorly, especially for samples of moderate sizes.

Another approach that we consider is a pure nonparametric estimator. In this case, the form of the distribution is not available, but just common information, e.g., the smoothness of the distribution is known. Among known nonparametric
estimators (histogram, projection, kernel estimators; see Section 2.2) just the kernel estimators are defined on the whole real line and, therefore, can be applied directly to estimate heavy-tailed PDFs. However, since the smoothing parameter (i.e., the bandwidth \( h \) of a kernel estimator) is fixed across the entire sample, kernel estimators cannot fit both the ‘tail domain’ and the ‘body’ of the PDF well enough, (Silverman, 1986).

In dealing with heavy-tailed PDFs it is natural to use the window width of the kernels that vary from one point to another. This approach is determined by the application of the variable bandwidth kernel estimator (Abramson, 1982; Devroye and Györfi, 1985; Hall and Marron, 1988; Hall, 1992).

An alternative to the variable bandwidth kernel estimator is given by a transform-retransform scheme or a preliminary transformation of the data and the estimation of the PDF of a new r.v. obtained by the transformation. We shall consider all these approaches below.

### 3.2 Combined parametric–nonparametric method

We use independent observations \( X^n = \{X_1, \ldots, X_n\} \) of some positive r.v. \( x \), for example, inter-arrival times between events, one-way delay or round-trip time measurements in high-speed networks, or session durations and file sizes during Web data transfer. These observations are governed by a common DF \( F(x) \) and PDF \( f(x) = F'(x) \). We assume that \( f(x) \) is heavy-tailed. For example, \( f(x) \) may belong to a mixture of long-tailed PDFs with regularly varying tail \( (1.4) \) as \( x \to \infty \), \( \gamma > 0 \).

We wish to estimate the whole PDF \( f(x) \) using a sample \( X^n \) of moderate size \( n \). For this purpose, we employ here the idea of a separate estimation of the ‘tail’ and ‘body’ of the PDF. Such a separation is reasonable since the risks related to the estimation of the tail and the body of the PDF have qualitatively different values.

We consider the combined estimate

\[
\tilde{f}(t, \gamma, N) = \begin{cases} 
 f_N(t), & t \in [0, X_{(n-k)}), \\
 f_\gamma(t), & t \in [X_{(n-k)}, \infty).
\end{cases}
\]  

(3.1)

Here \( X_{(n-k)} \) is some r.v. defined subsequently and \( f_N(t) \) is some nonparametric estimate of \( f(t) \) on the interval \( [0, X_{(n-k)}) \). The latter is represented by a finite series expansion in terms of trigonometric functions, for example, \( \varphi_k(t) = (4/\pi)^{1/2} \cos((2k-1)(\pi/2)t) \), \( t \in [0, 1] \), \( k = 1, 2, \ldots \):

\[
f_N(t) = \frac{1}{X_{(n-k)}} \sum_{j=1}^{N} \lambda_j \varphi_j \left( \frac{t}{X_{(n-k)}} \right).
\]  

(3.2)

Here \( N \) is a smoothing parameter determining the complexity of the estimate. Let

\[
f_\gamma(t) = (1/\gamma)t^{-1/\gamma-1} + (2/\gamma)t^{-2/\gamma-1}
\]  

(3.3)
be an estimate of the ‘tail’ of the PDF \( f(t) \). The estimate \( f^N(t) \) is calculated by that part of the sample located on the interval \([0, X_{(n-k)}]\) and a tail estimate \( f_\gamma(t) \) is calculated by the rest of the sample.

The estimate (3.1) may not be a real PDF, that is, \( \int_0^\infty \tilde{f}(t, \gamma, N) \, dt \neq 1 \) and \( \tilde{f}(t, \gamma, N) < 0 \) for some \( t \). In this case, one can take

\[
f^*(t, \gamma, N) = \begin{cases} \frac{\tilde{f}(t, \gamma, N)}{\int_0^\infty \tilde{f}(t, \gamma, N) \, dt}, & t \in A, \\ 0, & t \notin A, \end{cases}
\]

(3.4)

for \( A = \{ t \in [0, \infty) : \tilde{f}(t, \gamma, N) > 0 \} \) instead of \( \tilde{f}(t, \gamma, N) \) to provide \( \int_0^\infty f^*(t, \gamma, N) \, dt = 1 \) and \( f^*(t, \gamma, N) \geq 0 \). If

\[
\int_0^{X_{(n-k)}} f^N(t) \, dt = 1
\]

(3.5)

holds, then

\[
\int_0^\infty \tilde{f}(t, \gamma, N) \, dt = 1 + X_{(n-k)}^{-1/\gamma} + X_{(n-k)}^{-2/\gamma}
\]

(3.6)

follows.

The EVI \( \gamma \) is the most important parameter to describe the shape of the tail (see Section 1.2). In Hill’s estimator and many others (see Section 1.2.3) the parameter \( \gamma \) is calculated by the \( k + 1 \) largest values of the order statistics \( X_{(1)} \leq \ldots \leq X_{(n)} \) of the sample \( X^n \). The parameter \( k \) indicates \( X_{(n-k)} \) and, therefore, the part of the distribution which controls the extreme values of the underlying r.v. It can be estimated by different methods described in Section 1.2, e.g., by the bootstrap technique. One has first to estimate \( k \) to fit the ‘tail’ and adapt then the ‘body’ of the PDF as described in Section 3.2.1.

**Remark 1** The boundary between two parts \( f^N(t) \) and \( f_\gamma(t) \) requires additional smoothing to avoid a gap at the point \( X_{(n-k)} \). This can be done, for instance, by means of a kernel that has a special boundary property linking it with \( f_\gamma(X_{(n-k)}) \) if a kernel estimator is used as \( f^N(t) \). One can propose joint conditions for estimate (3.2) as a slope line between two points \((X_{(n-k-1)}, f^N(X_{(n-k-1)}))\) and \((X_{(n-k)}, f_\gamma(X_{(n-k)}))\), namely,

\[
\tilde{f}(t, \gamma, N) = \begin{cases} f^N(t), & t \in [0, X_{(n-k-1)}], \\ \frac{f_\gamma(X_{(n-k)}) - f^N(X_{(n-k-1)})}{X_{(n-k)} - X_{(n-k-1)}} (t - X_{(n-k-1)}) + f^N(X_{(n-k-1)}), & t \in (X_{(n-k-1)}, X_{(n-k)}), \\ f_\gamma(t), & t \in [X_{(n-k)}, \infty). \end{cases}
\]

The combined estimate has the disadvantage that it is necessary to select an appropriate tail model. It has two advantages: First, having selected an appropriate tail model, the ‘tail’ of the PDF can be accurately evaluated (Figure 3.3 shows an example of the poor selection of the tail model). Second, one is free to select many combinations of nonparametric and parametric estimator.
3.2.1 Nonparametric estimation of the density by structural risk minimization

To estimate the PDF on the interval \([0, X_{(n-k)})\), we shall use here a technique for dependence reconstruction (Vapnik and Stefanyuk, 1979; Stefanyuk, 1984). In the following, let \(n^* = n - k\). We suppose, without loss of generality, that the unknown PDF \(f(t)\) is continuous and located on \([0, 1]\). For this purpose, we transform that part of a sample \(X^n\) located on the interval \([0, X_{(n-k)})\) to \([0, 1)\) by the formula:

\[
t_{n-k-i+1} = \frac{X(i)}{X(n-k)}, \quad i = 1, \ldots, n-k.
\]

Constructed from the given data \(X^n = \{X_1, \ldots, X_n\}\), the empirical DF (2.9) is calculated at some independent uniformly distributed points \(\tau_1, \ldots, \tau_l\), i.e. \(\tau_i = i/(l+1), i = 1, \ldots, l\). Hence, we get the data \(\{(\tau_i, y_i), \ldots, (\tau_i, y_i)\}\), where \(y_i = F_n(\tau_i) = F(\tau_i) + \xi_i\) is assumed, that is, the values of the empirical DF are considered as the values \(F(\tau_i)\) of the original DF corrupted by the noise \(\xi_i\). Since the values \(y_i\) are correlated, the noise is correlated, too. The empirical DF is an unbiased estimate of \(F(t)\), therefore \(E\xi_i = 0\). The variance of the noise depends on the value of \(F(t)\), that is, it changes from one point to the next. Considering an optimal adaptation to the data in this case, it is recommended by the theory of the least-squares method to minimize the empirical risk

\[
(\eta - G(\alpha))^T (\eta - G(\alpha)) = (y - F(\alpha))^T R_y^{-1} (y - F(\alpha))
\]

instead of carrying out a minimization of the form \((y - F(\alpha))^T (y - F(\alpha))\). Here \(R_y\) is the covariance matrix of the vector \(y = (y_1, \ldots, y_l)^T\), where \(\eta = By\), \(G(\alpha) = BF(\alpha), B^T B = R_y^{-1}\) and \(F(\alpha) = (F(\tau_i, \alpha), \ldots, F(\tau_i, \alpha))^T\). The matrix \(B\) transforms the correlated observations \(\{y_i\}\) into the uncorrelated \(\{\eta_i\}\) with variance equal to 1. All asymptotic properties of the least-squares method, unbiasedness and minimal variance of the linear, unbiased estimates, are preserved.

The estimation of the PDF from data with correlated or independent noise is different. If we are dealing with independent data, increasing the sample size gives us more accurate estimates. If we use correlated data, then increasing the number of observations is subject to diminishing returns – the correlated points may ‘repeat’ and fail to provide any new information. The use of the covariance matrix takes into account the co-location of different parts of the distribution – the structure of the PDF – and helps to estimate multimodal PDFs (PDFs of mixtures of distributions). The idea of using correlated data is essentially used in a variety of methods (Čencov, 1982; Dubov, 1998; Kooperberg et al., 1994; Vapnik, 1982).

---

2 The asymptotical normality of \(\{\eta_i\}\) may be proved. Uncorrelated, normally distributed \(\{\eta_i\}\) are independent. The independence is required for the further application of the structural risk-minimization method.
Following this idea of transforming the correlated observations \( \{(\tau_i, y_i), i = 1, \ldots, l\} \), we will now determine the estimate

\[
g^N(t) = \sum_{j=1}^{N} \lambda_j \varphi_j(t)
\]

of the ‘body’ of the PDF by an application of the structural risk-minimization method; see Vapnik and Stefanyuk (1979) and Vapnik (1982). Let \( G(\tau, \alpha) = \int_{\tau_i} g^N(t) \, dt \) be the corresponding DF estimate and \( \lambda = (\lambda_1, \ldots, \lambda_N)^T \).

The original structural risk-minimization method requires independent measurements \( (\tau_i, y_i) \). The idea of this method is to formulate the optimal estimation by means of the given data as minimization of the mean risk

\[
\int (y - G(\tau, \alpha))^2 \, dP(\tau, y) \to \min_{\alpha},
\]

where the measure \( P(\tau, \eta) \) is unknown and \( \alpha = (\lambda, N) \) is the vector of the parameters. This task is performed by the minimization of its upper bound

\[
J = g(N, l) \frac{1}{I} \sum_{i=1}^{l} (y_i - G(\tau_i, \alpha))^2,
\]

(3.8)

where the penalty function \( g(N, l) \) depends on the Vapnik–Chervonenkis dimension and may have different forms for different classes of models. Such types of bounds follow from fundamental estimates of the deviation of the mean risk from its empirical analogue (for further details, see Vapnik, 1982).

Following the arguments leading to (3.7) in the case of correlated points, instead of \( J \) in (3.8) the minimization of

\[
J(N, \lambda) = \left[ \frac{l^{-1}(Y - F(\lambda))^T R^{-1}_\gamma(Y - F(\lambda))}{1 - \sqrt{l^{-1}((N+1)(1+\ln l - \ln(N+1)) - \ln \eta)}} \right]_\infty
\]

(3.9)

with respect to \( N \) and \( \lambda \) is used, where \( \eta > 0 \) is a confidence level,

\[
[z]_\infty = \begin{cases} z, & z > 0, \\ \infty, & z \leq 0, \end{cases}
\]

and

\[
Y = (Y_1, \ldots, Y_l)^T, \quad Y_i = y_i - \int_{0}^{\tau_i} \frac{\varphi_i(t)}{\psi_1} \, dt.
\]

By the choice of an optimal complexity \( N \) for a given number \( l \) of points \( \tau_i \) the structural risk-minimization method selects the values of these parameters which provide a lower minimum of the mean risk than those parameters corresponding to the minimum of the empirical risk.
As $y_i$ one can take the estimate $\Phi_n(t_i)$ of the unknown DF $F(t)$ at $t_i$ determined below in (3.11). Furthermore, we use:

$$F(\lambda) = (F_1^\lambda, \ldots, F_l^\lambda)^T, \quad F_i^\lambda = \int_0^{\tau_i} \left( \sum_{j=2}^{N} \lambda_j (\varphi_j(t) - \frac{\psi_j}{\psi_1} \varphi_1(t)) \right) dt,$$

$$F(\lambda) = A \cdot \lambda_1.$$ 

Here the elements of the $l \times (N - 1)$ matrix $A$ are given by

$$A_{i,j} = \int_0^{\tau_i} \left( \varphi_j(t) - \frac{\psi_j}{\psi_1} \varphi_1(t) \right) dt,$$

$$\psi_j = \int_0^1 \varphi_j(t) dt, \quad i = 1, \ldots, l; \ j = 2, \ldots, N.$$ 

$\Lambda_1$ is the $(N - 1) \times 1$ vector of parameters $\lambda_j, j = 2, \ldots, N$. The matrix

$$R_y^{-1} = \begin{pmatrix} r_1 & \rho_1 & 0 & \ldots & \ldots & 0 \\ \rho_1 & r_2 & \rho_2 & 0 & \ldots & \vdots \\ 0 & \ldots & \ldots & \ldots & \ldots & \vdots \\ \vdots & \ldots & \ldots & \ldots & 0 & \rho_{l-2} \\ 0 & \ldots & 0 & \rho_{l-1} & \rho_{l-1} & r_l \end{pmatrix}$$

(3.10)

with

$$r_1 = \frac{n^* F(\tau_2)}{F(\tau_1) (F(\tau_2) - F(\tau_1))}, \quad r_i = \frac{n^* (1 - F(\tau_{i-1}))}{(1 - F(\tau_i)) (F(\tau_i) - F(\tau_{i-1}))},$$

$$r_{i-1} = \frac{n^* (F(\tau_i) - F(\tau_{i-2}))}{(F(\tau_i) - F(\tau_{i-1})) (F(\tau_{i-1}) - F(\tau_{i-2}))}, \quad i = 3, 4, \ldots, l,$$

$$\rho_i = -\frac{n^*}{(F(\tau_{i+1}) - F(\tau_i))}, \quad i = 1, 2, \ldots, l - 1,$$

is used. However, the following estimate $\Phi_n(t)$ is used instead of the unknown DF $F(t)$:

$$\Phi_n(t) = \begin{cases} \frac{1}{2n^*} \left( \frac{t}{\tau_1} \right), & 0 < t \leq t_1 \\ \frac{m - 0.5}{n^*} + \frac{1}{n^*} \left( \frac{t - t_m}{t_{m+1} - t_m} \right), & t_m < t \leq t_{m+1}, \quad m = 1, \ldots, n^* - 1 \\ \frac{n^* - 0.5}{n^*} + \frac{1}{2n^*} \left( \frac{t - t_{n^*}}{1 - t_{n^*}} \right), & t_{n^*} < t \leq 1 \end{cases}$$

(3.11)
The minimization algorithm has two stages:

1. The DF $F(t)$ and $R_y^{-1}$ are estimated from the sample using (3.11) and (3.10).

2. In (3.9) $R_y^{-1}$ is replaced by its estimate and the parameters of the PDF estimate $g(t)$ are obtained by the minimization of $J(N, \lambda)$ with respect to $N$ and $\lambda$.

   The method preserves $\int_0^1 \sum_{j=1}^N \lambda_j \varphi(t) dt = 1$.

**Computational notes**

1. Let $\eta = 0.05$.

2. Stefanyuk (1984) recommended selecting $l = 5n/\ln n$ to provide the asymptotic minimum of the $L_2$ error as $n \to \infty$.

3. To avoid division by zero in the formula (3.11) of the estimate $\Phi_n(t)$ of the empirical DF, the points $\{t_m, m = 1, \ldots, n^*\}$ cannot repeat each other.\(^3\)

4. $\lambda_1$ is calculated by

   $$\lambda_1 = \frac{1 - \sum_{j=2}^N \lambda_j \psi_j}{\psi_1}.$$  

5. One minimizes the empirical risk $l^{-1}(Y - A\Lambda_1)^T R_y^{-1}(Y - A\Lambda_1)$ over $\Lambda_1 = (\lambda_2, \ldots, \lambda_N)^T$ for each fixed $N$. The minimum gives the following estimate:

   $$\lambda^*_N = (A^T R_y^{-1} A)^{-1} A^T R_y^{-1} Y.$$  

   Among the vectors $\lambda^*_N, N = 2, 3, \ldots, N_{\text{max}}$ (where $N_{\text{max}}$ is the maximum value of $N$ considered) one selects those corresponding to the minimum of $J(N, \lambda)$.

6. The empirical risk (the numerator of (3.9)) has to decrease with increasing $N$. If this risk increases, then the matrix of the system is nearly singular.

7. The minimum of (3.9) is not necessarily reached for a maximal $N$. For such $N$ the empirical risk is minimal, but the inverse denominator of (3.9) is maximal.

8. Usually, $2 \leq N \leq 20$.

9. Finally, the ‘body’ estimate of the PDF is calculated by the formula

   $$f^N(x) = \frac{1}{X_{(n-k)}} \sum_{j=1}^N \lambda_j \varphi_j \left( \frac{x}{X_{(n-k)}} \right),$$

   with $x \in [0, X_{(n-k)}]$.

\(^3\) For continuous $F(x)$, repetitions are impossible.
10. One can use another complete system of basis functions $\varphi_k(t), k = 1, 2, \ldots$, instead of the trigonometric functions.

Remark 2 Kernel estimates typically do not fit all modes of a multimodal PDF well enough. Usually, a kernel estimate over-smoothes one mode and fits another well, if, for instance, a mixture of two normal distributions is considered. Vapnik and Stefiyuk (1979) found that the approach considered works better than kernel estimates for the estimation of multimodal PDFs (see Figure 3.3).

3.2.2 Illustrative examples

To demonstrate the power of the combined estimator

$$\tilde{f}(t, \gamma, N) = \begin{cases} \sum_{j=1}^{N} \lambda_j \varphi_j(t), & t \in [0, X_{(n-k)}], \\ f_{\gamma}(X_{(n-k)}) - f^N(X_{(n-k-1)}) \left( t - X_{(n-k-1)} \right) / \left( X_{(n-k)} - X_{(n-k-1)} \right), & t \in (X_{(n-k-1)}, X_{(n-k)}), \\ (1/\gamma)t^{1/\gamma-1} + (2/\gamma)t^{-2/\gamma-1}, & t \in [X_{(n-k)}, \infty) \end{cases}$$

(3.13)

and its ability to estimate long-tailed PDFs and their mixtures, some illustrative examples, motivated by the measurements in Bolotin et al. (1999) and Roppel (1999), are presented.

For this purpose we have generated samples which follow a mixture of two distributions, i.e. the PDF is determined by $f(x) = 0.5f_1(x) + 0.5f_2(x)$. Particularly, we consider mixtures of

- a Burr distribution, $\text{Burr}(\gamma, \rho) = \text{Burr}(0.8, -2)$, with PDF $f_1(x) = \lambda \tau x^{\tau-1} (1 + x^{\tau})^{-\lambda-1}, \quad x > 0, \lambda > 0, \tau > 0$, and $\gamma = \frac{1}{\lambda \tau}, \rho = -\frac{1}{\lambda}$, and a gamma distribution $\text{Ga}(\alpha) = \text{Ga}(9)$ with PDF $f_2(x) = x^{\alpha-1} \exp(-x) / \Gamma(\alpha), \quad x > 0, \alpha > 0$

(see Figure 3.1);

- a Gamma distribution $\text{Ga}(2.5)$ as $F_1(x)$ with $f_1(x) = \frac{dF_1(x)}{dx}$ and a Pareto distribution with PDF $f_2(x) = \frac{\alpha}{k} \left( \frac{k}{k+x} \right)^{\alpha+1}$ and $k = 1, \alpha = 0.3$ (see Figure 3.2);

- the Gamma distributions $\text{Ga}(1.9)$ and $\text{Ga}(10)$ (see Figure 3.3).
Figure 3.1 Estimation of the PDF of a mixture of a gamma and a Burr distribution (dotted line) by the combined estimate (3.13) (solid line) and the kernel estimate with Epanechnikov’s kernel (dashed line): ‘body’ reconstruction (left); ‘tail’ reconstruction (right). The bandwidth is selected by the over-smoothing method (2.30).

Figure 3.2 Estimation of the PDF of a mixture of gamma and a Pareto distribution (dotted line) by the combined estimate (3.13) (solid line): ‘body’ reconstruction (left); ‘tail’ reconstruction (right). The over-smoothing bandwidth is given by $h = 1.279 \times 10^7$. The kernel estimate is over-smoothed and not presented.

Figure 3.3 Estimation of the PDF of a mixture of two gamma distributions (dotted line) by combined estimate (3.13) (solid line) and the kernel estimate with Epanechnikov’s kernel (dashed line): ‘body’ reconstruction (left); ‘tail’ reconstruction (right). The bandwidth is selected by the over-smoothing method (2.30). The tail model is wrongly selected.
Note that the first two mixtures are heavy-tailed due to the presence of Burr and Pareto distributions, and the last mixture is light-tailed. In all cases the sample size \( n = 200 \) has been used. The bootstrap values \( k_1 \in (5, 12, 4) \) and \( k \in (29, 68, 23) \) (see formulas (1.10)–(1.12)) in Hill’s estimator and the number of terms \( N \in (18, 4, 11) \) in expansion (3.2) corresponding to the minimum of \( J(N, \lambda) \), have been selected for the mixtures mentioned.

**Typical mistakes**

1. Sometimes the mixtures look like unimodal distributions (Figure 3.2). Therefore one tries to find an appropriate parametric model among well-known distributions, which is difficult.

2. Figure 3.3 demonstrates the wrong selection of the tail model and the estimator of the tail index. The gamma tail is lighter than the Pareto tail (3.3) that is used in (3.13). The Hill estimator cannot be applied here since the tail index of the light-tailed mixture of gamma distributions is negative. Due to these mistakes, the gap between the nonparametric and the parametric part is visible. This example demonstrates that a rough investigation of the heaviness of tails (Section 1.3.1) is necessary before estimation.

3. The kernel estimate with compactly supported Epanechnikov kernel (i.e., the kernel is defined on a finite interval) is truncated beyond the range of the sample and cannot be used to estimate the tail. In order to estimate the tail of the PDF better a specific transformation of the data is proposed (see Section 4.3).

**3.2.3 Web data analysis by a combined parametric–nonparametric method**

We apply the combined estimator (3.1) to four Web data characteristics described in Section 1.3.2. To simplify the calculations, the data were scaled, that is, the values were divided by the scaling parameter \( s \) (see Table 1.4). The values of the parameters of the combined estimate are presented for each r.v. in Table 3.1. Here, \( k_1 \) provides the minimum value of (1.12), and \( k \) is calculated by (1.10) and (1.11). In these formulas we take \( \alpha = 2/3, \beta = 1/2 \). The corresponding order statistic \( X_{n-k_1} \), which is, roughly speaking, the ‘boundary’ between the ‘tail’ and the ‘body’ of the PDF (see (3.1)) as well as Hill’s estimate of the tail index \( \hat{\gamma}_H(n, k) \) are also presented. The number of terms \( N \) in the expansion (3.2) provides the minimum of the functional (3.9). The vectors of the coefficients \( \lambda \) in (3.2) calculated by (3.12) are given in Table 3.2.

The combined estimates are presented in Figures 3.4–3.7. Each figure consists of two graphs to demonstrate better the behavior on the ‘tails’ and ‘bodies’ of the corresponding PDFs. The scaled values \( x/s \) are presented on the \( x \)-axis. Calculated by the formula \( f(x) = (1/s)g(x/s) \), the values of the PDF estimates are presented on
Table 3.1  Parameters of the combined estimate.

<table>
<thead>
<tr>
<th>r.v.</th>
<th>$k_1$</th>
<th>$k$</th>
<th>$X_{(n-k)}/s$</th>
<th>$\gamma^H(l,k)$</th>
<th>$N$</th>
<th>$s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>s.s.s.</td>
<td>6</td>
<td>42</td>
<td>0.214</td>
<td>0.952</td>
<td>15</td>
<td>$10^7$</td>
</tr>
<tr>
<td>d.s.s.</td>
<td>4</td>
<td>28</td>
<td>4.071</td>
<td>0.6</td>
<td>12</td>
<td>$10^3$</td>
</tr>
<tr>
<td>s.r.</td>
<td>4</td>
<td>31</td>
<td>0.108</td>
<td>0.615</td>
<td>8</td>
<td>$10^6$</td>
</tr>
<tr>
<td>i.r.t.</td>
<td>9</td>
<td>70</td>
<td>0.109</td>
<td>1.001</td>
<td>4</td>
<td>$10^3$</td>
</tr>
</tbody>
</table>


Table 3.2  Vectors of optimal coefficients.

<table>
<thead>
<tr>
<th>r.v.</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>s.s.s.</td>
<td>(1.584, 0.929, 0.688, 0.384, 0.365, 0.297, 0.373, 0.381, 0.365, 0.289, 0.225, 0.186, 0.263, 0.159, 0.153)$^T$</td>
</tr>
<tr>
<td>d.s.s.</td>
<td>(1.542, 0.635, 0.405, 0.175, 0.108, 0.172, 0.174, 0.157, 0.173, 0.125, 0.219, 0.189)$^T$</td>
</tr>
<tr>
<td>s.r.</td>
<td>(1.621, 1.079, 0.862, 0.590, 0.486, 0.329, 0.300, 0.076)$^T$</td>
</tr>
<tr>
<td>i.r.t.</td>
<td>(1.561, 0.763, 0.512, 0.121)$^T$</td>
</tr>
</tbody>
</table>


Figure 3.4  Estimation of the PDF of the sub-session size by the combined estimate. Reprinted from *Computer Networks*, 40(3), pp. 459–474, The estimation of heavy-tailed probability density functions, their mixtures and quantiles, Markovitch NM and Krieger UR, Figure 4, © 2002 Elsevier. With permission from Elsevier.
the y-axis. Here, \( g(y) \) is the estimate (3.1) resulting from the scaled data \( Y_i = X_i / s \), where \( X_i, i = \{1, \ldots, n\} \), are empirical measurements. Since \( \hat{\gamma}^H(n, k) > 0 \) for all r.v.s considered, one may conclude that their distributions have heavy tails. However, d.s.s. and s.r. have larger \( 1/\gamma \) and, hence, heavier tails, than s.s.s. and i.r.t.

### 3.3 Barron’s estimator and \( \chi^2 \)-optimality

A similar approach to a combined parametric–nonparametric method is realized in Barron et al. (1992). Specifically, let \( P_n = \{A_{n1}, \ldots, A_{nm_n}\} \) be partitions of the real line \((0, \infty)\) into finite intervals (bins) by quantiles \( G^{-1}(j/m_n), 1 \leq j \leq m_n - 1, \)
of an arbitrary distribution $G(x)$, $\delta_j = \int_{A_{nj}} dF_n(x) = (1/n) \sum_{i=1}^{n} 1_{A_{nj}}(X_i)$, $F_n(x)$ an empirical DF, $n$ the sample size. The estimator is defined as follows:

$$\hat{f}_B(x) = g(x) \frac{1/n + \delta_j}{1/n + 1/m_n}, \quad x \in A_{nj}, \quad 1 \leq j \leq m_n. \quad (3.14)$$

The consistency of the estimate is provided by the conditions $m_n \to \infty$ and $m_n/n \to 0$ as $n \to \infty$. The behavior of the DF beyond the range of the sample, as $x > X_{(n)}$, is unknown. Therefore, one has to use the asymptotic models of the DF, which follow from the extreme value theory (Gnedenko, 1943). In the estimate $\hat{f}_B(x)$, different parametric models are selected as DF $G(x)$ (e.g., lognormal, normal, Weibull distributions). The parameters of these models may be estimated by the ML or moment methods. Indeed, the choice of $g(x) = G'(x)$ has a strong effect on the estimate of the DF in the tail domain, when $x \in A_{nm_n} = (G^{-1}((m_n - 1)/m_n), \infty)$ (this is the area of sparse observations), since

$$\hat{F}(x) = \frac{1/n + \delta_{m_n}}{1/n + 1/m_n} \int_{-\infty}^{x} g(x) dx = \frac{1/n + \delta_{m_n}}{1/n + 1/m_n} (1 - G(x))$$

holds. In Berlinet et al. (1998) the consistency of $\hat{f}_B(x)$ in a sense of the $\chi^2$-distance under some assumptions on the PDF is proved. Two problems related to an optimal choice of partitions and the auxiliary function $g(x)$ are solved in Vajda and van der Meulen (2001). Following Györfi et al. (1998), upper bounds of the mean $\chi^2$-distance (2.5) over some classes of PDFs are minimized to find an optimal $m_n$.

By the way, a similar estimate

$$\hat{f}_s(x) = \sum_{i=1}^{m_n} \frac{\alpha + \delta_i}{\alpha + 1/m_n}, \quad x \in A_{ni},$$

with the separation of the domain of definition domain of PDF $f(x)$ into equal partitions has been obtained by a regularization method in Stefanyuk and Karandeev.
(1996). However, this estimate is intended for finite PDFs. A Bayesian approach to the choice of parameters, namely, the smoothing parameter $\alpha$ and the number of intervals $m_n$, has been used. This approach provides the minimum of the MISE for known prior distributions (Stefanyuk and Karandeev, 1996).

Barron’s estimator refers from two disadvantages. First, it is necessary to select an appropriate tail model. Second, the tail model $g(x)$ distorts the estimate of the ‘body’ of the PDF for samples of moderate size. This influence becomes weaker as the sample size increases (Küs and Vajda, 1996).

### 3.4 Kernel estimators with variable bandwidth

If $F(x)$ is heavy-tailed with PDF $f(x)$, the well-known PDF estimators such as the histogram and the kernel estimator perform quite poorly. Since the smoothing parameter (e.g., the bandwidth $h$ of a kernel estimator) is fixed across the entire sample, these estimators may provide a misleading estimation in the tail domain or over-smooth the body of the PDF. Some examples are given for suicide data by Silverman (1986, p.18) and in Figures 3.3 and 3.8.

To overcome this problem it is natural to use kernel estimators with kernels that vary from one point to another – the so-called variable bandwidth kernel estimator (Abramson, 1982; Devroye and Györfi, 1985; Hall and Marron, 1988; Hall, 1992)

$$
\hat{f}^A(x|h) = (nh)^{-1} \sum_{i=1}^{n} f(X_i)^{1/2} K \left( \frac{x - X_i}{f(X_i)^{1/2} / h} \right).
$$

Since $f(X_i)$ is unknown, the estimator

$$
\tilde{f}^A(x|h_1, h) = (nh)^{-1} \sum_{i=1}^{n} \hat{f}_{h_1}(X_i)^{1/2} K \left( \frac{x - X_i}{\hat{f}_{h_1}(X_i)^{1/2} / h} \right)
$$

(3.15)

Figure 3.8  Kernel estimates with Epanechnikov’s kernel and different bandwidth values $h$ for a Fréchet PDF with shape parameter $\gamma = 1.5$ (solid line): $h = 0.05$ (dotted line), $h = 1$ (dot-dashed line).
is used in practice. Usually, the nonvariable bandwidth kernel estimator (2.18) is used as a pilot estimator \( \hat{f}_h(x) \).

The variable bandwidth kernel estimator \( \hat{f}^A(x|h) \) provides the mean squared error

\[
\text{MSE}(\hat{f}^A(x)) = E(\hat{f}^A(x|h) - f(x))^2
\]

\[
= h^8 \left( \frac{K_3}{24} \right)^2 \left( \frac{d}{dx} \right) ^4 \frac{1}{f(x)} + f(x)^{3/2} \frac{c}{nh} + o((nh)^{-1} + h^8)
\]

as \( h \to 0 \) uniformly in \( x \in R \), if \( f(x) \) has four continuous derivatives and is bounded away from zero on \( R^e \equiv \{ x \in R : \text{for some } y \in R, \| x - y \| \leq \varepsilon \} \) (Silverman, 1986; Hall and Marron, 1988). Here, \( c = \int K^2(t)dt \) and \( K_3 \) is determined by Definition 14. Hence, the fastest achievable order \( n^{-8/9} \) of the MSE is attained if \( f(x) \) has four continuous derivatives, the kernel function \( K(x) \) satisfies

\[
\int t^4 |K(t)|dt < \infty, \quad \int K(x)dx = 1, \quad \sup_x |K(x)| < \infty,
\]

and \( h = c_1 n^{-1/9} \) for some constant \( c_1 \). For a nonvariable kernel estimator this improves to \( n^{-4/5} \) (Hall and Marron, 1988). Since the variance of any kernel estimate has rate \( O(1/(nh)) \), as \( nh \to \infty \), then the reduction of MSE arises by the reduction of the bias \( E\hat{f}_h(x) - f(x) \) which has order \( h^4 \) for the variable bandwidth kernel estimates.

It is proved by Hall and Marron (1988) for the estimator (3.15) used in practice that

\[
\tilde{f}^A(x|h_1, h) = \hat{f}^A(x|h) + cZ(nh)^{-1/2} + o((nh)^{-1/2}),
\]

where \( c \) is a constant, \( Z \) is a standard normal r.v., and \( h_1 \approx n^{-1/5} \). The value \( c = c(h_1) \) may be obtained from Hall and Marron’s formula (4.5) and the application of Lindeberg’s theorem to \( \tilde{f}^A(x|h_1, h) - \hat{f}^A(x|h) \), which is a sum of i.i.d. r.v.s (Petrov, 1975). Then the bias of \( \tilde{f}^A(x|h_1, h) \) is the same as for \( \hat{f}^A(x|h) \). The variance of \( \tilde{f}^A(x|h_1, h) \) is defined by

\[
\text{var} \left( \tilde{f}^A(x|h_1, h) \right) = \text{var} \left( \hat{f}^A(x|h) \right) + c^2(nh)^{-1} + o((nh)^{-1}),
\]

under the assumption \( E(Z \cdot \hat{f}^A(x|h)) = 0 \). The variance of \( \tilde{f}^A(x|h_1, h) \) is a little larger than the variance of \( \hat{f}^A(x|h) \). However, both are of the same order of magnitude.

The modified estimator

\[
\hat{f}^H(x|h) = \frac{1}{nh} \sum_{i=1}^n f(X_i)^{1/2} K \left( \frac{(X_i - X_j)f(X_j)^{1/2}}{h} \right) 1 \{ |x - X_i| \leq (1 + |x|) (\log h)^{-1} \}
\]

is presented in Hall (1992). This modification prevents very large \( X_i \)’s, that is, \( X_i \)’s a long way from \( x \). The bias of \( \hat{f}^H(x|h) \) has the same rate \( h^4 \).
It is remarkable that none of the estimators mentioned require the assumption that $K(x)$ is nonnegative, because the fourth order (see Definition 14, p. 70) of $K(x)$ (and thus, possible negativity) is not required. This implies that variable bandwidth kernel estimates have the fastest rate of convergence without the disadvantage of negativity. For nonvariable kernel estimates this rate can only be achieved using fourth-order kernels (such as $\int t^2 K(t) dt = 0$) which take negative values and, lead to negative kernel estimates (Silverman, 1986).

It is argued in Hall (1992) that the MISE is not a convenient measure of quality for the estimate $\hat{f}^A(x|h)$, since the asymptotic error $E \int \left( \hat{f}^A(x|h) - f(x) \right)^2 dx$ – or, more precisely, the variance of the estimate – is driven by the tail behavior of $f(x)$. As the tails of the distribution become lighter, the MISE converges to that of an nonvariable kernel estimate, $n^{-4/5}$.

Another important problem is the smoothing or selection of the bandwidth $h$ in kernel estimators. We find that $\text{MSE}\left(\hat{f}^A(x)\right)$ is minimal on

$$h^{\text{opt}} = \left( \frac{F_2}{8F_1} \right)^{1/9} n^{-1/9},$$

where

$$F_1 = (K_3/24)^2 \left( (d/dx)^4 (1/f(x)) \right)^{2}, \quad F_2 = cf(x)^{3/2}. \quad (3.21)$$

For such $h^{\text{opt}} \sim n^{-1/9}$ it evidently follows that $\text{MISE} \sim n^{-8/9}$. Indeed, the parameter $h^{\text{opt}}$ depends on the unknown derivative $(d/dx)^4 (1/f(x))$. The estimation of derivatives of the PDF is a complicated problem in itself. The estimation of an additional derivative is more difficult than estimating an additional dimension. For example, the optimal asymptotic mean integrated squared error (AMISE) rate for the second derivative is $O(n^{-4/9})$ which is the same (slower) rate as for the optimal AMISE of a five-dimensional multivariate frequency polygon PDF estimator (Scott, 1992, p. 132). Hence, for practical computation the data-dependent methods (e.g., cross-validation or the discrepancy method) for the selection of $h$ may be better if one is dealing with samples of moderate size.

The cross-validation method produces consistent nonvariable kernel estimates in the $L_1$ metric in the case of a distribution with bounded support (Chow et al., 1983). For heavy-tailed PDFs cross-validated estimates do not converge since $h \not\to 0$ as $n \to \infty$ (Devroye and Györfi, 1985). In Bowman (1984) the integrated squared error cross-validation method (i.e., a minimization (2.34) with respect to $h$) was found to estimate long-tailed distributions by means of variable bandwidth kernel estimators. It was shown in Schuster and Gregory (1981) by a simulation study that this method produces better estimates than the cross-validation for the Cauchy and Student’s $t_{(5)}$ distributions.

A weighted version of squared error cross-validation for the estimator $\widetilde{f}^A(x|h_1, h)$ was proposed in Hall (1992). According to this method the empirical
version of the functional
\[ WISE = \int \tilde{f}_-(x; h)^2 \omega(x) dx - 2 \int \tilde{f}_-(x; h)^2 f(x) \omega(x) dx \] (3.22)
has to be minimized with respect to \( h \) to choose \( h \). Here,
\[ \tilde{f}_-(x; h) = \frac{1}{nh} \sum_{j=1, j \neq i}^{n} \hat{f}_-(X_j, h_1)^{1/2} K \left( \frac{(x - X_j) \hat{f}_-(X_j, h_1)^{1/2}}{h} \right) \cdot 1 \left( |x - X_j| \leq Ah \right), \quad \forall A > 0, \] (3.23)
and \( \omega(x) \) is a bounded, nonnegative function (a weight). The estimate \( \tilde{f}_-(x; h) \) is the estimate \( \tilde{f}^A(x|h_1, h) \) that is calculated over the sample with one excluded observation.

For this method the optimal order \( n^{-1/9} \) of \( h \) providing the minimum of the MSE was not proved and, hence, the fastest MSE of rate \( n^{-8/9} \) was not obtained. From Theorem 3.1 of Hall (1992, p. 772) it follows that (in Hall’s notation) \( \hat{I} \not\to I \) (\( I = \int \tilde{f}^A f_0 \) is a weighted expectation of \( \tilde{f}^A \), \( \hat{I} \) is an empirical estimate of \( I \)) and hence, \( WISE \not\to WISE \) as \( n \to \infty \).

Novak (1999) and Naito (2001) give estimates that modify \( \hat{f}^A_h(x) \) and have the same bias \( O(h^4) \).

In contrast to the retransformed kernel estimators presented below, the variable bandwidth kernel estimators are not intended for the estimation of the PDF at infinity, at least by compactly supported kernels, because the latter estimators are defined on finite intervals which are approximately the same as the ranges of the samples.

**Example 8** We consider the estimate \( \hat{f}^A_h(x) \) with kernel given by some symmetric finite PDF, such as the Epanechnikov kernel. Due to the restriction \( |x| \leq 1 \) of this kernel, we have \( |x - X_j| \leq h/\hat{f}(X_j)^{1/2} \). Let us use the Gaussian PDF \( N(0, \hat{\sigma}^2) \) as a pilot estimate \( \hat{f}(x) \), where \( \hat{\sigma}^2 \) is the empirical variance. Then the maximal point \( x \) where the estimate may be computed is defined by the inequality
\[ x \leq X_{(n)} + h \left( \sqrt{2\pi \hat{\sigma} \exp \left( X_{(n)}^2/(2\hat{\sigma}^2) \right)} \right)^{1/2}, \]
that is, it depends on the maximal observation \( X_{(n)} \). Let us consider the Pareto distribution with PDF
\[ f(x) = \begin{cases} \alpha x^{-(\alpha+1)}, & x > 1, \\ 0, & x \leq 1, \end{cases} \]
with \( \alpha = 3 \). The variance of this distribution is equal to 3/2. The \( (1 - 10^{-5}) \) 100% quantile is equal to \( t_{(1-10^{-5})} = 46.41 \), the 95% quantile is \( t_{0.05} = 2.714 \). If the whole sample falls into the 95% confidence interval and \( X_{(n)} = t_{0.05} \), then \( x \leq t_{0.05} + \left( \sqrt{3\pi \exp(t_{0.05}^2/3)} \right)^{1/2} \approx 8.7 \) as \( h = 1 \). At the same time the endpoint of the distribution is approximately equal to 46.41.
It seems that the selection of a heavy-tailed symmetric PDF, such as Cauchy, as a kernel could extend the ability of variable kernel estimates with regard to the estimation of the ‘tail’.

### 3.5 Retransformed nonparametric estimators

For finite and light-tailed distributions a histogram is a good estimate of the corresponding PDF. But if the distribution is heavy-tailed, a histogram provides a misleading estimate in the ‘tail’ domain. The same is true for most of the common nonparametric PDF estimates, such as kernel, projection and spline estimates. In general, they have sharp peaks at ‘outliers’ and do not provide the correct rate of decay at infinity (Silverman, 1986). Variable bandwidth kernel estimates with compactly supported kernels are truncated beyond a finite interval that is determined by the largest observation of the sample.

It is obvious that nonparametric PDF estimates with good behavior in the ‘tail’ domain are required. This feature is highly significant if PDFs of many populations are compared. Another problem related to the comparison of PDFs is provided by classification (pattern recognition). 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 (Chapter 5). Since the object can arise in the ‘tail’ domain as well as in the ‘body’, a tail estimator with good properties is of great importance for the classification.

To improve the behavior of the PDF estimation at infinity one can apply a transform–retransform scheme or a preliminary transformation of the data and estimate the PDF of a new r.v. obtained by the transformation. We discuss here the estimation of a heavy-tailed PDF \( f(x) \) using a transform–retransform scheme that is an alternative to a variable bandwidth kernel estimation. This means that first the \( X \)-space data are transformed via the monotone increasing continuously differentiable ‘one-to-one’ transformation function \( T(x) \) to obtain \( Y_1, \ldots, Y_n(Y_i = T(X_i)) \). The derivative of the inverse function \( T^{-1} \) is assumed to be continuous.

The DF of \( Y_j \) is given by

\[
G(y) = P(Y_j \leq y) = P(T(X_j) \leq y) = P(X_j \leq T^{-1}(y)) = F(T^{-1}(y)),
\]

and its PDF is

\[
g_0(y) = G'(y) = f(T^{-1}(y))(T^{-1}(y))'.
\]

Furthermore, the PDF \( g_0(y) \) of \( Y_i \) is estimated by some estimator \( \hat{g}_0(y) \) and after back-transformation we get the PDF estimate of the \( X_i \) by the formula

\[
\hat{f}(x) = \hat{g}_0(T(x))T'(x).
\]

One may take any nonparametric estimator as \( \hat{g}_0(x) \). The PDF \( g_0(x) \) should be convenient for the estimation; for example, it should not go to infinity in the domain of definition. The latter can be ensured by the choice of \( T(x) \).
The background to the transformation idea is given by the need for different amounts of smoothing at different locations of a heavy-tailed PDF. Then retransformed PDF estimates with fixed smoothing parameters work like location-adaptive estimates. Therefore, such estimates may better evaluate the tail of heavy-tailed PDFs.

The selection of $T(x)$ is an important problem. By (3.24), a transformation $T(x)$ is completely determined by the distribution functions $G(x)$ and $F(x)$. One can select any ‘target’ $G(x)$, but $T(x)$ and $F(x)$ are unknown.

In Devroye and Györfi (1985) transformations to a finite interval, $T: R_+ \rightarrow [0, 1]$, were proposed. It was proved that both the transformation to an isosceles triangular PDF $\phi^{tri}(x)$ on $[0, 1]$, 

$$T(x) = \begin{cases} \sqrt{\frac{F(x)}{2}}, & F(x) \leq 0.5, \\ 1 - \sqrt{\frac{1-F(x)}{2}}, & F(x) > 0.5, \end{cases}$$

for kernel estimates with compact kernels, and the transformation $T(x) = F(x)$ to a uniform PDF $\phi^{uni}(x)$ for a histogram, provide the minimal convergence rate in metric space $L_1$:

$$\int_{-\infty}^{\infty} |f_n(x) - f(x)| \, dx = \int_{0}^{1} |g_n(x) - g(x)| \, dx.$$  

Since such $T(x)$ and, therefore, the distribution of $Y_j = T(X_j)$ depend on the unknown DF $F(x)$, it is impossible to obtain coinciding values of $g_0(x)$ and $\phi^{tri}(x)$ (or $\phi^{uni}(x)$). Hence, Devroye and Györfi (1985) proposed using some parametric family of DFs $\Psi_\gamma$ instead of $F$, which depends on some parameter $\gamma$ and to adapt $\gamma$ to the sample. However, the concrete models were not indicated and their influence on the decay rate at infinity of the retransformed estimates was not discussed.

Wand et al. (1991) and Yang and Marron (1999) consider the families of fixed transformations $T_\lambda(x)$ (independent of $F(x)$) given by

$$T_\lambda(x) = \begin{cases} x^\lambda \text{sign}(\lambda), & \text{if } \lambda \neq 0, \\ \ln x, & \text{if } \lambda = 0. \end{cases}$$

Here, $\lambda$ is the parameter minimizing the functional $\int_R (g''(y))^2 \, dy$, and $g(x)$ is the unknown PDF of the transformed r.v. $Y_1 = T_\lambda(X_1)$ that requires a preliminary estimation. Since the function $\int_R (g''(y))^2 \, dy$ shows the curvature of the PDF, such transformations are applied for better estimation of curvy but not necessarily heavy-tailed densities.

---

4 An empirical DF cannot be used as an estimate of $F(x)$ since its derivative does not exist at a finite number of points. Hence, formula (3.26) cannot be applied.
Markovitch and Krieger (2000) consider the fixed transformation \( T(x) = \frac{2}{\pi} \arctan(x) \), which provides good accuracy for some heavy-tailed PDFs. However, without assumptions on the type of the distribution any transformation may lead to a PDF that is difficult to estimate from a sample of moderate size and, hence, one cannot provide an accurate estimation of the ‘tails’.

To improve the estimation in the ‘tails’ a transformation \( T^*(x) : R_+ \rightarrow [0, 1] \) which is adapted to the data (via an estimate \( \hat{\gamma} \) of shape parameter \( \gamma \)) is proposed in Maiboroda and Markovich (2004). To construct such an estimate \( T^*(x) \) of \( T(x) \) one has to select a target DF \( G(x) \) and a fitted DF \( F(x) \). This adaptive transformation is considered in Section 4.3.

### 3.6 Exercises

1. Combined estimator.

   Generate \( X^n \) with sample size \( n = 500 \) according to some heavy-tailed distribution, for example, Fréchet, Weibull with shape parameter less than 1, or Burr. Calculate the combined estimate \( \tilde{f}(t, \gamma, N) \) by formula (3.1). For this purpose, use kernel estimator (2.18) with \( d = 1 \) as \( f^N(t) \), and the parametric model (3.3) as \( f_\gamma(t) \).

   Estimate the bandwidth \( h \) of the kernel estimator by the \( \omega^2 \) method, i.e. as a solution of the discrepancy equation (2.41). Use \( (nh)^{-1} \sum_{i=1}^{n} \int_{-\infty}^{\gamma_i} K \left( (x - X_i)/h \right) dx \) as \( \hat{F}(x) \).

   Estimate the shape parameter \( \gamma \) by Hill’s method (1.5). Select the number of largest order statistics \( k \) by means of a Hill plot (see Section 1.2.2). Use this value of \( k \) to determine the ‘boundary’ order statistic \( X_{(n-k)} \).

   Propose a boundary kernel of the \( \tilde{f}(t, \gamma, N) \) on the interval \( [X_{(n-k-1)}, X_{(n-k)}] \) to avoid the gap between \( f^N(t) \) and \( f_\gamma(t) \).

2. Repeat Exercise 1 but use the group estimator \( \gamma_i \) to estimate \( \gamma \). For this purpose, apply (1.19) and (1.28). Estimate the parameter \( m \) (i.e., the number of observations in each group) by means of a plot (see (1.29), Section 1.2.4).

   Compare the estimates obtained in both exercises.

3. Barron’s estimator (see Section 3.3).

   Generate \( X^n \) with sample size \( n \in \{50, 100, 500, 1000\} \) according to some heavy-tailed distribution. Consider the following parametric tail models.

   (a) Lognormal family with PDF

   \[
   g(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp \left( -\frac{(\ln x - \mu)^2}{2\sigma^2} \right),
   \]

   and DF

   \[
   G(x) = \frac{1}{\sigma_x} \Phi \left( \frac{\ln x - \mu}{\sigma} \right), \quad \text{for} \quad \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt.
   \]
120 HEAVY-TAILED DENSITY ESTIMATION

Use two variants of the parameter estimation:

(i) The maximum likelihood estimates are

\[ \mu_n = \frac{1}{n} \sum_{i=1}^{n} \ln X_i \quad \text{and} \quad \sigma_n^2 = \frac{1}{n} \sum_{i=1}^{n} \left( \ln X_i - \mu_n \right)^2. \]

(ii) The moment estimates are

\[ \mu_n = \frac{1}{2} \ln \frac{m_n^4}{m_n^2 + s_n^2} \quad \text{and} \quad \sigma_n^2 = \ln \left( 1 + \frac{s_n^2}{m_n^2} \right), \]

where \( m_n \) and \( s_n \) are the maximum likelihood estimates of the previous variant.

(b) Normal family. The maximum likelihood and moment estimates coincide, that is,

\[ \mu_n = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{and} \quad \sigma_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_n)^2. \]

(c) Weibull family with PDF

\[ g(x) = \beta \theta (x \theta)^{\beta-1} \exp \left( -(x \theta)^\beta \right), \quad \beta > 0, \theta > 0, \]

and DF

\[ G(x) = 1 - \exp \left( -(x \theta)^\theta \right). \]

Estimate parameters \( \beta \) and \( \theta \) by the maximum likelihood method.

Select the number of intervals \( m_n \) in Barron’s estimator (3.14), e.g., \( m_n \in \{5, 10, 20\} \). Construct the intervals \( A_{n1}, \ldots, A_{nm_n} \) by means of the quantiles \( G^{-1}(j/m_n), 1 \leq j \leq m_n - 1 \), of the DF \( G(x) \). One can select any other auxiliary DF \( G(x) \) to obtain these intervals. Calculate \( \delta_j \), the number of observations falling in each interval \( A_{nj} \). Calculate Barron’s estimate by formula (3.14).

Compare the estimation of a PDF using estimates with different tail models for different sample sizes. Draw conclusions regarding the reliability of the estimates for moderate size samples. Investigate how the accuracy of Barron’s estimate depends on the value \( m_n \) for different sample sizes.

4. Variable bandwidth kernel estimator.

Generate \( X^n \) according to some heavy-tailed distribution. Calculate the estimate by formula (3.15). For this purpose, calculate the auxiliary estimate \( \hat{f}_{h_1}(x) \) by formula (2.18). Select Epanechnikov’s kernel as \( K(x) \) for both estimators \( \hat{f}_{h_1}(x) \) and \( \tilde{f}^A(x|h_1, h) \). Calculate \( h_1 \) by formula (2.30). Select the parameter \( h \) in (3.15) in the following ways:
(a) Calculate $h$ by formulas (3.20) and (3.21). To do this, first estimate the PDF $f(x)$ and the derivative $(d/dx)^4(1/f(x))$ by a kernel method\(^5\).

(b) Use the weighted version of the squared error cross-validation method to estimate $h$ (see formulas (3.22) and (3.23)). The choice of $\omega$ could be

$$\omega(x) = \begin{cases} 1, & \text{for } \|\hat{\Sigma}^{-1/2}(x - \hat{\mu})\|^2 \leq z_\varepsilon, \\ 0, & \text{otherwise}, \end{cases}$$

where $\hat{\mu}$ and $\hat{\Sigma}$ denote the sample mean and variance, respectively, $\| \cdot \|$ is the Euclidean distance and $z_\varepsilon$ is the upper $(1 - \varepsilon)$-level critical point of the $\chi^2$ distribution with $p$ degrees of freedom, $\varepsilon \in \{0.1, 0.2\}$.

(c) Calculate $h$ by the $\omega^2$ method, i.e. as a solution of the discrepancy equation (2.41). Use $(nh)^{-1} \sum_{i=1}^n \hat{f}_{h1}(X_i)^{1/2} \int_{-\infty}^{x} K \left( (x - X_i) \hat{f}_{h1}(X_i)^{1/2}/h \right) dx$ as $\hat{F}(x)$.

Compare the estimates by different selection methods for $h$.

\(^5\) To estimate the $r$th derivative of the PDF one can use the estimator

$$\hat{f}^{(r)} = n^{-1}h^{-r-1} \sum_{i=1}^n K^{(r)} ((x - X_i)/h),$$

assuming that the kernel function $K(x)$ is smooth enough (Wand and Jones, 1995). For an extended discussion on the accuracy of the estimation of the PDF derivative, see Prakasa Rao (1983).
4

Transformations and heavy-tailed density estimation

In this chapter, we study the heavy-tailed PDF estimates based on the preliminary data transformations. Fixed and adaptive transformations are considered. To improve the behavior of a retransformed kernel estimate at infinity, boundary kernels are studied. To select smoothing parameters of the nonparametric PDF estimators the data-dependent discrepancy methods are investigated. These methods are applied both to nonvariable and variable bandwidth kernel estimators as well as to a projection estimator. The mean squared errors of these estimates are proved to be optimal.

4.1 Problems of data transformations

It is well known that kernel estimates provide a good asymptotic MISE and MSE for sufficiently smooth PDFs. For instance, the variable bandwidth kernel estimates give $\text{MISE} \sim n^{-8/9}$ and $\text{MSE} \sim n^{-8/9}$ even without a preliminary transformation of the data if the bandwidth $h$ is taken proportional to $n^{-1/9}$. However, this does not imply that the estimation in the tail domain will be good enough. Relatively large values of the PDF in the body generate the main contribution in the MSE (MISE), unlike small values in the tail. Hence, the MSE and MISE as well as popular
measures of the metric spaces $C$, $L_1$ and $L_2$ are not sensitive to the accuracy of the estimation at the tail.

For samples of moderate size a preliminary data transformation may improve the estimation of heavy-tailed PDFs at infinity (Section 4.3) or curvy PDFs (Section 3.5). The MISE of the retransformed PDF estimates is determined by the MSE of the PDF estimate of the new r.v. which is constructed by the transformation (Section 4.6).

Fixed transformations that do not require any knowledge of the distribution type, for example $\ln x$ or $(2/\pi) \arctan x$, are more attractive for practical applications (see Section 4.2). Nevertheless, they may result in discontinuous PDFs of transformed r.v.s, which are difficult to estimate. In general, the tail of the PDF cannot be estimated accurately by pure nonparametric methods since without imposing assumptions on the tail behavior the shape of the PDF of the transformed r.v. cannot be predicted. Considering kernel estimates, the rate of decay of retransformed estimates in the distribution tail may be close to that of the true PDF for certain boundary kernels and appropriate bandwidths (Section 4.5).

To improve the accuracy of retransformed estimates, the selection of the smoothing parameter (e.g., the kernel estimate bandwidth or the polygram bin width) constitutes the most important problem. It is better to estimate this parameter for the PDF of the transformed r.v. if the latter is compactly supported and sufficiently smooth.

Data-driven selection methods like the cross-validation and the $D$ and $\omega^2$ discrepancy methods are universal in the sense that they are applicable to any nonparametric PDF estimator. It is proved that the $D$ method may provide the optimal rates of the MSE of kernel estimates with variable and nonvariable bandwidths (Sections 4.7 and 4.8), while the $\omega^2$ method provides an optimal convergence rate of some projection estimate in the metric of the space $L_2$ (Section 4.9).

The proofs of all stated theorems are presented in Appendix B.

### 4.2 Estimates based on a fixed transformation

Here, the fixed transformation $\arctan x$ is considered. The estimates obtained by means of this transformation were investigated in a Monte Carlo study (Markovitch and Krieger, 2000). The transformation

$$ T(x) = \frac{2}{\pi} \arctan x, \quad T'(x) = \frac{2}{\pi(1+x^2)} \quad (4.1) $$

does not depend on the sample $X^n$ and satisfies the conditions on transformations assumed in Section 3.5. $T(x)$ generates an r.v. $Y = T(X)$ with a bounded PDF

---

1 $g(x)$ is calculated by formula (3.25).
Figure 4.1 PDFs of transformed r.v.s $Y$ generated by transformation (4.1) for different $f(x)$: standard exponential (solid line), Cauchy (solid horizontal line), Weibull with shape parameter 0.5 (dotted line), gamma with shape parameter 2 (dot-dashed line), lognormal with parameters (1, 1) (solid line with + marks). Based on Figure 1 in Markovich and Krieger (2000).

$g(x)$ for many heavy-tailed PDFs $f(x)$ (apart from the Weibull distribution; see Figure 4.1).

The estimate $g_n(x)$ may not obey the conditions of a PDF on $[0, 1]$ since part of the distribution may be located outside $[0, 1]$. However, one may normalize it, that is, use the estimate

$$\hat{g}_n(x) = \frac{g_n(x)}{\int_0^1 g_n(x)dx}$$

instead of $g_n(x)$. The risk in the metric space $L_1$ will decrease after such normalization. This implies, for the estimate

$$\hat{f}_n(x) = \hat{g}_n(T(x))T'(x), \quad (4.2)$$

that

$$\int_0^\infty |\hat{f}_n(x) - f(x)|dx = \int_0^1 |\hat{g}_n(x) - g(x)|dx \leq \int_0^1 |g_n(x) - g(x)|dx$$

(Devroye and Györfi, 1985).

The following algorithm to estimate a heavy-tailed PDF is considered:

- Construct the nonparametric estimate $g_n$, located on $[0,1]$, from the transformed sample $Y^n = \{Y_1, \ldots, Y_n\}$, $Y_i = T(X_i)$, $i = 1, \ldots, n$, and normalize if necessary.
• Calculate an estimate of the smoothing parameter of \( g_n \).

• To obtain the estimate of the PDF \( f(x) \), apply an inverse transformation (4.2).

For the purpose of the analysis the polygram (2.29) and kernel estimators with the Gaussian kernel and Epanechnikov’s kernel (2.21) are used. For transformation (4.1) these kernel estimates of the transformed r.v. \( Y_i \) are determined by

\[
\hat{g}_{h,n}^{(1)}(x) = \frac{1}{nh\sqrt{2\pi}} \sum_{i=1}^{n} \exp \left( -\frac{1}{2} \left( \frac{x - Y_i}{h} \right)^2 \right) 
\]

(4.3)

and

\[
\hat{g}_{h,n}^{(2)}(x) = \frac{3}{4nh} \sum_{i=1}^{n} \left( 1 - \left( \frac{x - Y_i}{h} \right)^2 \right) \theta(h + Y_i - x),
\]

(4.4)

respectively, where

\[
Y_i = \frac{2}{\pi} \arctan(X_i), \quad \theta(t) = \begin{cases} 
1, & t \geq 0, \\
0, & t < 0.
\end{cases}
\]

By (4.2) we obtain, after normalization, the final PDF estimates constructed from the transformed sample \( Y^n \):

\[
\hat{f}_{h,n}^{(1)}(x) = \frac{\sqrt{2}}{nh\pi^2 I_{[0,1]}^1(h)(1 + x^2)} \sum_{i=1}^{n} \exp \left( -\frac{1}{2} \left( \frac{2}{\pi} \arctan(x) - Y_i \right)^2 \right) 
\]

(4.5)

and

\[
\hat{f}_{h,n}^{(2)}(x) = \frac{3}{2\pi nh I_{[0,1]}^2(h)(1 + x^2)} \sum_{i=1}^{n} \left( 1 - \left( \frac{2}{\pi} \arctan(x) - Y_i \right)^2 \right) \theta \left( h + Y_i - \frac{2}{\pi} \arctan(x) \right). 
\]

(4.6)

Here

\[
I_{[0,1]}^1(h) = \frac{1}{n} \sum_{i=1}^{n} \left( \Phi \left( \frac{1 - Y_i}{h} \right) - \Phi \left( \frac{-Y_i}{h} \right) \right)
\]

is the integral of \( \hat{g}_{h,n}^{(1)}(x) \) on \([0, 1]\), \( \Phi(x) = 1/(\sqrt{2\pi}) \int_{-\infty}^{x} \exp \left( -\frac{u^2}{2} \right) du \) is the Gaussian DF and

\[
I_{[0,1]}^2(h) = \frac{3}{4nh} \sum_{i=1}^{n} \left( 1 - \frac{1}{3\pi^2} (1 + 3Y_i(Y_i - 1)) \right), \quad \text{for } h + Y_i > 1,
\]

\[
\frac{2}{3} h + Y_i \left( 1 - \frac{Y_i^2}{3\pi^2} \right), \quad \text{for } h + Y_i \leq 1,
\]

is the integral of \( \hat{g}_{h,n}^{(2)}(x) \) on \([0, 1]\).
Let $g_{L,n}(x)$ be a polygram constructed on $Y^n$ by formula (2.29). After the inverse transformation (4.2) we get (since no normalization is necessary)

$$f_{L,n}(x) = \frac{2}{\pi(1+x^2)} g_{L,n} \left( \frac{2}{\pi} \arctan x \right).$$

(4.7)

Let us now discuss the selection of the bandwidth $h$ determining the accuracy of the kernel estimates and a polygram. The parameters $h$ and $L$ of estimates $g_{h,n}^{(1)}(x)$, $g_{h,n}^{(2)}(x)$ and $g_{L,n}(x)$ may be calculated from a sample $Y^n$ using the cross-validation method (2.31) (or (2.35) for the Gaussian kernel) or the $\omega^2$ and $D$ discrepancy methods (see (2.38)–(2.40)). The idea of the $\omega^2$ method is to obtain $h$ (or $L$) from

$$\hat{\omega}^2(h) = \sum_{i=1}^{n} \left( G^h(Y_{(i)}) - \frac{i-0.5}{n} \right)^2 + \frac{1}{12n} = 0.05,$$

or, in the case of the $D$ method, from

$$\sqrt{n} D_n = \sqrt{n} \max(\hat{D}^+_n, \hat{D}^-_n) = 0.5,$$

where

$$\sqrt{n} \hat{D}^+_n = \sqrt{n} \max_{1 \leq i \leq n} \left( \frac{i}{n} - G^h(Y_{(i)}) \right), \quad \sqrt{n} \hat{D}^-_n = \sqrt{n} \max_{1 \leq i \leq n} \left( G^h(Y_{(i)}) - \frac{i-1}{n} \right),$$

and $Y_{(1)} \leq Y_{(2)} \leq \ldots Y_{(n)}$ are the order statistics of the transformed observations. For the normalized kernel estimate (4.3) we get

$$G_1^h(x) = \frac{1}{I_{[0,1]}^1(h)} \int_0^x g_{h,n}^{(1)}(t) dt = \frac{1}{n I_{[0,1]}^1(h)} \sum_{i=1}^{n} \left( \Phi \left( x - \frac{Y_i}{h} \right) - \Phi \left( -\frac{Y_i}{h} \right) \right),$$

while for the normalized kernel estimate (4.4) we get

$$G_2^h(x) = \frac{1}{I_{[0,1]}^2(h)} \int_0^x g_{h,n}^{(2)}(t) dt = \frac{3}{4 n h I_{[0,1]}^2(h)} \sum_{i=1}^{n} \begin{cases} \frac{x - \frac{1}{3h^2} (x - Y_i)^3 + Y_i^3}{x - \frac{1}{3h^2} (h^3 + Y_i^3)}, & h + Y_i \geq x, \\ \frac{x - \frac{1}{3h^2} (h^3 + Y_i^3)}{x - \frac{1}{3h^2} (h^3 + Y_i^3)}, & h + Y_i < x. \end{cases}$$

For the polygram one can calculate

$$G^L(x) = \int_0^x g_{L,n} (t) dt$$

instead of $G^h(Y_{(i)})$ in the formulas above. In Markovitch and Krieger (2000) the polygram and kernel estimates (4.5) and (4.6) with noncompact and compact kernel functions were compared for long-tailed distributions by a simulation study. Moreover, the $\omega^2$ and $D$ methods were compared with the cross-validation method on p. 77.

Regarding the comparison, samples of a gamma distribution with parameter $s = 2$, a lognormal distribution with $\mu = 1, \sigma = 1$ and a Weibull distribution
with \( s = 0.5 \) were generated. The gamma distribution is related to light-tailed distributions, but the lognormal and Weibull PDF are heavy-tailed. As characteristics of the estimates the loss functions in metric spaces \( L_1, L_2 \) and \( C \) were used.

The simulation study showed that the heavy-tailed Weibull PDF is difficult to estimate using such retransformed kernel estimates. For this PDF there is no uniform convergence for any of the estimates and smoothing methods considered. At the same time, the polygram demonstrates better accuracy than kernel estimates in \( L_1 \) and \( L_2 \). For the gamma and lognormal PDFs the polygram and the kernel estimate with the Gaussian kernel are preferable. They provide convergence in all metrics considered, whereas the kernel estimate with Epanechnikov’s kernel does not converge in \( C \) for the lognormal PDF as the sample size increases.

It follows from the simulation study that a polygram and kernel estimate (4.5) are preferable for the application to real data if the true PDF is not available. If one knows that the PDF is heavy-tailed, then a polygram may be recommended.

### 4.3 Estimates based on an adaptive transformation

Some important questions about the transform–retransform scheme (see Section 3.5) may arise: firstly, what family of distributions is a reasonable approximation of the true DF; secondly, what is a better target DF of the transformation considering the stability of the retransformed estimates to minor perturbations in the transformation; and thirdly, which nonparametric estimate best maintains the rate of tail decay of the true PDF. Here we discuss all these questions.

The adaptive transformation described in the following is derived from a specific assumption regarding the parametric model \( \Psi \gamma \) of the DF \( F(x) \); see Maiboroda and Markovich (2004). The system of heavy-tailed distributions is taken as \( \Psi \gamma \), where the EVI \( \gamma \) is estimated using Hill’s estimator.

#### 4.3.1 Estimation algorithm

The algorithm proceeds as follows:

- Estimate the EVI of \( X_j \) from the sample \( X^n \) (see Section 1.2), for example, using Hill’s estimator \( \hat{\gamma}_n = \hat{\gamma}^H (n, k) \).
- Construct the transformation \( T = T_{\hat{\gamma}_n} \) in the following way: if \( \xi \) has the fitted DF \( \Psi_{\hat{\gamma}_n} \) then \( T_{\hat{\gamma}_n}(\xi) \) has the target DF \( \Phi \), for example, a uniform or triangular one. (Here \( \hat{\gamma}_n \) is considered as a fixed value.)
- Construct the transformed sample \( Y_j = T_{\hat{\gamma}_n}(X_j), j = 1, \ldots, n \).
- Estimate the PDF of \( Y_1, \ldots, Y_n \) by some estimator \( \hat{g}_n(x) \).
- Estimate the PDF of \( X_j \) by (3.26).
4.3.2 Analysis of the algorithm

Given a realization of the algorithm, we must choose a family of DFs, the target distribution, and the estimate \( \hat{g}_n \). The family of fitted DFs must be chosen in such a way that the transformation \( T \) and its derivative can be easily evaluated. If \( \Psi_\gamma \) is the fitted DF and \( \Phi \) is the target DF, then by (3.24), \( T(x) = \Phi^{-1}(\Psi_\gamma(x)) \) and \( T^{-1}(x) = \Phi^{-1}(\Psi_\gamma^{-1}(x)) \) hold.

We assume that the GPD is the fitted DF and \( x < 0 / \Phi \) is the EVI of this distribution, and the estimate \( \hat{g}_n \) is a bounded function with \( \lim_{x \to -\infty} \ell(x) = \ell_\infty < \infty \) (\( \gamma \) is the EVI of this PDF). Then by (3.25),

\[
g(x) = \left(1 + \frac{1 - x}{\gamma} - 1\right)^{-1/(\gamma+1)} (1 - x)^{-\gamma-1} \cdot \ell \left(\frac{(1 - x)^{-\gamma} - 1}{\gamma}\right), \quad x \in [0, 1],
\]

with \( \gamma > 0 \) is chosen as a fitted DF. Evidently, the fitted PDF is determined by

\[
\psi_\gamma(x) = \Psi_\gamma'(x) = (1 + \hat{\gamma}x)^{-1/(\hat{\gamma}+1)}.
\]

This choice of distribution is widespread and motivated by the theorem of Pickands (1975); see Section 1.1. This theorem states that, for a certain class of distributions \( F(x) \in \text{MDA} \left( H_\gamma \right), \gamma \in \mathbb{R} \), of the r.v. \( X \) and for a sufficiently high threshold \( u \) of the r.v. \( X \), the conditional distribution of the overshoot \( Y = X - u \), provided that \( X \) exceeds \( u \), converges to the GPD.

We consider the uniform PDF \( \phi_{\text{uni}}(x) = 1\{x \in [0, 1]\} \) and the positive triangular PDF \( \phi_{\text{tri}}(x) = 2(1 - x)1\{x \in [0, 1]\} \) as our target PDFs. The corresponding DFs are therefore

\[
\Phi_{\text{uni}}(x) = x1\{x \in [0, 1]\} + 1\{x > 1\}, \quad \Phi_{\text{tri}}(x) = (2x - x^2)1\{x \in [0, 1]\} + 1\{x > 1\}.
\]

We wish to clarify two questions:

- Which transformation ensures a more stable estimation algorithm given deviations of an EVI estimate?
- What tail behavior is ensured by the inverse transformation?

Let the target be \( \phi = \phi_{\text{uni}}, \Phi = \Phi_{\text{uni}}. \) Then

\[
T_\gamma(x) = \Phi^{-1}(\Psi_\gamma(x)) = 1 - (1 + \hat{\gamma}x)^{-1/\hat{\gamma}}, \quad T_\gamma'(x) = (1 + \hat{\gamma}x)^{-1/(\hat{\gamma}+1)},
\]

\[
T_\gamma^{-1}(x) = ((1 - x)^{-\hat{\gamma}} - 1)/\hat{\gamma}, \quad (T_\gamma^{-1}(x))' = (1 - x)^{-\hat{\gamma}-1}.
\]

We suppose that the true PDF of \( X \) is given by

\[
f(x) = \ell(x)\psi_\gamma(x),
\]

where \( \ell(x) \) is a bounded function with \( \lim_{x \to -\infty} \ell(x) = \ell_\infty < \infty \) (\( \gamma \) is the EVI of this PDF). Then by (3.25),

\[
g(x) = \left(1 + \gamma \frac{(1 - x)^{-\hat{\gamma}} - 1}{\hat{\gamma}}\right)^{-(1/(\gamma+1))} (1 - x)^{-\hat{\gamma}-1} \cdot \ell \left(\frac{(1 - x)^{-\gamma} - 1}{\gamma}\right), \quad x \in [0, 1],
\]
follows as the PDF of \( Y = T_\hat{\gamma}(X) \). This PDF will be estimated by \( \hat{g}_n(x) \) for some value of \( \hat{\gamma} \). Hence, its behavior on its support \( x \in [0, 1] \) must be convenient for the estimation. If \( x \) is close to zero, then \( g(x) \) is bounded. Let us consider \( g(x) \) as \( x \uparrow 1 \). Note that in this case \( \ell(T_\hat{\gamma}^{-1}(x)) \to \ell_\infty < \infty \) since \( T_\hat{\gamma}^{-1}(x) \to \infty \). Then for \( x \uparrow 1 \) it follows that

\[
g(x) \simeq \ell_{\infty} \left( 1 + \gamma \frac{(1-x)^{-\hat{\gamma}} - 1}{\hat{\gamma}} \right)^{-1/(1+1)} \left( 1 - x \right)^{-\hat{\gamma} - 1}.
\]

For \( \gamma = \hat{\gamma} \) we have \( g(x) \simeq \ell_{\infty} \), that is, \( g(x) \) behaves as a uniform PDF in the neighborhood of \( x = 1 \) (up to an insignificant multiplier \( \ell_{\infty} \)).

Let \( \gamma \neq \hat{\gamma} \). For \( x \uparrow 1 \) and any constant \( C < \infty \), we have \( (1-x)^{-\hat{\gamma}} + C \sim (1-x)^{-\hat{\gamma}} \). Hence,

\[
g(x) \simeq \ell_{\infty} \left( \gamma / \hat{\gamma} \right)^{-1/(1+1)} (1 - x)^{\hat{\gamma}/\gamma - 1}.
\]

Therefore, if \( \hat{\gamma} \) overestimates the true EVI \( \gamma \), that is, \( \hat{\gamma}/\gamma - 1 > 0 \), then the PDF \( g(x) \) behaves nicely in the neighborhood of \( x = 1 \) (provided it is not like the uniform distribution). But if we underestimate \( \gamma (\hat{\gamma} < \gamma) \) then \( g(x) \to \infty \) as \( x \uparrow 1 \). In such situations PDF estimates perform very poorly, since they are designed for the estimation of finite values.

We now suppose that the target PDF is a triangular one. Then

\[
\Phi^{-1}(x) = (\Phi^{tri})^{-1}(x) = 1 - \sqrt{1 - x},
\]

\[
T_\hat{\gamma}(x) = 1 - (1 + \hat{\gamma}x)^{-1/(2\hat{\gamma})}, \quad T_\hat{\gamma}^{-1}(x) = ((1-x)^{-2\hat{\gamma}} - 1)/\hat{\gamma},
\]

\[
T_\hat{\gamma}'(x) = 0.5(1 + \hat{\gamma}x)^{-1/(2\hat{\gamma} - 1)}, \quad (T_\hat{\gamma}^{-1}(x))' = 2(1-x)^{-2\hat{\gamma} - 1}.
\]

If the PDF of \( X \) is of the form (4.10), then

\[
g(x) \simeq \ell_{\infty} \left( \gamma / \hat{\gamma} \right)^{-1/(1+1)} (1 - x)^{2\hat{\gamma}/\gamma - 1}, \quad \text{as } x \uparrow 1,
\]

arises as the PDF of \( Y = T_\hat{\gamma}(X) \). For \( \hat{\gamma} = \gamma \) the PDF \( g(x) \to 0 \) as \( x \uparrow 1 \) and its rate of decay is the same as for the triangular PDF, \( \sim C(1-x) \). If \( \hat{\gamma} \neq \gamma \) and \( \hat{\gamma} > \gamma/2 \) then \( g(x) \) tends to zero as \( x \uparrow 1 \). The rate of this decay is not asymptotically equal to \( C(1-x) \), but PDF estimates will normally work in this case. These problems can only arise if \( \hat{\gamma} < \gamma/2 \). If \( \gamma \) is estimated by a consistent estimator (e.g., by Hill’s) such rough underestimation will be a rare event. In fact, the probability of getting a value \( \hat{\gamma} \) which is far from \( \gamma \) tends to zero as the sample size \( n \uparrow \infty \).

We suppose now that we have estimated \( g(x) \) by \( \hat{g}_n(x) \). What is the value of the EVI if the PDF estimate \( \hat{f}_n(x) \) is defined by (3.26)? Let \( \hat{g}_n(x) \) be a histogram (or a polygram) estimate. Since for both transformations \( T_\gamma(x) \to 1 \) as \( x \to \infty \), we have

\[
\hat{f}_n(x) \simeq \hat{g}_n(1) T_\gamma'(x) \quad \text{as } x \to \infty \text{ by (3.26)}.
\]

Hence, the accuracy of the estimation in the tail depends on the behavior of the estimate \( \hat{g}_n(x) \) in the neighborhood of \( x = 1 \).

We suppose that \( \hat{g}_n(1) > 0 \). This property holds almost surely for large \( n \) if \( R_+ \) is the support of \( X \). Note that \( \hat{g}_n(1) \) is the height of the rightmost bar of the histogram.
If the target PDF is uniform, this means that \( \hat{f}_n(x) \approx \hat{g}_n(1)(1 + \hat{\gamma}x)^{-\frac{1}{\hat{\gamma}+1}} \), that is, the estimate has the EVI which we have chosen for the fitted PDF. In our algorithm it is given by \( \hat{\gamma} \).

The same is true for a kernel estimate when \( \hat{g}_n(1) \) is replaced by \( \hat{g}_n(1, h) \). The kernel estimate with a compactly supported kernel (e.g., Epanechnikov’s kernel (2.21)) \( \hat{g}_n(x) \) depends on the smoothing parameter \( h \) and \( \hat{g}_n(1, h) \sim O(h^{-1}) \) for sufficiently large \( h > 1 - \frac{\hat{\gamma}}{X_n} \), whereas \( \hat{g}_n(1, h) = 0 \) holds for other \( h \). The situation can be explained by Figure 4.2. Here, we observe boundary effects of the kernel estimate applied to a PDF with a compact support due to the truncation of the kernel near the boundaries.

If the target PDF is triangular, we have

\[
\hat{f}_n(x) \approx 0.5\hat{g}_n(1)(1 + \hat{\gamma}x)^{-\frac{1}{(\hat{\gamma}+1)}}.
\]

In this case, the EVI is twice as large as needed. To remove this effect, we can use a smoothed histogram (or polygram). For such \( \hat{g}_n(x) \) we have \( \hat{g}_n(x) = C_n(1 - x) \) for \( x \approx 1 \), that is \( \hat{g}_n(x) \approx \phi^{\text{tri}}(x) \). Here \( C_n \) is the slope of the line which connects the center of the top of the rightmost histogram bar with the point \((1,0)\). For the triangular target PDF we then get

\[
\hat{f}_n(x) \approx 0.5(1 + \hat{\gamma}x)^{-\frac{1}{(\hat{\gamma}+1)}}.
\]
\[
\hat{f}_n(x) \simeq 0.5C_n(1 - (1 - (1 + \hat{\gamma}x)^{-1/(2\hat{\gamma})}))(1 + \hat{\gamma}x)^{-(1/2\hat{\gamma}+1)}
\]
\[
\simeq \frac{C_n}{2}(1 + \hat{\gamma}x)^{-(1/\hat{\gamma}+1)},
\]
so that the EVI of the estimate coincides with its estimate of the fitted PDF.

For a kernel estimate this rate in the tail may be given by a correct selection of a kernel near the boundary and a smoothing parameter \( h \). One can take the triangular kernel \( K(x) = (1 - |x|)1_{\{|x| \leq 1\}} \) as a boundary kernel. Then, by (2.18),
\[
\hat{g}_n(x) \simeq (1 - (x - T_{\hat{\gamma}}(X_{(n)}))/h)/h
\]
holds for boundary points \( x \in (T_{\hat{\gamma}}(X_{(n)}), 1] \). For \( h = 1 - T_{\hat{\gamma}}(X_{(n)}) \) it follows that \( \hat{g}_n(x) \simeq (1 - x)/h^2 \). Therefore, we get the same tail of \( \hat{f}_n(x) \) as for a smoothed polygram.

The estimation of the body and the tail of the PDF of the Fréchet(0.3) distribution (1.32) by the retransformed kernel estimate with Epanechnikov’s kernel and the retransformed polygram is shown in Figure 4.3. The sample size is \( n = 50 \). Here, \( h = n^{-1/5} = 0.457 > h_1 = 1.01 - T_{\hat{\gamma}}(X_{(n)}) = 0.112 \). The maximal observation \( X_{(n)} \) in the sample is equal to 9.235. Hill’s estimate \( \hat{\gamma}(n, k) \) of the EVI is equal to 0.278, where \( k = 11 \) is obtained by the bootstrap method (Caers and Van Dyck, 1999) with resample size \( B = 50 \). The parameter \( L \) of the polygram is equal to 15. The tail domain is shown on a logarithmic scale on both axes.

The value \( h_1 \) provides a better estimation of the tail of the PDF than \( h \) due to the better estimation of the PDF \( g(x) \) of the transformed r.v. at the boundary. Experience tells us that the triangular and Epanechnikov’s kernels provide similar results for the same \( h \). If a value of \( h \) smaller than \( h_1 \) is used, it leads to truncation of the kernel (see Figure 4.2). Hence, the estimate is equal to 0 in the tail.

![Figure 4.3](image-url) ‘Body’ (left) and ‘tail’ (right) estimation of a Fréchet PDF (solid line) by retransformed estimates: kernel estimate with \( h \) (dotted line), kernel estimate with \( h_1 \) (solid line with + marks), polygram (dashed line). The polygram nearly coincides with the PDF in the tail domain. The kernel estimate with \( h_1 \) is best in the ‘body’.
Generally speaking, a boundary kernel should coincide with a target PDF, which is nearly the same as \( g(x) \) for \( \hat{y} \approx y \). Let us explain why the polygram is preferable to the histogram in this case. As a matter of fact, a PDF \( g \) with \( g(x) \to 0 \) as \( x \to 1 \) is estimated. If the lengths of the bins are equal, very few observations fall into the rightmost bin. Hence, the histogram estimate is not stable in the tail. The polygram dynamically adapts the bin widths to the data and works better.

Of course, the algorithm considered can perform in inadequately if the underlying PDF has light tails. Therefore, it is a good idea to test this hypothesis before the PDF is estimated. A comprehensive list of references on such tests can be found in Jurečková and Picek (2001) and Dietrich et al. (2002). It is useful to provide a preliminary data analysis as is done in Section 1.3.

In summary, we conclude that among the alternatives considered the best combination of parameters of the algorithms is determined by \( \phi^{\text{tri}} \) as target PDF and a smoothed polygram (or a kernel estimate with compactly supported kernel) as PDF estimate on \([0,1]\).

The next question concerns the outcome if one applies the transformation (4.11) but the true PDF does not belong to a Pareto class. Typical distributions with heavy tails are distributions with regularly varying tails (like (4.10)), lognormal-type tails and Weibull-like tails (Mikosch, 1999). Applying the transformation (4.11) to the lognormal and Weibull PDFs, one can conclude by (3.25) that the corresponding PDF \( g(x) \) of the transformed r.v. is continuous in the neighborhood of \( x = 1 \). However, without the assumption on the class of the tail, an accurate estimation of the tail by means of a nonparametric method is impossible, because in this case one cannot select a suitable boundary kernel.

### 4.3.3 Further remarks

Our first aim here has been to present a nonparametric method which provides a good estimation of the ‘tails’ of heavy-tailed PDFs. The transformation to a triangular PDF and a smoothed polygram (or a kernel estimate) is established as the best combination of an estimation. It allows us to get a stable estimate of a PDF with regard to small perturbations of the transformation due to a rough EVI estimation, and to retain the tail decay of the true PDF after the inverse transformation.

Retransformed nonparametric estimates have the following advantages:

1. Estimates with a fixed smoothing parameter \( h \) work like estimates with a variable \( h \).

2. The data transformation allows us to apply nonparametric (histogram, polygram, projection) estimates which are only suitable for PDFs with finite support to improve the accuracy of the estimates for heavy-tailed PDFs.
3. The transformation approach is the most suitable one for heavy-tailed PDFs or PDFs with sharp features, such as high skewness (Wand et al., 1991; Wand and Jones, 1995; Yang and Marron, 1999).

It makes no sense to apply the transformation approach to ‘simple’ PDFs. Such PDFs may be better estimated by standard nonparametric methods. The accuracy of a retransformed estimate is defined by the accuracy of the PDF estimator of a transformed r.v.; in particular, the mean integrated absolute error (MIAE) is invariant to the transformation (see property (2.2)), so $MIAE_x = MIAE_{T(x)}$ (Devroye and Györfi, 1985). The latter may be worse than the accuracy of an untransformed PDF estimate. For example, suppose that a Gaussian r.v. is transformed into a triangular distributed r.v. It was shown in Markovich (1989) that a kernel estimator with the Gaussian kernel estimates the normal PDF better than a triangular one. It is clear that one can also give examples of the preference of the transformed approach.

An alternative approach to estimating heavy-tailed PDFs is determined by variable bandwidth kernel methods (Abramson, 1982; Devroye and Györfi, 1985). However, the latter estimators are not intended for accurate tail estimation. In order to recognize the heaviness of the tail one may use one of the tests specified in Jurečková and Picek (2001).

Simplicity of calculations is very important in practice. In this respect the use of fixed transformations (like that in Section 4.2 or the classical $T(x) = \ln x$), leading to bounded, convenient PDFs of a transformed r.v. that require no assumptions on the distribution, may be sometimes preferable.

However, even if information about the behavior of the distribution at the tail is available, fixed transformations cannot provide PDFs of the transformed r.v.s with predictable features. The latter may lead to bad estimation of the PDF. In this respect, adaptive transformations are more flexible. Here, a transformation that is adapted to the data under the assumption that the true distribution belongs to a Pareto class is considered. In Section 4.3.2, PDF estimation in the class of distributions with regularly varying tails (4.10) is investigated. Due to the choice of $\ell(x)$ this class is rather wide and includes the Pareto distribution as well. We leave the consideration of heavy-tailed PDFs with other types of tails as an open problem.

The Pareto parameter (or an EVI) reflects the shape of the tail and is estimated by sparse data. It may also be estimated by Hill’s estimator. The fitting capabilities of a transformation family strongly depend on the accuracy of Hill’s estimator. It is known that the latter is accurate only for sufficiently large sample sizes. Therefore, more accurate EVI estimates could be preferable.

To provide the correct tail decay at infinity the smoothed polygram is used. The selection of a boundary kernel and a smoothing parameter for it is proposed for a kernel estimate. A smoothed polygram may be preferable, especially for limited sample sizes and for tail estimation. For sufficiently large samples, kernels give a better estimate for the ‘body’ of a PDF. The quality
of the kernel estimates may be improved by further boundary corrections to improve the tail estimation, a better EVI estimation, and a better smoothing procedure.

### 4.4 Estimating the accuracy of retransformed estimates

We consider the MISE on the interval $\Omega$,  

\[
\text{MISE}^h(\hat{\gamma}, \Omega) = \mathbb{E} \int_{\Omega} (\hat{f}(x) - f(x))^2 \, dx
\]

(4.12)

\[
= \mathbb{E} \int_{\Omega} (\hat{g}_h(T_\gamma(x)) - g(T_\gamma(x)))^2 T_\gamma'(x) \, dT_\gamma(x)
\]

\[
= \mathbb{E} \int_{\Omega^*} (\hat{g}_h(y) - g(y))^2 T_\gamma'(T_\gamma^{-1}(y)) \, dy,
\]

as a measure of the quality of the estimate $\hat{f}(x)$. Here $\Omega^* = T_\gamma(\Omega)$ and $g(x)$ is the PDF,

\[
g(x) = f(T_\gamma^{-1}(x))(T_\gamma^{-1}(x))',
\]

(4.13)

which is actually estimated instead of

\[
g_0(x) = f(T_\gamma^{-1}(x))(T_\gamma^{-1}(x))' \tag{since $\hat{\gamma} \neq \gamma$,}
\]

and $\hat{g}_h(x)$ is some estimate of $g(x)$ with the smoothing parameter $h$.

For fixed transformations and nonrandom intervals $\Omega^*$ the MISE has a simpler form than (4.12):

\[
\text{MISE}^h(\Omega) = \int_{\Omega^*} T'(T^{-1}(y))E(\hat{g}_h(y) - g(y))^2 \, dy. \tag{4.14}
\]

If $0 < T'(T^{-1}(x)) \leq c$ holds on $\Omega^*$ for the transformation $T$ (not necessarily fixed), then we have

\[
\text{MISE}^h(\Omega) \leq c \int_{\Omega^*} E(\hat{g}_h(y) - g(y))^2 \, dy \tag{4.15}
\]

for a nonrandom $\Omega^*$. This means that the order of the MISE of the retransformed estimates at $\Omega$ is at least not worse than the order of the MSE of $\hat{g}_h(y)$. For example, a kernel estimator or a polygram can be used as $\hat{g}_h(x)$.

**Example 9** Note that for the adaptive transformation (4.11), the function $T_\gamma'(T_\gamma^{-1}(x)) = 0.5 (1 - x)^{1+2\gamma}$ is bounded on $[0,1]$. 

4.5 Boundary kernels

If the tail shape is known, then, irrespective of the transformation, the order of decay of transformed estimates on tails may be close to that of the true PDF \( f(x) \) for a proper choice of kernels and smoothing parameter \( h \) near the bounds of variation of the transformed r.v. \( Y_1 = T(X_1) \), (Markovich, 2005c). Let us assume that the PDF of the r.v. \( X_1 \) belongs to class\(^2\)

\[
f(x) = \begin{cases} 
\ell(x)(1 + \gamma x)^{-(1/\gamma+1)}, & x \geq c > 0, \\
0, & x < c,
\end{cases}
\]

(4.16)

where \( \ell(x) \) is a slowly varying function (see Definition 11).

We consider transformation (4.11) first and obtain the PDF estimate for the r.v. \( X_1 \),

\[
\hat{f}(x) = \hat{g}_h(T\_\gamma(x))T\_\gamma'(x) = 0.5\hat{g}_h(T\_\gamma(x))(1 + \hat{\gamma}x)^{-1/(2\gamma)-1}.
\]

(4.17)

We need to find \( \hat{f}(x) \sim (1 + \hat{\gamma}x)^{-1/\gamma-1} \), which is close to (4.16) for a sufficiently accurate estimate \( \hat{\gamma} \). Taking (4.17) into account, the accuracy of the tail of the estimate \( \hat{f}(x) \) depends on the behavior of the estimate \( \hat{g}_h(x) \) near \( x = 1 \) since \( T\_\gamma(x) \to 1 \) as \( x \to \infty \).

It is known that the kernel estimates (2.18), when they are applicable to PDFs, concentrated on bounded intervals, exhibit boundary effects due to the truncation of kernels at the support boundary (this is discussed in detail in Section 4.3 and in Figure 4.2). This effect is suppressed with boundary kernels. They are useful in improving the estimation of PDFs that take zero values or suffer from discontinuity (a bounded jump) at the boundary. To overcome this effect, Scott (1992) used modified bi-weight kernels \( K_b(x) = \frac{15}{16} (1 - x^2)^2 \). Schuster (1985) studied the mirror imaging of the sample to negative values.

These methods can be successfully applied only if the location of the discontinuity or zero value of the PDF is known. In general, such information can be obtained from a histogram. A boundary kernel is constructed under certain constraints on the kernel. For example, the kernel and its derivative vanish at the boundary if the PDF vanishes. Since such a boundary kernel does not depend on the PDF, there is no guarantee that the estimation would be good. Following Simonoff (1996), the bias of the estimate at the boundary can be reduced by using a linear combination \( B_1 \) or \( B_2 \) of two kernels \( K(x) \) and \( L(x) \)(\( K(x) \) is the usual kernel used in PDF estimation at interior points of an interval and is defined on \([0,1]\), whereas \( L(x) \) is a kernel differing from \( K(x) \), but related to it in some way):

\[
B_1(x) = \frac{c_1(p)K(x) - a_1(p)L(x)}{c_1(p)a_0(p) - a_1(p)c_0(p)},
\]

(4.18)

\(^2\) The value \( x \) is bounded away from zero to avoid problems, particularly if \( \ell(x) = \ln x \).
where

\[ a_l(p) = \int_{1-p}^{1} u'K(u)du, \quad c_l(p) = \int_{1-p}^{1} u'L(u)du, \quad 0 < p < 1, \]

or

\[ B_2(x) = \frac{(a_2(p) - a_1(p)x)K(x)}{a_0(p)a_2(p) - a_1^2(p)}, \quad (4.19) \]

if \( L(x) = xK(x) \). This is only possible if the second derivative of the estimated PDF is continuous.

The bias of a kernel estimate with kernel \( B_1 \) or \( B_2 \) in the boundary domain is of order \( O(h^2) \) and the variance is of order \( O((nh)^{-1}) \) (the same as at the interior points of the interval). Now we must find such boundary kernels that allow us to get for retransformed estimates the same decay rate at infinity as for the true PDF. This is not possible unless the shape of the distribution tail is known. Since the approaches described above do not use such information, they are not always useful for this purpose.

In particular, the order of decay of the tail of the retransformed estimate (2.18) with kernel \( B_2(x) \) depends on the kernel \( K(x) \) for class (4.16) and transformation (4.11). Expression (4.17) shows that the kernel must be chosen such that

\[ \hat{g}_h(T_{\tilde{\gamma}}(x)) \sim (1 + \tilde{\gamma}x)^{-1/(2\tilde{\gamma})}, \quad x \to \infty, \quad (4.20) \]

to obtain an order of decay at infinity close to that of (4.16). This can be done in two ways (Markovich, 2005c):

(i) the kernel at boundary points \( (Y_{(n)}, 1] \) in the kernel estimate \( \hat{g}_h(x) \) must be chosen equal to the ‘target’ PDF \( g(x) \), or

(ii) for an arbitrary kernel the parameter \( h \) must be determined from the equation

\[ \frac{1}{h}K\left( \frac{T(x) - Y_{(n)}}{h} \right) T'(x) = \hat{f}(x), \quad (4.21) \]

where \( Y_{(n)} \) is the maximal order statistic corresponding to the transformed sample.

Let us consider the first item. For a kernel estimate we obtain at the boundary points \( y \in (Y_{(n)}, 1] \),

\[ \hat{g}_h(y) \simeq \frac{1}{h}K\left( \frac{y - Y_{(n)}}{h} \right), \]

that is, \( \hat{g}_h(y) \) is approximated at boundary points by the last term of the sum (2.18). From (4.11) we find for a triangular kernel \( K(x) = 2(1 - x)1\{x \in [0, 1] \} \) coinciding with the ‘target’ PDF and for the Epanechnikov kernel, that

\[ K\left( \frac{T_{\tilde{\gamma}}(x) - Y_{(n)}}{h} \right) \sim (1 + \tilde{\gamma}x)^{-1/(2\tilde{\gamma})}, \quad x \to \infty, \]
if \( h = 1 - Y(n) \). Use of \( B_2(x) \) in \( \hat{g}_h(y) \) gives (4.20) for \( h = 1 - Y(n) \), that is,

\[
\hat{g}_h(T(y)(x)) \simeq \frac{1}{h} B_2 \left( \frac{T(y)(x) - Y(n)}{h} \right)
= \frac{1}{h} \left( a_2(p) - a_1(p) \frac{T(y)(x) - Y(n)}{h} \right) K \left( \frac{T(y)(x) - Y(n)}{h} \right) \sim (1 + \hat{\gamma}x)^{-1/(2\hat{\gamma})}.
\]

Then by (4.17) we obtain

\[
\hat{f}(x) \simeq (1 + \hat{\gamma}x)^{-(1/\hat{\gamma}+1)}.
\]

Hence, the tail of the estimate \( \hat{f}(x) \) coincides with the tail of (4.16) both for triangular and Epanechnikov’s kernels if \( \hat{\gamma} \) is close to \( \gamma \). But for the biweight kernel \( K_h(x) \),

\[
K \left( \frac{T(y)(x) - Y(n)}{h} \right) \sim (1 + \hat{\gamma}x)^{-1/\hat{\gamma}}, \quad \text{as } x \to \infty,
\]

and (4.20) is not satisfied.

According to the second item, \( h \) can be found at the boundary by the equality

\[
\frac{1}{h} B_2 \left( \frac{T(y)(x) - Y(n)}{h} \right)(1 + \hat{\gamma}x)^{-1/(2\hat{\gamma})-1} = (1 + \hat{\gamma}x)^{-(1/\hat{\gamma}+1)}.
\]

Let us examine the transformation \( T(x) = (2/\pi) \arctan x \). The PDF of the transformed r.v. is defined by (3.25) as

\[
g(x) = \ell \left( \tan \left( \frac{\pi x}{2} \right) \right) \frac{\pi}{2} \frac{1 + \tan^2((\pi/2)x)}{1 + \gamma \tan((\pi/2)x))^{1/\gamma+1}}; \quad (4.22)
\]

\( g(x) \) is continuous in \([0,1]\) if \( \gamma \leq 1 \). If \( \gamma > 1 \), then \( g(x) \to \infty \) as \( x \to 1 \) and it is not easy to estimate \( g(x) \) near \( x = 1 \). Moreover, \( g''(x) \) is continuous on \([0,1]\) for some \( \ell(x) \) if \( \gamma \leq 1/3 \).\(^3\) In the latter case, the estimate with kernel \( B_1 \) or \( B_2 \) could be applied to \( g(x) \) at boundary points to reduce the bias. Otherwise, one can use boundary kernels similar to those recommended in Scott (1992). In any case, since \( T(x) \) does not depend on \( \gamma \), the estimate

\[
\hat{f}(x) = 2 \hat{g}_h \left( \frac{2}{\pi} \arctan x \right) / (\pi(1 + x^2)) \quad (4.23)
\]

does not depend on \( \gamma \) if \( K(x) \) or \( h \) is chosen independently of \( \gamma \). Then the necessary order of the decay for class (4.16) is not ensured.

\(^3\) The parameter \( \gamma \) can be found by some other method from Section 1.2.

\(^4\) The derivative \( \ell(\tan(\pi x/2))'' \) is not continuous for \( \ell''(x) = \exp((\ln(1+x))^{1/2} \cos((\ln(1+x))^{1/2})) \), unlike such \( \ell(x) \) where \( \lim_{x \to \infty} \ell(x) = \ell_\infty, 0 < \ell_\infty < \infty \) (e.g., positive constants or functions converging to positive constants) and \( \ell(x) = \ln x \). The function \( \ell''(x) \) is an example of a slowly varying function that oscillates at infinity, that is, \( \lim_{x \to \infty} \inf \ell(x) = 0 \) and \( \lim_{x \to \infty} \sup \ell(x) = \infty \) (Mikosch, 1999).
Choosing the boundary kernel

\[ K(y) = \frac{\pi}{2} \frac{1 + \tan^2(\gamma y/2)}{(1 + \hat{\gamma}\tan(\gamma y/2))^{1/\hat{\gamma}+1}}, \]

equal to \( g(x) \), we obtain for any \( \gamma > 0 \),

\[ \hat{g}_h(y) \simeq \frac{\pi}{2h} \frac{1 + \tan^2(\gamma y/2h)}{(1 + \hat{\gamma}\tan(\gamma y/2h))^{1/\hat{\gamma}+1}}. \]

Let \( h = 1 - \pi Y_{(n)}/(2\arctan x) \) for some \( x > \tan((\pi/2)Y_{(n)}) \). Then (4.23) tends to the tail of (4.16) as \( x \to \infty \). Here the window width \( h \) depends on \( x \).

The last kernel is somewhat artificial since it is similar to (4.22). We can select a \( \gamma \)-dependent bandwidth \( h \) for the usual kernel \( K(x) \) to ensure a proper order of decay at infinity. For example, if \( \gamma \leq 1/3 \), then at the boundary \( h \) can be found from the equality

\[ \hat{f}(x) \simeq \frac{1}{h} B_2 \left( \frac{2/\pi \arctan(x) - Y_{(n)}}{h} \right) \frac{2}{\pi(1+x^2)} = (1 + \hat{\gamma}x)^{-(1/\hat{\gamma}+1)}. \]

The algorithm for boundary kernel selection is as follows:

1. Select an appropriate transformation function \( T(x) : R_+ \to [0, 1] \).
2. Determine the class of the distribution of r.v. \( X \).
3. Determine the PDF \( g(x) \) of the retransformed r.v. \( Y = T(X) \) by formula (3.25).
4. Determine the kernel \( K(x) \) at boundary points \( (Y_{(n)}, 1] \) coinciding with \( g(x) \).
   An alternative is to select \( h \) for some \( K(x) \) from (4.21).
5. Use kernels \( B_1 \) or \( B_2 \) if \( g''(x) \) is continuous.

### 4.6 Accuracy of a nonvariable bandwidth kernel estimator

We consider a nonvariable bandwidth kernel estimator

\[ \hat{f}_h(x) = (nh)^{-1} \sum_{i=1}^{n} K((x - X_i)/h), \quad (4.24) \]

where the kernel \( K : R \to R \) is a real function, \( h \) is its bandwidth. Let \( K(x) \) be symmetric and vanish outside a compact set. For simplicity, we assume that \( K(x) \)

---

5 The class of the tail can be determined by some parametric test (Jurečková and Picek, 2001). Very roughly, one can find it by the investigation of the sign of the tail index or EVI.
is defined on $[-1, 1]$. Suppose, that $K(x)$ is of second order (see Definition 14),
that is,

$$
\int K(x)dx = 1, \quad \int xK(x)dx = 0,
$$

$$\int x^2K(x)dx = K_1 \neq 0.
$$

The derivative of the PDF $f(x)$ is assumed to be continuous and its second derivative
is bounded. We write

$$K_1 f(x) = \int u^2K(u)f''(x + \theta hu)1\{|u| \leq 1\}du, \quad 0 < \theta < 1.
$$

To get the bias of $\hat{f}_h(x)$ we use the substitution $(y - x)/h = u$, the symmetry
of $K(x)$ ($K(x) = K(-x)$) and apply the Taylor expansion to $f(x + hu)$. It
follows that

$$\text{bias}(\hat{f}_h(x)) = E \hat{f}_h(x) - f(x) = \frac{1}{h} \int K \left(\frac{x - y}{h}\right) f(y)1\{|x - y| \leq h\}dy - f(x) \int K(u)1\{|u| \leq 1\}du$$

$$= \int (f(x + hu) - f(x)) K(u)1\{|u| \leq 1\}du$$

$$= \int \left(huf'(x) + \frac{(hu)^2}{2}f''(x + \theta hu)\right) K(u)1\{|u| \leq 1\}du = h^2 \frac{K_1 f(x)}{2}.
$$

Let

$$K^* = \int K^2(u)1\{|u| \leq 1\}du, \quad K_1^* = \int uK^2(u)1\{|u| \leq 1\}du,$$

$$K_2 f(x) = \int u^2K^2(u)f''(x + \theta hu)1\{|u| \leq 1\}du, \quad 0 < \theta < 1.
$$

Applying the Taylor expansion to $f(x + hu)$, we then get the variance of $\hat{f}_h(x)$:

$$\text{var}(\hat{f}_h(x)) = E \hat{f}_h^2(x) - \left(E \hat{f}_h(x)\right)^2$$

$$= n^{-1}h^{-2} \int K^2 \left(\frac{x - y}{h}\right) f(y)1\{|x - y| \leq h\}dy$$

$$- n^{-1} \left(E \hat{f}_h(x) - f(x) + f(x)\right)^2$$

$$= (nh)^{-1} \int K^2(u)f(x + hu)1\{|u| \leq 1\}du - n^{-1} \left(f(x) + h^2 \frac{K_1^2 f(x)}{2}\right)
$$

$^6$ The simplest example of such kernels is given by a symmetric kernel with compact support.
\[(nh)^{-1} \left( f(x)K^* + hK^*_1 f'(x) + h^2 \frac{K^*_1(x)}{2} \right) - n^{-1} \left( f(x) + h^2 \frac{K^*_1(x)}{2} \right) = (nh)^{-1} f(x)K^* + O \left( n^{-1} \right). \]

Hence, the MSE for a nonvariable bandwidth kernel estimate (4.24) with second-order kernel obeys

\[
\text{MSE}(\hat{f}_h) = E(\hat{f}_h(x) - f(x))^2 = h^4 \left( K^*_1(x) \right)^2 / 4 + (nh)^{-1} f(x)K^* + O \left( n^{-1} \right). \tag{4.27}
\]

Obviously, \( \text{MSE} \sim n^{-4/5} \) if the bandwidth is \( h = h(n) = Dn^{-1/5} \), where \( D \) depends on the unknown PDF \( f(x) \), \( f''(x) \), and the kernel \( K(x) \).

### 4.7 The \( D \) method for a nonvariable bandwidth kernel estimator

For moderate sample sizes, a data-dependent choice of the smoothing parameter of the PDF estimate is a more practical tool than one derived from theory such as \( h = h(n) = Dn^{-1/5} \), where \( D \) is a positive constant.

A well-known data-dependent method is given by cross-validation. However, this method has slow convergence rates and high sampling variability (see Park and Marron, 1990). An alternative data-dependent smoothing tool is determined by the discrepancy method (see Section 2.2.4). Here, we shall find the rate of \( h \) for the nonvariable bandwidth kernel estimate \( \hat{f}_h(x) \) (4.24), when \( h \) is defined by the following version of the discrepancy equation. Let the bandwidth \( h \) be selected from the discrepancy equation

\[
\sup_{x \in \Omega^*} |F_n(x) - F_h(x)| = n^{-\alpha}, \quad \text{as } 0 < \alpha < 1/2, \tag{4.28}
\]

where \( \Omega^* \subseteq (-\infty, \infty) \) is some finite interval, \( F_h(x) = \int_{(-\infty,x] \cap \Omega^*} \hat{f}_h(t)dt \). Without loss of generality, one can take \( \Omega^* = [0, 1] \).

**Theorem 3** Let \( X^n = \{X_1, \ldots, X_n\} \) be i.i.d. r.v.s with a PDF \( f(x) \) that is supported on \( \Omega^* = [0, 1] \). We assume that for \( x \in \mathbb{R}, K(x) \), is continuous, positive, vanishes outside the interval \( [-1, 1] \), and satisfies

\[
\sup_x K(x) \leq C < \infty, \quad \int_{\mathbb{R}} K(x)dx = 1.
\]

Then any solution \( h^*_n = h^*_n(n) \) of (4.28) obeys the condition

\[
h^*_n \to 0, \quad \text{as } n \to \infty.
\]

**Theorem 4** Let the PDF \( f(x) \) be estimated by the nonvariable bandwidth kernel estimator (4.24). Let \( f(x) \) be located on a finite interval \( \Omega^* \). Assume that the
conditions on $K(x)$ given in Theorem 3 hold. In addition, we assume that \( K(x) \) is of second order, \( f(x) \) has two continuous derivatives \( f'(x), f''(x) \), and

\[
\eta_2 \geq |f''(x)|, \quad |f'(x)| \geq \eta_1 > 0, \quad \text{for } x \in \Omega^*, \tag{4.29}
\]

where \( \eta_1 \) and \( \eta_2 \) are constants such that \( \eta_2 \geq \eta_1 (C/2 + 1/4) \). Then any solution \( h_* = h_*(n) \) of equation (4.28) obeys the condition

\[
1 - P\{\rho_1 n^{-\alpha/2} \leq h_* \leq \rho_2 n^{-\alpha/2} \} \leq 2 \exp\left(-n^{1-2\alpha}/(2(2C + 1)^2)\right), \tag{4.30}
\]

where \( \rho_1 = (\eta_2 K_1)^{-1/2}, \rho_2 = 2((2C + 1)K_1 \eta_1)^{-1/2} \) are constants.

\textbf{Remark 3} The PDF \( f(x) = A(2 - (x + 0.1)^2), x \in [0, 1], \) gives an example of a PDF which satisfies the condition (4.29). Here, \( A \) is a normalizing constant giving \( \int_0^1 f(x)dx = 1 \). In fact, \( |f''(x)| = | - 2A | \leq 2A = \eta_2, \ |f'(x)| = | - 2A(x + 0.1)| \geq A/5 = \eta_1 > 0 \). For Epanechnikov’s kernel \( K(x) = \frac{3}{4}(1 - x^2)1\{|x| \leq 1\} \), we have \( K(x) \leq \frac{2}{3} = C \). Then, \( \eta_2/\eta_1 = 10 \geq C/2 + 1/4 = 5/8 \).

\textbf{Theorem 5} Let the PDF \( f(x) \) be estimated by the kernel estimator (4.24). We assume that the conditions on \( f(x) \) and \( K(x) \) given in Theorem 4 hold and \( \alpha = 2/5 \) in (4.28). Then for any solution \( h_* \) of (4.28), we have

\[
P\left\{ \lim_{n \to \infty} n^{4/5}\text{MSE}(\hat{f}_{h_*}) \leq c^* \right\} = 1,
\]

where \( c^* \) is some constant that is independent of the sample size \( n \).

\textbf{Remark 4} It is assumed that \( f(x) \) is located on a finite interval \( \Omega^* \). This implies that, according to (4.14) or (4.15), one requires a preliminary data transformation in order to retain the results of theorems for the heavy-tailed case.

### 4.8 The D method for a variable bandwidth kernel estimator

#### 4.8.1 Method and results

We consider the estimator \( \tilde{f}^A(t \mid h_1, h) \) defined by (3.15). Simple calculus shows that the bandwidth \( h \) which minimizes the corresponding MSE (3.16) at \( x \) is given by \( h(x) = Dn^{-1/9} \), where \( D > 0 \) depends on the unknown PDF \( f(x) \), on \( (d/dx)^4(1/f(x)) \), and on the kernel \( K(x) \). The estimation of the fourth derivative of the inverse PDF is an awkward problem in itself. To avoid it, we shall consider the data-dependent discrepancy method again (Markovich, 2006a).

Let \( h_* \) be a solution of the equation

\[
\sup_{-\infty < x < \infty} |F_n(x) - F^A_{h,h_1}(x)| = \delta n^{-1/2}, \tag{4.31}
\]

where \( F_{h,h_1}^A(x) = \int_{-\infty}^x \tilde{f}^A(t \mid h_1, h)dt \). Further, we assume that the estimate (4.24) is taken as \( \hat{f}_{h_1}(x) \) in (3.15).
Theorem 6  Let $X^n = \{X_1, \ldots, X_n\}$ be i.i.d. r.v.s with PDF $f(x)$. Select the nonrandom bandwidth $h_1 = cn^{-1/5}$, $c > 0$ in $\hat{f}_{h_1}(x)$. We assume that for $x \in \mathbb{R}$, $K(x)$ is continuous and satisfies
\[
\sup_x |K(x)| < \infty, \quad \int_{\mathbb{R}} K(x)dx = 1.
\]
Then any solution $h_n = h_n(n)$ of (4.31) obeys the condition
\[
h_n \to 0, \quad \text{as } n \to \infty.
\]

Theorem 7 Suppose that the PDF $f(x)$ has $m - 1$ continuous derivatives and its $m$th derivative is bounded for a positive integer $m$. Let $f(x)$ be estimated by a variable bandwidth kernel estimate $\hat{f}^A(x|h_1, h)$ as in (3.15). Assume that the conditions on $K(x)$ given in Theorem 6 hold. In addition, we assume that $K(x)$ has order $m + 1$ (see Definition 14) and $\int_{\mathbb{R}} |K(x)|dx = A < \infty$ holds. Let the nonrandom bandwidth $h_1$ in $\hat{f}_{h_1}(x)$ obey the conditions $h_1 \to 0$, $nh_1 \to \infty$ as $n \to \infty$. Then any solution $h_n = h_n(n)$ of (4.31) obeys the condition
\[
P\{h_n > \rho n^{-1/(\alpha(m + 1))}\} < \exp \left(-2n^{1-2/\alpha}\right), \tag{4.32}
\]
where $\rho = (2(1 + A\delta)/G)^{1/(m+1)}$ is a constant, $G = (1/(m+1)) \sup_x |\int_{-\infty}^{\infty} f^{(m)}(x - hy\theta) y^{m+1} K(y)dy|$, $0 < \theta < 1$, for any $\alpha > 2$.

Remark 5 Pareto (4.9), exponential, and normal PDFs are examples of PDFs that satisfy the conditions of Theorem 7.

Let $\mathcal{M}$ be a compact set in $\mathbb{R}$. Given $\varepsilon > 0$, we use the following notation of Hall and Marron (1988):
\[
\mathcal{M}^\varepsilon \equiv \{x \in \mathbb{R} : \text{for some } y \in \mathcal{M}, \|x - y\| \leq \varepsilon\},
\]
where $\|\cdot\|$ is the usual Euclidean norm.

Theorem 8 Let $f(x)$ and $1/f(x)$ have four continuous derivatives and $f(x)$ be bounded away from zero on $\mathcal{M}^\varepsilon$. Let the PDF $f(x)$ be estimated by a variable bandwidth kernel estimate $\hat{f}^A(x|h_1, h)$ as in (3.15). Assume that the conditions on $K(x)$ given in Theorem 7 hold for $m = 3$. Furthermore, assume that $K(x)$ is symmetric, has two bounded derivatives, and vanishes outside a compact set. Assume that the nonrandom bandwidth $h_1$ in (3.15) obeys $h_1 = c_n n^{-1/5}$, where $c_n > 0$ is some constant. Then, for any solution $h_n$ of (4.31), we have
\[
P \left\{ \limsup_{n \to \infty} n^{4/9} \left( E\hat{f}^A(x|h_1, h_n) - f(x) \right) \leq \psi(x) \right\} = 1,
\]
where $\psi(x) = (K_3/24) (d/dx)^4 (1/f(x)) \rho^A$, and $\rho$ is defined in Theorem 7.

Corollary 1 Assume that the conditions of Theorem 8 hold. Assume that $E(Z \cdot \hat{f}^A(x|h)) = 0$, where $Z$ is a standard normal r.v. Then $\text{MSE}(\hat{f}^A(x|h_1, h_n))$ may reach order $n^{-8/9}$ if a maximal solution of (4.31) $h_n$ has order $n^{-1/9}$.
Remark 6 Since the function of the r.v. \( X_1 \) (i.e., one term in the sum \( \hat{f}^A(x|h) \)) and the normally distributed r.v. \( Z \) are independent, the condition \( E(Z \cdot \hat{f}^A(x|h)) = 0 \) is not rigorous.

Remark 7 In Theorem 8 we assume that \( K(x) \) has a compact support. This assumption is not reliable if tail estimation is the object of interest. In this case, a data transformation is required.

By way of an illustrative example, Figure 4.4 shows the prevalence of the retransformed kernel estimate in the tail domain in comparison with the variable bandwidth kernel estimate without a preliminary transformation. The adaptive transformation (4.11) is used. One can observe the truncation of estimate (3.15) beyond the sample maximum (\( X_{\text{max}} = 13.3 \)).

### 4.8.2 Application to Web traffic characteristics

We apply the kernel estimators (3.15) and (4.24) to the Web data, where \( h \) is estimated by the discrepancy method (2.42); see Markovich (2006a). These data have already been described in Table 1.4. To simplify the calculation the data were scaled, i.e. all values were divided by the scaling parameter \( s \).

To check whether the measurements corresponding to the s.s.s., s.r., d.s.s. and i.r.t. samples are derived from heavy-tailed distributions, we have estimated the EVI \( \gamma \) by Hill’s method (1.5). In Table 4.1 one can see the Hill estimates \( \hat{\gamma} = \hat{\gamma}^H(n, k) \). The numbers of retained data \( k \) for all data sets are selected by the bootstrap method (Markovich, 2005a; see also pp. 36, 37 above). Observing \( \hat{\gamma} \), one may conclude that
the estimates of the tail index $\alpha = 1/\gamma$ are always less than 2 for all data sets considered. It follows from the extreme value theory (Embrechts et al., 1997) that at least the $\beta$th moments, $\beta \geq 2$, of the distribution of s.s.s., d.s.s., s.r., i.r.t. may be not finite if we believe that the distribution has a regularly varying tail. The positive sign of $\hat{\gamma}$ indicates that distributions of the Web traffic characteristics considered are heavy-tailed.

Figure 4.5 The retransformed standard kernel estimate (4.24) (solid line) and variable bandwidth estimate (3.15) (dotted line) for the s.s.s. (left) and d.s.s. (right) data sets. For both estimates the bandwidth $h$ is selected by discrepancy method (2.42). The data transformation (4.11) is used. The curves nearly coincide for d.s.s. Reprinted from Proceedings of 2nd Conference on Next Generation Internet Design and Engineering, Valencia, Estimation of heavy-tailed density functions with application to WWW-traffic, Markovich NM, Figure 1, (c) 2006 IEEE. With permission from IEEE.
Hence, we may use (4.11) to transform the data. The PDF $g_0(x)$ of a new r.v. has been estimated by (4.24) and (3.15) with Epanechnikov’s kernel. The retransformed estimate of the unknown PDF $f(x)$ was calculated by (3.26):

$$\hat{f}(x) = 0.5g_0(1 - (1 + \hat{y}x)^{-1/(2\hat{\gamma})}(1 + \hat{y}x)^{-1/(2\hat{\gamma})}^{-1}.$$  

The bandwidths $h_s$ and $h_v$ in Table 4.1 have been selected by the discrepancy method (2.42) and correspond to estimates (4.24) and (3.15), respectively. The value $h_1$ of the nonvariable kernel estimate $\hat{f}_{h_1}(x)$ in (3.15) is calculated by (2.30). For Epanechnikov’s kernel we get $R(K) = 3/5$, $\mu_2(K) = 1/5$. This formula provides the minimal upper bound of the theoretical value of $h$ that corresponds to the optimal MSE $\sim n^{-4/5}$ of estimate (4.24).

The retransformed kernel estimates (4.24) and (3.15) have been calculated for the d.s.s. and s.s.s., s.r. and i.r.t. samples (see Figures 4.5 and 4.6). The estimate $f(x) = g(x/s)/s$ is shown, where $g(x/s)$ is the retransformed estimate constructed from scaled data. A logarithmic scale is used for both the $X$- and $Y$-axes.

The curves of the retransformed kernel estimate (4.24) corresponding to all data sets apart from d.s.s. and of the retransformed kernel estimate (3.15) for the sample s.r. are truncated for large values of $x/s$ because the kernel is not wide enough. Such boundary effects are typical of kernel estimates that are used for compactly

![Figure 4.6](image-url)  

**Figure 4.6** The retransformed standard kernel estimate (4.24) (solid line) and variable bandwidth estimate (3.15) (dotted line) for the s.r. (left) and i.r.t. (right) data sets. For both estimates the bandwidth $h$ is selected by discrepancy method (2.42). Data transformation (4.11) is used. Reprinted from *Proceedings of 2nd Conference on Next Generation Internet Design and Engineering*, Valencia, Estimation of heavy-tailed density functions with application to WWW-traffic, Markovich NM, Figure 2, (c) 2006 IEEE. With permission from IEEE.
supported PDFs. In this case, the kernel estimate of the PDF \( g_0(x) \) located on \([0,1]\) may be equal to zero at the neighborhood of 1 beyond the maximum observation of the sample. This reflects on the retransformed estimate. It becomes equal to zero in the tail and the logarithms of these values go to \(-\infty\). In Maiboroda and Markovich (2004) it was shown that the choice \( h = 1.01 - T_{\hat{\gamma}}(X(n)) \) in the neighborhood of 1, where \( T_{\hat{\gamma}}(x) \) is the transformation (4.11) and \( X(n) \) is the maximal observation in the sample, may improve the boundary problems. One can compare the values of \( h_s, h_v \) and \( 1.01 - T_{\hat{\gamma}}(X(n)) \) in Table 4.1. Obviously, the discrepancy method selects larger values \( h \) that are closer to \( 1.01 - T_{\hat{\gamma}}(X(n)) \), for estimate (3.15) than for estimate (4.24). Hence, the retransformed variable bandwidth estimate provides a better estimation of the PDF at the tail domain for Web traffic characteristics.

4.9 The \( \omega^2 \) method for the projection estimator

Let us consider the projection estimator (2.14). The selection of the smoothing parameter \( \gamma \) of this estimator was studied in Vapnik et al. (1992). The latter parameter plays the same role as the bandwidth \( h \) of kernel estimates.

**Theorem 9** Let \( X_1, \ldots, X_n \) be a sample of i.i.d. r.v.s with PDF \( f(x) \in \mathcal{V} \). If we take

\[
\gamma = n^{-1/(2k+2)}
\]

in (2.14), then the asymptotic rate of convergence of the estimates \( \hat{f}_{\gamma}(x, X^n) \) to \( f(x) \) is given by the expressions

\[
P \left\{ \lim_{n \to \infty} n^{(k+1/2)/(2k+2)} \frac{\| \hat{f}_{\gamma}(x, X^n) - f(x) \|}{\sqrt{\ln n}} \leq c \right\} = 1,
\]

\[
P \left\{ \lim_{n \to \infty} n^{(k+1/2)/(2k+2)} E \| \hat{f}_{\gamma}(x, X^n) - f(x) \| \leq c \right\} = 1,
\]

where \( c \) is a quantity that is independent of \( n \).

By \( \| \cdot \| \) we mean the \( L_2 \)-norm, while by \( E \) we mean the expectation with respect to the measure \( f(x) \). The proof is given in Appendix B.

We note that the convergence rate, indicated in the second relation, coincides with the maximal attainable rate in the class \( \mathcal{V} \) according to Čencov (1982). Experience shows that for samples of moderate size \( n \) the selection of \( \gamma \) derived from Theorem 9 leads to unsatisfactory PDF estimates.

Assume now that \( \gamma \) in (2.14) is selected by the variant of the \( \omega^2 \) method (see Section 2.2.4) which consists of the following steps:

\[\text{See Example 5 (Section 2.1) for the definition of class } \mathcal{V}\]
1. If the inequality
\[ \frac{n}{2\pi^2} \sum_{j=1}^{\infty} \left( \frac{a_j}{j} \right) \geq \delta \] (4.33)
is satisfied, then \( \gamma \) is obtained from the equality
\[ \hat{\omega}_n^2 = n \int_0^1 (F_n(x) - F\gamma(x))^2 \hat{f}(x) \, dx = \delta, \] (4.34)
where \( F\gamma(x) = \int_0^x \hat{f}_\gamma^\text{pr}(t, X^n) \, dt \), \( \hat{\omega}_n^2 \) is the estimator of the von Mises–Smirnov statistic
\[ \omega_n^2 = n \int (F_n(x) - F(x))^2 f(x) \, dx, \]
and
\[ \hat{f}(x) = \begin{cases} \hat{f}_\gamma^\text{pr}(x), & \text{if } \hat{f}_\gamma^\text{pr}(x) \geq \beta, \\ \beta, & \text{if } \hat{f}_\gamma^\text{pr}(x) < \beta; \end{cases} \] (4.35)
\( \beta > 0 \) is some arbitrary constant and \( \delta \) is the mode of the distribution \( \omega_n^2 \).
Such a \( \gamma \) need not be unique.

2. If inequality (4.33) is not satisfied, then we take
\[ \gamma = n^{-1/(2k+2)}. \]

**Theorem 10** Let \( X_1, \ldots, X_n \) be a sample of i.i.d. r.v.s with PDF \( f(x) \in \Phi \). If (4.33) is satisfied, then for a sufficiently large sample size, there exists, with probability arbitrarily close to unity, at least one \( \gamma \) satisfying (4.34). It is contained in the interval \( \left[ (G/n)^{1/(2k+3)}, \infty \right) \), where \( G = G(f) \) is some constant. If there exist various \( \gamma \) satisfying (4.34), then for their maximum \( \gamma_{\max} \) we have the inequality
\[ P\{ \gamma_{\max} < (G/n)^{1/(2k+3)} \} < 3n^{-9/8}. \]

**Theorem 11** Let \( X_1, \ldots, X_n \) be a sample of i.i.d. r.v.s with PDF \( f(x) \in \Phi \). If the regularization parameter \( \gamma \) of the PDF estimate \( \hat{f}_\gamma^\text{pr}(x, X^n) \) is obtained by the \( \omega^2 \) method, then we have the equality
\[ P \left\{ \lim_{n \to \infty} n^{(k+1)/2k+3} \| \hat{f}_\gamma^\text{pr}(x, X^n) - f(x) \| \leq c \right\} = 1, \]
where \( c \) is a quantity that is independent of \( n \).

**Remark 8** Although this rate of convergence is smaller than the maximum possible in the class \( \Phi \), in practice the \( \omega^2 \) method enables us to obtain reliable results for a wide range of \( f(x) \), even from rather moderate samples.

**Remark 9** The class \( \Phi \) contains the PDFs with compact support [0,1]. In order to apply the projection estimator (2.14) and the \( \omega^2 \) method to a heavy-tailed PDF, one has to transform the data to a compact interval by means of some transformation function \( T(x) \).
4.10 Exercises

1. Retransformed kernel estimates.

Generate $X^n$ according to some heavy-tailed distribution (e.g., Pareto, Weibull with shape parameter less than 1, lognormal), or take heavy-tailed real data. Estimate the shape parameter $\gamma$ by Hill’s method (1.5). Transform the sample $X^n$ into a new one $Y^n$ by the transformations $T(x) = \ln x$, $T(x) = (2/\pi) \arctan x$ and $T_\gamma(x) = 1 - (1 + \gamma x)^{-1/(2\gamma)}$ ($Y_i = T(X_i)$, $i = 1, \ldots, n$), where $\hat{\gamma}$ is the estimate of the EVI $\gamma$.

Calculate the standard kernel estimate $\hat{g}_h(x)$ by formula (4.24).

Take $h = \sigma n^{-1/5}$, where $\sigma$ is an empirical standard deviation calculated from the sample $Y^n$ and $K(x) = \frac{2}{4}(1 - x^2)1\{|x| \leq 1\}$.

Calculate the PDF $\hat{f}_h(x)$ of the initial r.v. $X_1$ by the formula

$$\hat{f}_h(x) = \hat{g}_h(T(x)) T'(x).$$

(4.36)

For generated data, compare the retransformed estimates for different heavy-tailed PDFs using the loss functions in the metric spaces $L_1$, $L_2$ and $C$:

$$\chi^1 = \int_{-\infty}^{\infty} |\hat{f}_h(x) - f(x)| dx = \int_{0}^{1} |\hat{g}_h(x) - g(x)| dx,$$

$$\chi^2 = \int_{-\infty}^{\infty} (\hat{f}_h(x) - f(x))^2 dx,$$

$$\chi^3 = \sup_{i=1, \ldots, n} |\hat{f}_h(X_i) - f(X_i)|,$$

where $f_n(x)$, $g_n(x)$ are the estimates of the PDF and $f_0(x)$, $g_0(x)$ are the exact models of the PDF arising from the initial and the transformed r.v..

For each sample size $n = 50, 100,$ and $300$ construct $l = 25$ realizations. Calculate the statistics

$$\bar{\rho}_j = \frac{1}{l} \sum_{i=1}^{l} \chi_j^i, \quad \sigma_j^2 = \frac{1}{l-1} \sum_{i=1}^{l} \left( \chi_j^i - \bar{\rho}_j \right)^2, \quad l = 25, j = 1, 2, 3.$$

2. Repeat Exercise 1 with polygram (2.29) instead of the kernel estimate. Exclude the transformation $T(x) = \ln x$ from consideration since it leads to an infinite interval $(-\infty, \infty)$.

3. Comparison of smoothing methods for retransformed kernel estimates.

Generate $X^n$ according to some heavy-tailed distribution or take heavy-tailed real data. Transform the sample $X^n$ to $Y^n$ by the transformation $T_\gamma(x) = 1 - (1 + \gamma x)^{-1/(2\gamma)}$. Using $Y^n$, calculate a kernel estimate $\hat{g}_h(x)$ by (4.24) and then $\hat{f}_h(x)$ by (4.36). Take Epanechnikov’s kernel function $K(x)$ and $h = \sigma n^{-1/5}$ as in Exercise 1.

Also, find $h$ of the estimate $\hat{g}_h(x)$ as a solution of the discrepancy equations

$$\sum_{i=1}^{n} \left( \hat{G}_h(Y_i) - \frac{i-0.5}{n} \right)^2 + \frac{1}{12n} = 0.05, \quad \omega^2 \text{ method},$$

where $\hat{G}_h(Y_i)$ is the estimate of the EVI $\gamma$. Calculate the statistics

$$\bar{\rho}_j = \frac{1}{l} \sum_{i=1}^{l} \chi_j^i, \quad \sigma_j^2 = \frac{1}{l-1} \sum_{i=1}^{l} \left( \chi_j^i - \bar{\rho}_j \right)^2, \quad l = 25, j = 1, 2, 3.$$
\( \sqrt{n \hat{D}_n} = \sqrt{n} \max(\hat{D}_n^+, \hat{D}_n^-) = 0.5, \quad D \text{ method,} \)

where

\[
\sqrt{n \hat{D}_n^+} = \sqrt{n} \max_{1 \leq i \leq n} \left( \frac{i}{n} - \hat{G}_h(Y_{(i)}) \right), \quad \sqrt{n \hat{D}_n^-} = \sqrt{n} \max_{1 \leq i \leq n} \left( \hat{G}_h(Y_{(i)}) - \frac{i-1}{n} \right),
\]

\( \hat{G}_h(x) = \int_{-\infty}^{x} \hat{g}_h(t) \, dt, \) and \( Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(n)} \) are the order statistics.

For the generated data, draw plots of \( \hat{f}_h(x) \) and compare the selection of \( h \) by the \( D \) and \( \omega^2 \) methods and \( h = \sigma n^{-1/5} \).

4. Repeat Exercise 3 with variable bandwidth kernel estimate (3.15) as the estimate of the PDF of a new r.v. \( Y_1 \). Use kernel estimate (4.24) as a pilot estimate \( \hat{f}_{h_1}(x) \) in (3.15). Calculate the value \( h_1 \) by the formula \( \sigma n^{-1/5} \), where \( \sigma \) is an empirical standard deviation constructed from the sample \( Y^n \).

5. Boundary kernels.

Generate 100 Fréchet distributed r.v.s with DF

\[ F(x) = \exp \left( -\gamma x \right)^{-1/\gamma} 1 \{ x > 0 \}, \quad \gamma = 1. \]

Calculate Hill’s estimate \( \hat{\gamma} \) of \( \gamma \). Transform the sample \( X^n \) to a new one \( Y^n \) by the transformation \( T_{\hat{\gamma}}(x) = 1 - (1 + \hat{\gamma} x)^{-1/(2\hat{\gamma})} \) \( (Y_i = T(X_i), i = 1, \ldots, n) \). Estimate the PDF \( g(x), x \in [0, 1] \) of the new r.v. \( Y_1 \) by kernel estimator (4.24).

Select Epanechnikov’s kernel \( K(x) = \frac{3}{4} (1 - x^2), \) \( |x| \leq 1 \) for the interval \([0, Y_{(n)}]\). Select \( h \) as \( \sigma n^{-1/5} \), where \( \sigma \) is an empirical standard deviation constructed from the sample \( Y^n \) or calculate \( h \) by the \( \omega^2 \) method.

Select the kernel \( K(y) = (1/h) B_2((y - Y_{(n)})/h) \) for the boundary domain \((Y_{(n)}, 1]\). Here, the kernel \( B_2 \) is determined by formulas (4.18), (4.19), where \( K(x) \) is Epanechnikov’s kernel. Use \( h = 1 - Y_{(n)} \) for this case.

Find the estimate of \( X_1 \) by the inverse transform formula (4.36).

Compare the estimate with the true PDF in the boundary domain.
In this chapter, the retransformed density estimates are applied to the classification problem. The Bayesian classification algorithm is considered. The retransformed kernel and polygram estimators are used to estimate heavy-tailed PDFs of each class. A new criterion for the quality of a density estimation is implemented. The classifiers obtained are compared in a simulation study. Possible applications of this classification technique to Web traffic data analysis and Web prefetching schemes are considered.

5.1 Classification and quality of density estimation

From the practical point of view, the ability of a PDF estimate to solve a specific problem is a much more important characteristic than its deviation from the true PDF in some functional space. Here, we wish to compare the PDF estimates in terms of the probability of classification (pattern recognition) error. This is a measure of the quality of the classifiers that are constructed by means of these PDF estimates (Maiboroda and Markovich, 2004, p. 579).
We assume that the observed object $O$ belongs to one of $M$ different populations $P_k$, $k = \{1, \ldots, M\}$. The true population is unknown. Let $p_k = P(O \in P_k)$ be the a priori probability that $O$ belongs to a population $P_k$. We observe some random characteristic (feature) $X \in R_+$ of $O$. For $O \in P_k$ we denote the PDF of $X$ by $f_k(x)$. Our aim is to estimate the type $k$ of the population that $O$ belongs to, from the observation $X$. The solution of this problem is given by a classifier, which is a function $\eta : R_+ \rightarrow \{1, \ldots, M\}$ which assigns the estimated type $k \in \{1, \ldots, M\}$ of the population to a value of the characteristic $X = x$.

We suppose that the penalty for making a mistake depends on the type of the true population $k$ and the value of the characteristic $x$, and denote it by $q_k(x)$. Then the probability of a misclassification is given by the mean of the loss

$$\mathcal{L}(\eta) = E \sum_{k=1}^{M} 1\{O \in P_k, \eta(X) \neq k\} q_k(X) = \sum_{k=1}^{M} p_k \int_{\eta(x) \neq k} q_k(x) f_k(x) dx. \tag{5.1}$$

The smallest probability of a misclassification is attained for the Bayesian classifier (Devroye and Györfi, 1985)

$$\eta_B(x) = k \quad \text{if} \quad p_k q_k(x) f_k(x) \geq p_i q_i(x) f_i(x) \quad \text{for all} \quad i \neq k, i, k \in \{1, \ldots, M\}.$$ 

Usually, the true $f_k(x)$ and $p_k$ are unknown, but there are some consistent estimates $\hat{f}_k(x), \hat{p}_k$ available. Then the empirical Bayesian classifier is used:

$$\eta_{EB}(x) = k \quad \text{if} \quad \hat{p}_k q_k(x) \hat{f}_k(x) \geq \hat{p}_i q_i(x) \hat{f}_i(x) \quad \text{for all} \quad i \neq k, i, k \in \{1, \ldots, M\}.$$ 

It is obvious that $\mathcal{L}(\eta_{EB}) \geq \mathcal{L}(\eta_B)$. Moreover, $\mathcal{L}(\eta_{EB})$ defined by (5.1) is a r.v. since $\eta_{EB}(x)$ is a random function depending on those data from which the estimates $\hat{f}_k(x)$ were evaluated.

The penalties $q_k(x)$ are defined by the loss caused by misclassification in a real classification problem. If $q_k(x) = 1$ for all $k$, one obtains the probability of misclassification as measure of the risk. Since observations in the tail domain are rare, the improvement of the classification in the tail provides a negligible decrease in the probability of misclassification. The effect of accurate tail classification is only significant if the misclassification of tail observations is more dangerous than errors in the body domain. For example, the classification at outliers (large claims or huge files) could be important for insurance or Web traffic analysis, respectively. Therefore, we consider such $q_k(x)$ as are larger in the tail and smaller in the body, that is, $q_k(x) \rightarrow \infty$ as $x \rightarrow \infty$. At the same time, the Bayesian approach can be applied only if the condition

$$\int_{0}^{\infty} q_k(x) f_k(x) dx < \infty \tag{5.2}$$

holds.

Penalties $q_k(x)$ should be determined such that the Bayesian classifier is switched from one population to another in the tail region, that is, $q_i(x) f_i(x) = q_j(x) f_j(x), i \neq j$, for sufficiently large $x$ in order to avoid the dominance of certain
Figure 5.1 Selection of penalty functions for two Pareto PDFs (4.9) with parameters $\gamma_1 = 0.3$ and $\gamma_2 = 1.2$: $q_1(x)\psi_{\gamma_1}(x)$ (solid line), $q_2(x)\psi_{\gamma_2}(x)$ (dotted line).

$q_i(x)f_i(x)$ over others at tail points. Otherwise, the Bayesian classifier assigns these points to the same class and an enhanced PDF estimation accuracy has almost no effect. Figure 5.1 shows an example of such a selection for two populations. Here $q_1(x) = \sqrt{x+10}$, $q_2(x) = \sqrt{x}$, and $p_1 = p_2 = 0.5$.

With regard to PDF estimation, classification means that we estimate $M$ different PDFs. Then we use the estimates in the empirical Bayesian classifier for objects associated with these PDFs. That estimate is the best one which provides minimal probabilistic classification error.

However, the classifier $\eta_{EB}$ is not sensitive to the quality of the PDF estimates; for example, PDF estimates of different accuracy that assign the objects to the same class may have the same value $\mathcal{L}(\eta_{EB})$. Therefore, as a criterion for the quality of a PDF estimate, we can use the estimate of $\mathcal{L}(\eta_{EB})$ given by

$$\mathcal{L}(\hat{\eta}_{EB}) = \sum_{i=1}^{M} \hat{p}_i \int_0^\infty q_i(x) \hat{f}_i(x) dx - \int_0^\infty \max_i \hat{p}_i q_i(x) \hat{f}_i(x) dx$$

(Markovich, 2002). Here $\mathcal{L}(\hat{\eta}_{EB})$ is the probability of a misclassification by the Bayesian classifier of the true PDF estimates $\hat{f}_i(x)$ and the a priori probabilities $\hat{p}_i$, $i = 1, \ldots, M$, of classes. Therefore, $\mathcal{L}(\hat{\eta}_{EB})$ fluctuates near $\mathcal{L}(\eta_B)$. The greater the accuracy of the PDF estimate, the closer is $\mathcal{L}(\hat{\eta}_{EB})$ to $\mathcal{L}(\eta_B)$. Hence,

$$|\mathcal{L}(\hat{\eta}_{EB}) - \mathcal{L}(\eta_B)| = \sum_{i=1}^{M} \int_0^\infty \left( \hat{p}_i q_i(x) \hat{f}_i(x) - p_i q_i(x) f_i(x) \right) dx$$

$$+ \int_0^\infty \left( \max_i p_i q_i(x) f_i(x) - \max_i \hat{p}_i q_i(x) \hat{f}_i(x) \right) dx$$

$$\leq \sum_{i=1}^{M} \int_0^\infty q_i(x) \left( \hat{p}_i \hat{f}_i(x) - p_i f_i(x) \right) dx$$
CLASSIFICATION AND RETRANSMFORMED DENSITY ESTIMATES

\[ + \max_i \left| \int_0^\infty q_i(x) \left( p_i f_i(x) - \hat{p}_i \hat{f}_i(x) \right) dx \right| \]

\[ \leq (M + 1) \max_i \left| \int_0^\infty q_i(x) \left( \hat{p}_i \hat{f}_i(x) - p_i f_i(x) \right) dx \right|. \] (5.3)

Let \( T : R^1 \rightarrow [0, 1] \) be a strictly monotonically increasing one-to-one transformation. The inverse transformation \( T^{-1} \) and derivatives \( T', (T^{-1})' \) are assumed to be continuous. Let \( g_i(y) \) be the PDF of the transformed r.v. \( Y = T(X) \) if \( O \in P_i \), and \( G_i(y) \) and \( G_i^n(y) \) be the corresponding DF and empirical DF, respectively. By virtue of (3.26), we have

\[ \left| \int_0^\infty q_i(x) \left( p_i f_i(x) - \hat{p}_i \hat{f}_i(x) \right) dx \right| = \left| \int_0^1 q_i(T^{-1}(x)) (p_i g_i(x) - \hat{p}_i \hat{g}_i(x)) dx \right| \]

\[ \leq \int_0^1 |q_i(T^{-1}(x)) (p_i g_i(x) - E(\hat{p}_i \hat{g}_i(x)))| dx \]

\[ + \int_0^1 |q_i(T^{-1}(x)) (E(\hat{p}_i \hat{g}_i(x)) - \hat{p}_i \hat{g}_i(x))| dx. \] (5.4)

Thus, the deviation of \( \mathcal{L}(\hat{\eta}_{EB}) \) from \( \mathcal{L}(\eta_B) \) is dependent on the accuracy of the PDF estimation of the transformed r.v. in \( L_1 \).

5.2 Convergence of the estimated probability of misclassification

We consider the rate of convergence of \( \mathcal{L}(\hat{\eta}_{EB}) \) to \( \mathcal{L}(\eta_B) \). We make the following assumptions:

1. All the r.v.s are positive.

2. The penalties \( q_i(x) \) satisfy, in addition to (5.2), the condition

\[ \int_0^\infty q_i(x) dT(x) < \infty. \] (5.5)

These conditions are satisfied if, for example, \( q_i(x) = (d_i + x)^\sigma, f_i(x) \simeq (1 + \gamma x)^{-(1 + \gamma]}, \hat{f}_i(x) \simeq (1 + \hat{\gamma} x)^{-(1 + \hat{\gamma})}, T(x) = 1 - (1 + \hat{\gamma} x)^{-(\hat{\gamma})^{-1}}, d_i > 0 \) is a constant, and \( 0 < \sigma < (2 \gamma)^{-1} \).

The following theorems are proved in Markovich (2002); the proofs are reproduced in Appendix C.

**Theorem 12** Let \( X_1, \ldots, X_n \) be i.i.d. r.v.s with PDF \( f(x) \). The transformation (4.11) to a triangular PDF \( g(x) = 2(1 - x)1 \{ x \in [0, 1] \} \) is considered. If the PDF of the transformed r.v. \( Y = T(X) \) is estimated by the kernel estimate

\[ \hat{g}(x) = \frac{1}{nh} \sum_{i=1}^n K \left( \frac{x - x_i}{h} \right), \]
where \(|K(x)| \leq c\), \(K(x) \neq 0\) for \(|x| \leq 1\), and \(K(x) = 0\) otherwise, \(K_h(x) = (1/h)K(x/h)1\{|x| \leq h\}, \int_0^h K_h(x)dx = \int_0^1 K(u)du = 1\) for \(0 < h < 1\), and the bandwidth is \(h = n^{-\beta}, \beta > 0\), \(0 < d < \min(0, 5; \beta)\), then

\[
P\left\{ \lim_{n \to \infty} n^d |\mathcal{L}(\tilde{\eta}_{EB}) - \mathcal{L}(\eta_B)| \leq c_1 \right\} = 1,
\]

where \(c_1\) is a positive constant independent of \(n\).

**Theorem 13**  Let \(X_1, \ldots, X_n\) be i.i.d. r.v.s with PDF \(f(x)\). If the PDF \(g(x)\) of the transformed r.v. \(Y = T(X)\) is such that \(g(x) \leq c < \infty, |g'(x)| < \infty\) for \(\{x : g(x) > 0\}\), and is estimated by the polygram (2.29), where \(m = \frac{n+1}{2}, L = n^\beta, 0 < \beta < 1\) is the smoothing parameter, and \(0 < d < \frac{\beta}{2}\), then

\[
P\left\{ \lim_{n \to \infty} n^d |\mathcal{L}(\tilde{\eta}_{EB}) - \mathcal{L}(\eta_B)| \leq c_1 \right\} = 1,
\]

where \(c_1\) is a positive constant independent of \(n\).

**Remark 10**  Polygram and kernel estimates have identical asymptotic convergence rates that are no worse than \(n^{-d}\), where \(0 < d < 0.5\).

### 5.3 Simulation study

The proximity of \(\mathcal{L}(\tilde{\eta}_{EB})\) (and \(\mathcal{L}(\eta_{EB})\)) to \(\mathcal{L}(\eta_B)\) is compared for the kernel estimate and a polygram (2.29). The fixed transformation \(T(x) = (2/\pi)\arctan(x)\) and adaptive transformation \(T_\gamma(x)\) (4.11) are applied to both estimates. Pairs of populations with PDFs \(f_1(x)\) and \(f_2(x)\) are used. For example, the latter may be the populations of file sizes of HTML and streaming video connections arising from Web sessions.

The mean loss due to misclassification for two classes is given by (Maiboroda and Markovich, 2004, pp. 580–582)

\[
\mathcal{L}(\eta) = p_1 \int_0^\infty q_1(x)f_1(x)1\{\eta(x) = 2\}dx + p_2 \int_0^\infty q_2(x)f_2(x)1\{\eta(x) = 1\}dx.
\]

Here \(\eta : R_+ \to \{1, 2\}\) is the classifier. Since the exact Bayesian risk of the misclassification \(\mathcal{L}(\eta_B)\) (the best possible) can be computed for known PDFs, the relative accuracy of the classifier \(\eta_{EB}(x)\) and a PDF estimate is taken equal to

\[
J_c(\eta_{EB}) = \frac{(1/m) \sum_{i=1}^m \mathcal{L}_i(\eta_{EB})}{\mathcal{L}(\eta_B)} - 1,
\]

and

\[
J_d(\eta_{EB}) = \left| \frac{(1/m) \sum_{i=1}^m \mathcal{L}_i(\tilde{\eta}_{EB})}{\mathcal{L}(\eta_B)} - 1 \right|,
\]

where

\[
\mathcal{L}(\tilde{\eta}_{EB}) = \tilde{p}_1 \int_0^\infty q_1(x)\tilde{f}_1(x)1\{\eta_{EB}(x) = 2\}dx + \tilde{p}_2 \int_0^\infty q_2(x)\tilde{f}_2(x)1\{\eta_{EB}(x) = 1\}dx.
\]
The mean is taken over all $m$ samples of the Monte Carlo study. The worse the PDF estimate, the larger the value $J_d(\eta_{\text{EB}})$ is. The larger $J_*(\eta_{\text{EB}})$ is, the worse the classifier $\eta_{\text{EB}}(x)$ is.

**Investigation of the quality of the classifiers**

Using the adaptive transformation $T_\gamma(x)$, the performance of an empirical Bayesian classifier can be improved only if

- the estimator of the EVI $\hat{\gamma}_n$ is sufficiently accurate (for Hill’s estimator this means that the EVI $\gamma$ and the sample size should be large enough);
- the EVIs of the compared PDFs differ noticeably;
- the misclassification losses $q_i(x)$ are significantly high at the tail domain (in comparison to the losses in the body domain).

We generated samples of the known PDFs of a Pareto distribution (4.9) with EVI $\gamma \in \{1, 3\}(f_p)$ and a Fréchet distribution with DF (1.32), with $\gamma \in \{1, 2\}(f_f)$. We also considered Pareto(2)–Fréchet(0.3) mixtures PDFs ($f_{pf}$). Here $p_1 = p_2 = 0.5$ were taken as proportions of the classes. We used the penalty functions $q_k(x) = x^{1/4}$, which ensure the convergence of (5.2). In Table 5.1, $J_*(\eta_{\text{EB}})$ is shown for two types of transformed polygram ($Pl$ and $Plf$) and transformed kernel estimates with Epanechnikov’s kernel ($Ke$ and $Kef$). $Pl$ and $Ke$ are calculated using the transformation $T_\gamma(x)$, but $Plf$ and $Kef$ are calculated using the fixed transformation $T(x)$. The ratios of $J_*(\eta_{\text{EB}})$ for the polygram and kernel estimates and both transformations are given in the last two columns.

A smoothing parameter is calculated by the formula

$$h = \sigma n^{-1/5}$$

($\sigma$ is the standard deviation of the transformed data), which performs well with Epanechnikov’s kernel; see Devroye and Györfi (1985). The number of quantile intervals in the polygram was chosen as $[n/10]$ for the sample size $n \in \{50, 100, 300, 500\}$. Here $[r]$ denotes the integer part of $r$. The number of Monte Carlo repetitions is given by $m = 1000$. We used Hill’s estimate for the EVI with the smoothing parameter $k = 0.1n$.

The Pareto–Fréchet mixture was investigated in more detail. Namely, for $f_{pf}^1$ the integral in the measure $\mathcal{L}(\eta_{\text{EB}})$ is taken over $[0, \infty)$ and for $f_{pf}^2$ over $[6, \infty)$, i.e., only in the tail domain. For ease of understanding, the results of Table 5.1 are presented in Figures 5.2 and 5.3.

The simulation study shows the following:

- The adaptive transformation (4.11) always improves the quality of the classification of the kernel estimate (in comparison to the fixed arctan transformation), but makes the polygram worse for relatively small samples.
• The superiority of $T_{\gamma}$-transformed estimates becomes more evident if the sample size increases.

• A kernel estimate is better than a polygram if the classification on the unrestricted domain is considered. However, if the tail domain classification is significant, then a smoothed polygram is preferable for relatively small samples.

Investigation of the quality of the PDF estimates

Samples with known PDFs were generated: Pareto PDFs ($f_p$) with EVI $\gamma = \{0.3; 1.2\}$, Burr PDFs ($f_b$) with parameters $\gamma = 1, \rho = -1$ and $\gamma = 4, \rho = -2$, $f(x) = \lambda \tau x^{-\gamma}(1 + x^\gamma)_{-\lambda}^{-1}$, where the EVI $\gamma$ and parameter $\rho$ are defined by $\gamma = 1/(\lambda \tau)$ and $\rho = -1/\lambda$, respectively. Mixed pairs of distributions Burr(1, -1)–Pareto(1.2) ($f_{pb}$) were also studied. Penalty functions satisfying (5.2) were used, namely $q_1(x) = (d_1 + x)^6$ and $q_2(x) = (d_2 + x)^6$, where $d_1, d_2 > 0$, and

<p>| Table 5.1 | Loss due to misclassification. |
|---|---|---|---|---|---|---|---|---|</p>
<table>
<thead>
<tr>
<th>PDF</th>
<th>$n$</th>
<th>$J_c(\eta_{EB}) \cdot 10^3$</th>
<th>$Pl$</th>
<th>$Plf$</th>
<th>$Ke$</th>
<th>$Kef$</th>
<th>$Pl/ Plf$</th>
<th>$Ke/ Kef$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_p$</td>
<td>50</td>
<td>131</td>
<td>62</td>
<td>57</td>
<td>69</td>
<td>2.11</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>54</td>
<td>51</td>
<td>41</td>
<td>52</td>
<td>1.06</td>
<td>0.79</td>
<td></td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>47</td>
<td>46</td>
<td>22</td>
<td>36</td>
<td>1.02</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>39</td>
<td>38</td>
<td>11</td>
<td>26</td>
<td>1.03</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>$f_f$</td>
<td>50</td>
<td>153</td>
<td>135</td>
<td>142</td>
<td>168</td>
<td>1.13</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>65</td>
<td>58</td>
<td>55</td>
<td>69</td>
<td>1.12</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>43</td>
<td>41</td>
<td>18</td>
<td>26</td>
<td>1.05</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>23</td>
<td>25</td>
<td>9</td>
<td>20</td>
<td>0.92</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>$f_{pf}^1$</td>
<td>50</td>
<td>347</td>
<td>485</td>
<td>72</td>
<td>77</td>
<td>0.72</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>321</td>
<td>467</td>
<td>31</td>
<td>43</td>
<td>0.69</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>257</td>
<td>410</td>
<td>8</td>
<td>13</td>
<td>0.63</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>215</td>
<td>379</td>
<td>6</td>
<td>12</td>
<td>0.57</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>$f_{pf}^2$</td>
<td>50</td>
<td>41</td>
<td>32</td>
<td>55</td>
<td>73</td>
<td>1.28</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>10</td>
<td>9</td>
<td>27</td>
<td>51</td>
<td>1.11</td>
<td>0.53</td>
<td></td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>9</td>
<td>15</td>
<td>9</td>
<td>25</td>
<td>0.6</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>4</td>
<td>7</td>
<td>5</td>
<td>21</td>
<td>0.57</td>
<td>0.24</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.2 $J_c(\eta_{EB}) \cdot 10^3$ over $[0, \infty)$ for estimates $Pl$ (solid line), $Plf$ (dotted line), $Ke$ (solid line with + marks), $Ke f$ (dot-dashed line): (left) Pareto(1)–Pareto(3) PDFs and (right) Fréchet(1)–Fréchet(2) PDFs. Reprinted from *Computational Statistics*, 19(4), pp. 569–592, Estimation of heavy-tailed probability density function with application to Web data, Maiboroda RE and Markovich NM, Figure 2, © 2004 Physica-Verlag, A Springer Company. With kind permission of Springer Science and Business Media.

Figure 5.3 $J_c(\eta_{EB}) \cdot 10^3$ for Pareto(2)–Fréchet(0.3) PDFs for estimates $Pl$ (solid line), $Plf$ (dotted line), $Ke$ (solid line with + marks), $Ke f$ (dot-dashed line): (left) over $[0, \infty)$ and (right) over $[6, \infty)$. Reprinted from *Computational Statistics*, 19(4), pp. 569–592, Estimation of heavy-tailed probability density function with application to Web data, Maiboroda RE and Markovich NM, Figure 3, © 2004 Physica-Verlag, A Springer Company. With kind permission of Springer Science and Business Media.
The modeling shows that:

- the adaptive transformation (4.11) leads to better estimation than the fixed transformation, though the EVI estimation is rough;
- the kernel estimate is worse than the polygram for relatively small samples, but better for large samples.
5.4 Application of the classification technique to Web data analysis

In this section, some examples of a potential application of the classification procedure to Internet data are provided (Maiboroda and Markovich, 2004, pp. 583–585).

5.4.1 Intelligent browser

When a user gets access to the Web, he generates one or several sessions. A session can consist of several http requests. An http request is generated each time the user clicks on a link. Sometimes, several http requests can be generated at the same time because the browser automatically loads images from a Web page.

Let us consider an ‘intelligent browser’ for a wireless environment such as a Universal Mobile Telecommunications System (UMTS) network. This can select which image to load depending on the typical behavior of the user. Specifically, suppose the browser first offers the user the information about the size of a picture. The user can ask the browser to show him a complete picture or decide not to look at this picture at all.

We assume that in order to maintain information about user behavior one can observe the work of the user for some fixed period of time. Then one maintains two data sets: the sizes of rejected pictures (i.e. the ones which the user did not want to open) and the sizes of accepted pictures (opened by the user after preliminary information from the browser).

Furthermore, the PDFs $f_1(x)$ and $f_2(x)$ of both samples can be estimated, for example, by the method described in Section 4.3. Using these PDF estimates one can construct an empirical Bayesian classifier $\eta_{EB}(x)$ to decide for the user whether the picture should be opened completely for browsing. This is a typical classification problem. The mean penalty of the misclassification for two classes is given by:

$$\mathcal{L}(\eta_{EB}) = p_1 \int_{0}^{\infty} q_1(x)f_1(x)1\{\eta_{EB}(x) = 2\}dx + p_2 \int_{0}^{\infty} q_2(x)f_2(x)1\{\eta_{EB}(x) = 1\}dx.$$

If the classifier has made a mistake and the browser opened the picture (i.e., $\eta_{EB}$ assigns the picture to the second class) which is not useful for the user (i.e., the picture is actually related to the first class), then the mean loss due to the browser is equal to $p_1 \int_{0}^{\infty} q_1(x)f_1(x)1\{\eta_{EB}(x) = 2\}dx$. Similarly, if the browser did not open a useful picture because the classifier assigned it to the first class then the mean loss due to the browser is determined by $p_2 \int_{0}^{\infty} q_2(x)f_2(x)1\{\eta_{EB}(x) = 1\}dx$. Here, $p_1$ and $p_2$ are the proportions of the pictures related to the first and second class in the common measurement. The penalty functions $q_1(x)$ and $q_2(x)$ could be defined as the financial losses of the network. Then $\mathcal{L}(\eta_{EB})$ reflects the quality of the classification.
5.4.2 Web data analysis by traffic classification

The http requests may be of different types: ordinary Web pages based on HTML descriptions, advanced Synchronized Multimedia Integration Language (SMIL) presentations with large images, and multimedia streams. We suppose that separate observations of all sources are available, for example, file sizes are measured. Then one can estimate the file size PDFs of the sources and provide the classifier \( \eta_{EB}(x) \). Furthermore, the classification of any new http request can be given and the best service of the request may be provided.

5.4.3 Web prefetching

Web prefetching aims to reduce the Web user’s perceived latency. First, the user’s preferences and accesses (e.g., favorite objects) are investigated. Then the prefetching engine lists the objects which should be prefetched. The prefetching engine is located in the web browser or in an intermediate web proxy server. As a result, the favorite objects are predownloaded and kept in cache. Mozilla Firefox is an example of a Web browser with the prefetching function (Padmanabhan and Mogul, 1996). Algorithms for predicting user preferences are given, for example, in Davison (2002) and Padmanabhan and Mogul (1996).

The idea of the ‘intelligent browser’ can be applied for such prediction. Suppose the user demands one object but the prefetching engine predicts another. Then the prefetching process is interrupted and the user’s request is satisfied. If the user demands the object from the prefetching list then the request is provided to the user with zero service time. The objects from the prefetching list can be considered as the first class, the other objects as the second class.

5.5 Exercises

1. Retransformed kernel estimates and classification.

   Generate two samples \( X^n_1 \) and \( X^n_2 \) of size \( n = 100 \) with known PDFs of a Pareto distribution (4.9) with \( \gamma_1 = 1 \) (\( f_1(x) \)) and \( \gamma_2 = 3 \) (\( f_2(x) \)), respectively.

   - Estimate \( f_1(x) \) and \( f_2(x) \) by the retransformed kernel estimator with Epanechnikov’s kernel (2.21). Use the adaptive transformation (4.11), where \( \gamma \) is estimated by Hill’s estimate (1.5). Apply formulas (3.26) and (4.24).

   - Take \( p_1 = p_2 = 0.5 \) as a priori proportions of the classes.

   - Take as smoothing parameter in (4.24) \( h = \sigma n^{-1/5} \), where \( \sigma \) is a standard deviation calculated from new samples \( T_{\hat{\gamma}_1}(X^n_1) \) and \( T_{\hat{\gamma}_2}(X^n_2) \).

   - Use the penalty functions \( q_1(x) = \sqrt{x + 10} \), \( q_2(x) = \sqrt{x} \).
• Mix samples $X^n_1$ and $X^n_2$ together and classify the points $X_i$ of the obtained sample by the classifier $X_i$ belongs to class \[
\begin{cases} 
1, & \text{if } p_1 q_1 (x) \hat{f}_1 (x) \geq p_2 q_2 (x) \hat{f}_2 (x), \\
2, & \text{if } p_1 q_1 (x) \hat{f}_1 (x) \leq p_2 q_2 (x) \hat{f}_2 (x). 
\end{cases}
\]

• Calculate the loss function 
\[ \varphi(x) = \begin{cases} 
0.5 q_1 (x) f_1 (x), & \text{if } q_1 (x) \hat{f}_1 (x) < q_2 (x) \hat{f}_2 (x), \\
0.5 q_2 (x) f_2 (x), & \text{otherwise}.
\end{cases} \]

• Calculate the risk of misclassification using the formula 
\[ R = \int_0^\infty \varphi(x) dx = \int_0^1 \varphi(u(x)) u'(x) dx, \]
where $u(x) = 1/(1-x)^2 - 1$ may be taken for simplicity of calculation.

2. Carry out the classification in Exercise 1 for the penalty functions $q_1 (x) = (d_1 + x)^\delta$ and $q_2 (x) = (d_2 + x)^\delta$, where $d_1, d_2 > 0$, $\delta = 0.5 \min\{1/\hat{\gamma}_1, 1/\hat{\gamma}_2\}$, $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are estimates of $\gamma_1$ and $\gamma_2$. Select the constants $d_1$ and $d_2$ from the switching condition of the Bayesian rule (see Figure 5.1) at the $(1-\alpha)$-quantile of the first population $x^* = F_1^{-1}(1-\alpha)$ ($F_1 (x)$ is the DF of the first population), that is, 
\[ (d_1 + x^*)^\delta f_1 (x^*) = (d_2 + x^*)^\delta f_2 (x^*), \]
where $x^* = (\alpha^{-\hat{\gamma}_1} - 1)/\hat{\gamma}_1$, and 
\[ d_1 = (d_2 + x^*) (f_2 (x^*)/f_1 (x^*))^{1/\delta} - x^*, \quad d_2 > x^* \left( (f_2 (x^*)/f_1 (x^*))^{-1/\delta} - 1 \right). \]
Select $\alpha \in \{0.1, 0.25\}$. Compare the risks of misclassification with Exercise 1.
Estimation of high quantiles

This chapter discusses estimators of the high quantiles for heavy-tailed distributions. The relative bias and the mean squared errors as well as confidence intervals of estimates are compared in 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 a modification of Weissman’s estimate. An application to WWW traffic data is considered.

6.1 Introduction

Suppose we have a sequence of i.i.d. observations \( X^n = \{X_1, X_2, \ldots, X_n\} \) from an unknown DF \( F(x) \).

**Definition 15** For the continuous DF \( F(x) \) the quantile \( x = x_p \) of level \( 1 - p \), \( p \in (0, 1) \), is the solution of the equation

\[
1 - F(x) = p.
\]

Our aim is to evaluate a high quantile, that is, a quantile corresponding to a probability \( p \) close to zero when \( F(x) \) is heavy-tailed.

In practice it is often necessary to evaluate the risk of large but possibly rare losses. The analysis of such risks is important in indicating the thresholds of
parameters in complex multi-component systems such as economic and ecological systems, atomic power stations, the Internet etc.

High quantiles are usually located at the boundary or beyond the range of the sample. The classical approach of using the empirical DF \( F_n(x) \) in (6.1) as well as weighted quantile estimators\(^1\) is not valid for high quantiles since \( F_n(x) = 1 \) holds for \( x \geq X_{(n)} \). Here \( X_{(n)} \) denotes the largest order statistic corresponding to \( X^n \).

The lack of information beyond the range of the sample creates the main problem in the estimation of high quantiles. Since \( F_n(X_{(n)}) = 1 \), for \( p < 1/n \) it is impossible to estimate the quantiles without knowledge of the behavior of \( F \) at infinity. The main idea behind all estimators for high quantiles is to select first some auxiliary pilot estimate of a quantile inside the range of the sample (one can use one of the order statistics close to the boundary as a pilot estimate) and to move this pilot estimate to the right. For this purpose, special scaling parameters are estimated.\(^2\)

Obviously, in order to extrapolate the pilot quantile beyond the sample range, one needs to use some model of the tail of the distribution. Such models are not available in many applications. Therefore, the asymptotic tail models (1.1)–(1.3) based on the distribution of the largest order statistic are usually used.

### 6.2 Estimators of high quantiles

To satisfy (1.1) it is necessary and sufficient, by Proposition 3.3.2 in Embrechts et al. (1997), that

\[
\lim_{n \to \infty} nF(b_n + a_n x) = -\ln H_\gamma(x), \quad x \in \mathbb{R}, \quad a_n > 0, \quad b_n \in \mathbb{R}.
\]  

(6.2)

It is evident from (1.3) and for \( \gamma \neq 0 \), that

\[
\lim_{t \to \infty} t \left( 1 - F(a(t) x + b(t)) \right) = (1 + \gamma x)^{-1/\gamma}
\]

and

\[
1 - F(u) \approx \frac{1}{t} \left( 1 + \gamma \frac{u - b(t)}{a(t)} \right)^{-1/\gamma}.
\]

For the \((1-p)th\) quantile the approximation

\[
\hat{x}_p = \hat{b}(n/k) + \hat{a}(n/k) \frac{(k/(pn))^{\hat{\gamma}} - 1}{\hat{\gamma}}
\]

may be used (Dekkers et al., 1989).\(^3\) Here, \( \hat{\gamma} \) is an estimate of \( \gamma \).

---

\(^1\) Let \( X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)} \) be the order statistics of the sample. Usually \( X_{(k)} \), with \( k = \lfloor np \rfloor + 1 \) (\( \lfloor a \rfloor \) denotes the integer part of \( a \)), is used as an estimate of the \( p \)th quantile \( x_p \). It leads to the estimates that are interpolations of order statistics, e.g., to a weighted average \( x_p = (1 - g)X_{(j)} + gX_{(j+1)} \), where \( j = \lfloor np \rfloor \) and \( g = np - j \) (Dielman et al., 1994).

\(^2\) The same principle can be applied to the estimation of the small quantiles when \( p \) is close to one.

\(^3\) Asymptotic normality of \( \hat{x}_p^d \) has been proved for the specific functions \( \hat{a}(n/k) \) and \( \hat{b}(n/k) \); see de Haan and Rootzén (1993).
The GPD (1.16) and the Pareto-type tail model (Hall, 1990; Hall and Weissman, 1997)

\[ 1 - F(x) = cx^{-1/\gamma} (1 + dx^{-\beta} + o(x^{-\beta})) , \]

where \( \gamma > 0, \beta > 0, c > 0, -\infty < d < \infty, \) are often used to model the tail of the distribution \( F(x) \). Different estimators of high quantiles (e.g., Dekkers et al., 1989; Beirlant et al., 1999; Weissman, 1978) follow from these assumptions on the tail. The form of the tail can be detected by rough methods of heavy-tailed data analysis such as the QQ plot (Embrechts et al., 1997) or by nonparametric tests (Jurečková and Picek, 2001). We shall consider some well-known estimators.

In the POT estimator the GPD is used as a distribution of excesses over some high threshold \( u \):

\[ x^\text{POT}_p = u + \hat{\sigma} \left( \frac{p}{1 - F_n(u)} \right)^{-\hat{\gamma}} - 1 , \]

where \( \hat{\sigma} \) and \( \hat{\gamma} \) are estimates of parameters of the GPD, and \( F_n(u) \) is the empirical DF evaluated at \( u \) (McNeil and Saladin, 1997; McNeil et al., 2005).

In Weissman (1978) the estimator

\[ x^w_p = X_{(n-k)} \left( \frac{k+1}{n+1} \right)^{\hat{\gamma}} , \quad k = 1, \ldots, n-1 , \]

is obtained from (6.2) for the Pareto tail model, that is, for the first type of tail in (1.2), where \( X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)} \) are order statistics corresponding to the sample \( X^n \).

In Markovitch and Krieger (2002a) an estimator was proposed that differs slightly from \( x^w_p \). To obtain this estimator the idea of combining the estimator (3.4) with (3.1) for a heavy-tailed PDF was applied. This implies that the PDF is approximated by the Pareto-type model

\[ f_\gamma(x) = \frac{1}{\gamma} x^{-1/\gamma-1} + \frac{2}{\gamma} x^{-2/\gamma-1} \]

as \( x \geq X_{(n-k)} \), and by some nonparametric estimator \( f^N(t) \) for \( x < X_{(n-k)} \) such that (3.5) holds. We note that for \( x_p > X_{(n-k)} \),

\[ p = P\{x > x_p\} = P\{x > x_p|x > X_{(n-k)}\} P\{x > X_{(n-k)}\} . \]

\( P\{x > X_{(n-k)}\} \) may be replaced by the number of excesses over the threshold \( X_{(n-k)} \), namely, \( k/n \). From (3.1) and (3.4) the estimate

\[ \tilde{F}(x) = \frac{1}{c(\gamma)} \left( \int_0^{\min(x,X_{(n-k)})} f^N(t)1(f^N(t) > 0)dt + \int_{\min(x,X_{(n-k)})}^x f_\gamma(t)dt \right) \]

of the DF \( F(x) \) follows, where

\[ c(\gamma) = \int_0^\infty \tilde{f}(t, \gamma, N)1(\tilde{f}(t, \gamma, N) > 0)dt \]
may be approximated by \( \int_{0}^{\infty} \tilde{f}(t, \gamma, N) dt \) (see (3.6)):

\[
c(\gamma) \approx 1 + X_{(n-k)}^{-1/\gamma} + X_{(n-k)}^{-2/\gamma}.
\]  

(6.7)

Since

\[
\int_{0}^{X_{(n-k)}} f^{N}(t)\mathbf{1}(f^{N}(t) > 0) dt = c(\gamma) - \int_{X_{(n-k)}}^{\infty} f_{\gamma}(t) dt,
\]

we have for \( x > X_{(n-k)} \) by (6.6) that

\[
\tilde{F}(x) = \frac{1}{c(\gamma)} \left( \int_{0}^{X_{(n-k)}} f^{N}(t)\mathbf{1}(f^{N}(t) > 0) dt + \int_{X_{(n-k)}}^{x} f_{\gamma}(t) dt \right) = 1 - \frac{1}{c(\gamma)}(x^{-1/\gamma} + x^{-2/\gamma}).
\]

Let

\[
1 - F_{\gamma}(x) = x^{-1/\gamma} + x^{-2/\gamma}.
\]

Then we can use

\[
P \{ x > x_{p} | x > X_{(n-k)} \} = \frac{1 - F(x_{p})}{1 - F(X_{(n-k)})} \approx \frac{1 - \tilde{F}(x_{p})}{1 - F_{\gamma}(X_{(n-k)})} < \frac{1}{c(\gamma)} \left( \frac{X_{(n-k)}}{x_{p}} \right)^{1/\gamma} \left( 1 + \left( \frac{X_{(n-k)}}{x_{p}} \right)^{1/\gamma} \right).
\]

Hence, since

\[
p \approx \frac{k}{nc(\gamma)} \left( \left( \frac{X_{(n-k)}}{x_{p}} \right)^{1/\gamma} + \left( \frac{X_{(n-k)}}{x_{p}} \right)^{2/\gamma} \right)
\]

holds approximately, one can expect that the statistic

\[
x_{p}^{c} = X_{(n-k)} \left( -0.5 + \sqrt{0.25 + \frac{pnc(\hat{\gamma})}{k}} \right)^{-\hat{\gamma}}
\]  

(6.8)

approximates \( x_{p} \). Here, \( \hat{\gamma} \) is an estimate of the EVI \( \gamma \) that determines the shape of the tail. The estimate \( x_{p}^{c} \) differs from Weissman’s estimate \( x_{p}^{w} \) in the normalizing multiplier, reflecting the fact that the estimate \( \tilde{F}(x) \) of the DF \( F(x) \) includes not only the parametric estimate of the tail domain as in \( x_{p}^{w} \), but also the “body” estimate. Since it is assumed that \( \int_{0}^{X_{(n-k)}} f^{N}(t) dt = 1 \), we are able to use any parametric or nonparametric estimate of the PDF “body” with such a property.

The common disadvantage of the high quantile estimators is their sensitivity to the choice of threshold. This may be the value of \( u \) in the estimator \( x_{POT}^{c} \) or the number of order statistics \( k \) in \( x_{p}^{c} \) and \( x_{p}^{w} \). An estimate of \( k \) is also required to estimate the EVI \( \gamma \). One can apply Hill’s estimator (1.5) or other estimators such
as the moment estimator, the ratio estimator, or the UH estimator that are valid not only for positive $\gamma$ (see Section 1.2).

When $k$ increases (i.e., $X_{(n-k)}$ decreases), the variance of the EVI estimate decreases but the bias increases. However, when $k$ decreases (i.e., fewer data are used) the bias tends to 0 but the variance increases.

Theoretically, an optimal value of $k$ has to minimize the MSE $E\left(\hat{x}_p(k) - x_p\right)^2$. An exact expression for MSE is not available since $x_p$ is unknown. Thus, finding $k$ is usually a matter of minimizing the asymptotic MSE, that is, the asymptotic expectation $as.E\left(\hat{x}_p(k) - x_p\right)^2$ or more precisely its bootstrap estimator (Ferreira et al., 2000) or $as.E\left(\log\left(\hat{x}_p(k)/x_p\right)\right)^2$ (Beirlant et al., 1999).

In Matthys et al. (2004) it is the estimate of the asymptotic MSE of $x^w_p$ given in (6.10) with incorporated ML estimates of the unknown parameters $\gamma$, $b_{n,k}$, $\rho$ of the distribution that is minimized to find $k$. The ML method is used in the framework of an exponential regression model for log-spacings of order statistics. The latter method is extended to censored data.

Hall and Weissman (1997) proposed to minimize the MSE $E(\tilde{F}_\theta(\tilde{F}^{-1}(p)) - p)^2$, where $\tilde{F}_\theta$ is some tail estimate.

A simulation study with heavy-tailed distributions has shown that the quantile estimate $x^c_p$ is better than $x^{POT}_p$ and $x^w_p$ for the highest quantiles, and has demonstrated smaller mean squared errors (Markovich and Krieger, 2002a). The estimate $x^{POT}_p$ is essentially less accurate due to the estimation of both GPD parameters by the ML method in addition to the threshold $u$. It is shown in Matthys and Beirlant (2001) that the GPD shape parameter $\gamma$ and the high quantile estimates are sensitive to the parameter computation method – the ML method (Smith, 1987), the PWM method (Hosking and Wallis, 1987) and the EPM (Castillo et al., 2006).

Formally, the threshold $u$ is not random. One estimates the parameters of the GPD for a fixed $u$. Nevertheless, one can select some quantile of the unknown distribution as a threshold and replace it by an empirical quantile that is random (McNeil and Saladin, 1997). The alternative is to select one of the sample points $X_{(n-k)}$ (Beirlant et al., 2004, p. 149).

### 6.3 Distribution of high quantile estimates

Here, we consider the distributions of $x^c_p$ (6.8) and $x^w_p$ (6.5). It is evident that

$$\frac{x^c_p}{x_p} = \frac{X_{(n-k)}}{x_p \left(-0.5 + \sqrt{0.25 + \frac{\text{poc}(\gamma)}{k}}\right)^\gamma}$$

---

and

$$\log \left( \frac{x_p^c}{x_p} \right) = \log X_{(n-k)} - \log x_p - \hat{\gamma} \log \left( -0.5 + \sqrt{0.25 + \frac{pnc(\hat{\gamma})}{k}} \right)$$

$$\simeq \log X_{(n-k)} - \log x_p - \hat{\gamma} \log \left( \frac{c(\hat{\gamma})}{a_n} \right),$$

where $a_n = k/(pn)$. Note that

$$\log \left( \frac{x_p^c}{x_p} \right) \simeq \log \left( \frac{x_p^w}{x_p} \right) - \hat{\gamma} \log c(\hat{\gamma}). \quad (6.9)$$

It is proved in Markovich (2005b) that the distributions of the logarithms of ratios $x_p^c$ and $x_p^w$ to the true value of the quantile $x_p$ are asymptotically normal.

Let us use Hill’s estimate (1.5) as $\hat{\gamma}$. Following Dekkers and de Haan (1989), to derive the asymptotics we require that $k/(pn)$ have positive limit as $n \to \infty$.

**Theorem 14 (Markovich, 2005b)** Let the tail distribution be of Pareto type (6.3) and $k, p, \gamma \to \infty, k_n \to 0, p = p_n \sim c^* \cdot k_n \to 0, c^* > 0$, as $n \to \infty$. Then

$$\frac{\log(x_p^w/x_p) - a}{\delta_w} \to^d N(0, 1),$$

$$\frac{\log(x_p^c/x_p) - (a + ((k+1)/n)\gamma)}{\delta_c} \to^d N(0, 1),$$

where

$$a = \gamma dc^{-\gamma\beta} \left( \left( \frac{k+1}{n} \right)^\gamma - p^{\gamma\beta} \right),$$

$$\delta_w^2 = \frac{\gamma^2}{k} \left( 1 - \gamma\beta dc^{-\gamma\beta} \left( \frac{k+1}{n} \right)^\gamma \right)^2 + \left( \log \left( \frac{k}{np} \right) \right)^2,$$

$$\delta_c^2 = \delta_w^2 + \frac{\gamma^2}{k} \left( \frac{k+1}{n} \right)^\gamma \left( \left( \frac{k+1}{n} \right)^\gamma - 2 \left( 1 - \gamma\beta dc^{-\gamma\beta} \left( \frac{k+1}{n} \right)^\gamma \right) \right).$$

The proof of the theorem is given in Appendix D.

Theorem 14 shows that the expectation of the distribution of $\log(x_p^c/x_p)$ is larger than that of $\log(x_p^w/x_p)$, while the variance is less. The difference becomes negligible as the sample size increases.

Asymptotic normality of the estimate $\log x_p^w$ is also given in Matthys and Beirlant (2003) for the class of heavy-tailed distributions with regularly varying tails, that is, $1 - F(x) = x^{-1/\gamma} \ell(x)$, where $\ell(x)$ is a slowly varying function. The asymptotic MSE of $\log \left( x_p^w/x_p \right)$,
\[
\text{as}.E\left( \log \left( \frac{x^w_p}{x_p} \right) \right)^2 = \frac{\gamma^2}{k+1} \left[ 1 + \left( \log \frac{k+1}{(n+1)p} \right)^2 \right] + \frac{b^2_{n,k}}{1 - \rho} \left[ \frac{1}{\rho} \left( 1 - \left( \frac{k+1}{p(n+1)} \right)^\rho \right) + \frac{1}{1 - \rho} \log \frac{k+1}{p(n+1)} \right]^2,
\]

(6.10)

where

\[
b_{n,k} = b \left( (n+1)/(k+1) \right),
\]

(6.11)

is obtained under a specific assumption on \( \ell(x) \) denoted by \( (R_t) \): There exists a real constant \( \rho \leq 0 \) and a rate function \( b \) satisfying \( b(x) \to 0 \) as \( x \to \infty \), such that for all \( \lambda \geq 1 \), as \( x \to \infty \), \( \log(\ell(\lambda x)/\ell(x)) \sim b(x)k^\rho(\lambda) \), with \( k^\rho(\lambda) = (\lambda^\rho - 1)/\rho \), which is to be read as \( \log \lambda \) if \( \rho = 0 \).

The distribution (6.3) satisfies assumption \( (R_t) \) with

\[
b(x) = \rho \gamma dc^\rho x^\rho [1 + o(1)],
\]

(6.12)

where \( \rho = -\gamma \beta \). Then the result of Theorem 14 can be rewritten in this notation as

\[
\text{as}.E\left( \log \left( \frac{x^w_p}{x_p} \right) \right)^2 = \frac{\gamma^2}{k} \left[ 1 + \left( \log \frac{k}{np} \right)^2 \right] + \frac{b^2_{n,k}}{1 - \rho} \left[ \frac{1}{\rho^2} \left( 1 + \left( \frac{k+1}{pn} \right)^\rho \right)^2 \frac{1}{k} \left( 1 + \frac{2}{\rho d} \left( \frac{k+1}{cn} \right)^\rho \right) \right].
\]

(6.13)

This is similar to (6.10), apart of the second term in the final brackets. It follows from (6.10)–(6.12) that

\[
b^2_{n,k} \frac{1}{1 - \rho} \log \frac{k+1}{p(n+1)} \sim - \left( \frac{cn}{k} \right)^{2\rho} \log c^\ast.
\]

(6.14)

We derive from (6.13) that

\[
b^2_{n,k} \frac{1}{k} \left( 1 + \frac{2}{\rho d} \left( \frac{k+1}{cn} \right)^\rho \right) \sim \left( \frac{cn}{k} \right)^{2\rho} \frac{1}{k}.
\]

This latter term goes to zero faster than (6.14).

### 6.4 Simulation study

#### 6.4.1 Comparison of high quantile estimates in terms of relative bias and mean squared error

For the comparison of \( x^\text{POT}_p \), \( x^w_p \) and \( x^c_p \) we use some of the distributions applied in McNeil and Saladin (1997), namely, the Pareto distribution with \( 1/\gamma \in \{1, 2\} \) and the Student distribution with \( 1/\gamma \in \{1, 2\} \). Following McNeil and Saladin (1997),
we use as characteristics of the quantile estimates $\hat{x}_p \in \{x_{p}^{\text{POT}}, x_{p}^{w}, x_{p}^{c}\}$ the empirical estimates of the bias and the mean squared error of the estimates expressed as proportions of the true value $x_p$:

$$\%\text{Bias} = \frac{1}{N_R} \sum_{i=1}^{N_R} \hat{x}_{p_i} - x_p$$

$$\%\text{RMSE} = \sqrt{\frac{1}{N_R} \sum_{i=1}^{N_R} (\hat{x}_{p_i} - x_p)^2}.$$ 

Here $N_R = 25$ is the number of repetitions in our Monte Carlo study. The parameters $\sigma$ and $\gamma$ of $x_{p}^{\text{POT}}$ were calculated by the ML method and the threshold $u$ as an empirical quantile of the underlying distribution. The EVI $\gamma$ of $x_{p}^{c}$ and $x_{p}^{w}$ was calculated by Hill’s estimate (1.5) and $k$ from the minimum of the bootstrap estimate of $E \left( \hat{x}_p(k) - x_p \right)^2$.

$$\text{MSE}(n_1, k_1) = E \left\{ \left( \hat{x}_p^{*}(n_1, k_1) - \hat{x}_p(n, k) \right)^2 | X^n \right\}$$

with respect to $k_1$. Here, $\hat{x}_p^{*}(n_1, k_1)$ is the estimate of the quantile calculated from the resample of size $n_1$ with parameter $k_1$. Such a resample is drawn randomly from the sample $X^n$ with replacement. For the bootstrap, 25 resamples were used. The values of the auxiliary parameters $\alpha$ and $\beta$ for the calculation of the size of the resamples $n_1 = n^\beta$ and relation $k = k_1 \left( \frac{n}{n_1} \right)^\alpha$ between $k$ and $k_1$ were taken equal to $\alpha = 2/3$, $\beta = 1/2$, similar to Section 1.2.4.

The results of the simulation are shown in Table 6.1. The results for $x_{p}^{\text{POT}}$ are compiled from McNeil and Saladin (1997). The simulation study illustrates that the quantile estimate $x_{p}^{c}$ is better than $x_{p}^{\text{POT}}$ and $x_{p}^{w}$ especially for the highest quantiles, demonstrating the smaller mean squared error results. This conclusion is in agreement with that which follows from Theorem 14.

### 6.4.2 Comparison of high quantile estimates in terms of confidence intervals

Theorem 14 does not allow us to construct asymptotic confidence intervals, since $\alpha, \delta_w$ and $\delta_c$ depend on the unknown parameters of the distribution. Here, we describe the nonasymptotic bootstrap confidence intervals considered in Markovich (2005b). One can find more about bootstrap confidence sets in Shao and Tu (1995, Chapter 4).

It follows from Theorem 14 that the logarithms of both estimates $x_{p}^{w}$ and $x_{p}^{c}$ have asymptotically normal distributions. However, in order to get better confidence intervals for finite samples one has to assume that the estimates of quantile $\hat{x}_p$ constructed over the set of the samples are normally distributed. The mean and variance of the normal distribution are constructed by these estimates $\hat{x}_1^{N_R}, \ldots, \hat{x}_{N_R}^{N_R}$.
Table 6.1 Simulation results of quantile estimation.

<table>
<thead>
<tr>
<th>Quantile estimate</th>
<th>Sample size</th>
<th>%Bias</th>
<th>%RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 − ( p = 0.99 )</td>
<td>1 − ( p = 0.999 )</td>
</tr>
<tr>
<td>( x_{POT} )</td>
<td>250</td>
<td>2.5</td>
<td>23.60</td>
</tr>
<tr>
<td>( x_p )</td>
<td>-9.7</td>
<td>-22.9</td>
<td>18.0</td>
</tr>
<tr>
<td>( x^w )</td>
<td>-9.3</td>
<td>-6.6</td>
<td>24.7</td>
</tr>
<tr>
<td>( x_{POT} )</td>
<td>500</td>
<td>2.58</td>
<td>19.16</td>
</tr>
<tr>
<td>( x_p )</td>
<td>-9.7</td>
<td>-9.6</td>
<td>15.7</td>
</tr>
<tr>
<td>( x^w )</td>
<td>-7.3</td>
<td>10.4</td>
<td>19.4</td>
</tr>
</tbody>
</table>

---

Pareto distribution (1/\( \gamma = 2 \))

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_{POT} )</td>
<td>250</td>
<td>10.52</td>
<td>207.49</td>
<td>83.28</td>
<td>1611.14</td>
</tr>
<tr>
<td>( x_p )</td>
<td>-27.5</td>
<td>-25.5</td>
<td>35.5</td>
<td>58.2</td>
<td></td>
</tr>
<tr>
<td>( x^w )</td>
<td>8.4</td>
<td>16.5</td>
<td>45.5</td>
<td>93.9</td>
<td></td>
</tr>
<tr>
<td>( x_{POT} )</td>
<td>500</td>
<td>7.69</td>
<td>51.70</td>
<td>50.25</td>
<td>254.02</td>
</tr>
<tr>
<td>( x_p )</td>
<td>-19.8</td>
<td>-25.0</td>
<td>32.9</td>
<td>44.7</td>
<td></td>
</tr>
<tr>
<td>( x^w )</td>
<td>12.0</td>
<td>12.1</td>
<td>40.5</td>
<td>65.5</td>
<td></td>
</tr>
</tbody>
</table>

---

Standard lognormal distribution

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_p )</td>
<td>250</td>
<td>0.62</td>
<td>3.09</td>
<td>18.84</td>
<td>54.28</td>
</tr>
<tr>
<td>( x^w )</td>
<td>1.4</td>
<td>29.545</td>
<td>2.301</td>
<td>34.152</td>
<td></td>
</tr>
<tr>
<td>( x_{POT} )</td>
<td>500</td>
<td>1.267</td>
<td>35.472</td>
<td>2.935</td>
<td>40.896</td>
</tr>
<tr>
<td>( x_p )</td>
<td>-1.15</td>
<td>1.85</td>
<td>12.95</td>
<td>39.26</td>
<td></td>
</tr>
<tr>
<td>( x^w )</td>
<td>1.033</td>
<td>23.119</td>
<td>1.895</td>
<td>25.645</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( x_{POT} )</td>
<td>250</td>
<td>1.12</td>
<td>4.61</td>
<td>25.17</td>
<td>81.24</td>
</tr>
<tr>
<td>( x_p )</td>
<td>-15.2</td>
<td>32</td>
<td>37.9</td>
<td>69.8</td>
<td></td>
</tr>
<tr>
<td>( x^w )</td>
<td>34.1</td>
<td>122</td>
<td>52.4</td>
<td>157.9</td>
<td></td>
</tr>
<tr>
<td>( x_{POT} )</td>
<td>500</td>
<td>0.95</td>
<td>3.22</td>
<td>20.72</td>
<td>65.16</td>
</tr>
<tr>
<td>( x_p )</td>
<td>-22.2</td>
<td>2.7</td>
<td>25.7</td>
<td>32.5</td>
<td></td>
</tr>
<tr>
<td>( x^w )</td>
<td>17.7</td>
<td>62.8</td>
<td>30.4</td>
<td>86.3</td>
<td></td>
</tr>
</tbody>
</table>

---

Student distribution 1/\( \gamma = 2 \)

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_p )</td>
<td>250</td>
<td>14.75</td>
<td>131.52</td>
<td>69.82</td>
<td>495.95</td>
</tr>
<tr>
<td>( x^w )</td>
<td>-17.4</td>
<td>-28.9</td>
<td>40.4</td>
<td>57.7</td>
<td></td>
</tr>
<tr>
<td>( x_{POT} )</td>
<td>500</td>
<td>3.98</td>
<td>31.86</td>
<td>40.16</td>
<td>163.80</td>
</tr>
<tr>
<td>( x_p )</td>
<td>-27.7</td>
<td>-31.7</td>
<td>36.6</td>
<td>42.1</td>
<td></td>
</tr>
<tr>
<td>( x^w )</td>
<td>20.9</td>
<td>31.0</td>
<td>47.7</td>
<td>72.7</td>
<td></td>
</tr>
</tbody>
</table>

---

Student distribution 1/\( \gamma = 1 \) (Cauchy)

where $N_R$ is the number of samples. Then one can calculate the tolerance limits of the confidence intervals by the well-known formula

$$
(\text{mean}(\hat{x}_p) - \rho \cdot \text{StDev}(\hat{x}_p); \text{mean}(\hat{x}_p) + \rho \cdot \text{StDev}(\hat{x}_p)),
$$

where mean($\hat{x}_p$) and StDev($\hat{x}_p$) are the empirical mean and standard deviation of the $N_R$ estimates (Smirnov and Dunin-Barkovsky, 1965). As before (pp. 24, 25), such an interval is constructed in such a way that $100 \cdot (1 - p)\%$ part of the distribution falls into it with probability $P$. The value $\rho$ is calculated by (1.34)–(1.36). We have $\rho_{\infty} = 1.645$ for $p = 0.1$, which corresponds to a 90% confidence interval. Then $\rho = 1.776$ holds when $N_R = 500$.

For both estimates $x_p^w$ and $x_p^c$, Hill’s estimator (1.5) is used to estimate the EVI $\gamma$. The number of largest order statistics $k$ for the latter estimate and for the order statistic $X_{(n-k)}$ in (6.5) and (6.8) is obtained from the minimum of the bootstrap estimate of the mean squared error of the quantile estimation, that is, $E(\hat{x}_p(k) - x_p)^2 = \text{bias}(\hat{x}_p)^2 + \text{variance}(\hat{x}_p)$. To construct this bootstrap estimate we use resamples of smaller sizes $n_1$ than $n$ (see the bootstrap method in Section 1.2.2 for details).

As values of the auxiliary parameters $\alpha$ and $\beta$ for the calculation of the size of a resample $n_1 = n^\beta$ and the relation $k = k_1(n/n_1)^\alpha$ among $k$ and $k_1$, $\alpha = 2/3$ and $\beta = 1/2$ are selected, as in Markovitch and Krieger (2002a).

Then one can find the minimum of the estimate of the MSE,

$$
\text{MSE}(n_1, k_1) = E\{((\hat{x}_p(n_1, k_1) - \hat{x}_p(n, k))^2 | X^n) = (b^*(n_1, k_1))^2 + \text{var}^*(n_1, k_1),
$$

with respect to $k_1$, where $b^*(n_1, k_1)$ and $\text{var}^*(n_1, k_1)$ are the bootstrap estimates of the bias and variance, $\hat{x}_p^*(n_1, k_1)$ is the quantile estimate with parameter $k_1$, constructed from the resample $X_{n_1}^n$ of the size $n_1$ that is less than $n$. Such resamples are drawn randomly from the sample $X^n$ of the size $n$ with replacement.

Since the DF $F(x)$ is unknown, one has to use instead of $b^*(n_1, k_1)$ and $\text{var}^*(n_1, k_1)$ their empirical estimates:

$$
b^*(n_1, k_1) = \frac{1}{B} \sum_{b=1}^B \hat{x}_p^*(n_1, k_1) - \hat{x}_p(n, k),
$$

$$
\text{var}^*(n_1, k_1) = \frac{1}{B - 1} \sum_{b=1}^B \left( \hat{x}_p^*(n_1, k_1) - \frac{1}{B} \sum_{b=1}^B \hat{x}_p^*(n_1, k_1) \right)^2.
$$

The means, standard deviations, mean squared errors and 90% confidence intervals of the estimates $x_p^w$ and $x_p^c$ are given in Tables 6.2 and 6.3 for different heavy-tailed distributions and $N_R = 500$. A Pareto distribution with DF $F(x) = 1 - x^{-1/\gamma}$, $x > 0$, and parameter $\gamma \in \{1/2, 1\}$ and Weibull distribution with

---

5 In the expression for $\text{MSE}(n_1, k_1)$ the sample $X^n$ is fixed and the expectation is calculated under all theoretically possible resamples $X_{n_1}^n$. 

---
DF \( F(x) = 1 - sx^{s-1} \exp(-x^s) \) and \( s = 1/\gamma = 0.5 \) were investigated. Sample sizes \( n \in \{100, 1000\} \) were taken.

The true values \( x_p^{\gamma} \) of quantiles of level \( 1 - p \) are given in Table 6.4. From Tables 6.2 and 6.3 one can conclude that

- the bias of the estimate \( x_p^w \) is less than that of \( x_p^c \), but the variance is larger for \( x_p^w \);
- the MSE of the estimate \( x_p^c \) tends to be less than that of \( x_p^w \), especially for smaller sample sizes;
- the confidence intervals of \( x_p^w \) are wider than those of \( x_p^c \);
- the means of both estimates are far away from the true value of a 99.9% quantile for a Weibull distribution; however, the confidence interval is better for \( x_p^c \), especially for smaller samples.

### Table 6.2

Tolerant 90% confidence intervals of estimates \( x_p^w \) and \( x_p^c \) for heavy-tailed distributions: 500 samples of \( n = 100 \) observations each.

<table>
<thead>
<tr>
<th>PDF</th>
<th>((1 - p) \cdot 100%)</th>
<th>(x_p^c)</th>
<th>Confidence interval</th>
<th>(x_p^w)</th>
<th>Confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(\text{mean}(x_p^c))</td>
<td>(StDev(x_p^c))</td>
<td>(\text{mean}(x_p^w))</td>
<td>(StDev(x_p^w))</td>
</tr>
<tr>
<td>Pareto</td>
<td>99</td>
<td>75.814</td>
<td>(46.949)</td>
<td>117.957</td>
<td>(90.903)</td>
</tr>
<tr>
<td>(\gamma = 1)</td>
<td></td>
<td>(−7.567, 159.195)</td>
<td>MSE = 2.789 \cdot 10^3</td>
<td>(−43.487, 279.401)</td>
<td>MSE = 8.586 \cdot 10^3</td>
</tr>
<tr>
<td></td>
<td>99.9</td>
<td>963.094</td>
<td>(1.553 \cdot 10^3)</td>
<td>1.616 \cdot 10^3</td>
<td>(2.661 \cdot 10^3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(−1.795 \cdot 10^3)</td>
<td>MSE = 2.413 \cdot 10^6</td>
<td>(−3.11 \cdot 10^3)</td>
<td>MSE = 7.460 \cdot 10^6</td>
</tr>
<tr>
<td>Pareto</td>
<td>99</td>
<td>8.259</td>
<td>(2.116)</td>
<td>10.132</td>
<td>(3.107)</td>
</tr>
<tr>
<td>(\gamma = 1/2)</td>
<td></td>
<td>(4.501, 12.017)</td>
<td>MSE = 7.509</td>
<td>(4.614, 15.65)</td>
<td>MSE = 9.671</td>
</tr>
<tr>
<td></td>
<td>99.9</td>
<td>26.562</td>
<td>(13.045)</td>
<td>34.002</td>
<td>(21.559)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(3.394, 49.73)</td>
<td>MSE = 195.786</td>
<td>(4.287, 72.291)</td>
<td>MSE = 470.450</td>
</tr>
<tr>
<td>Weibull</td>
<td>99</td>
<td>25.398</td>
<td>(16.759)</td>
<td>38.719</td>
<td>(9.372)</td>
</tr>
<tr>
<td>(\gamma = 2)</td>
<td></td>
<td>(−4.366, 55.162)</td>
<td>MSE = 298.420</td>
<td>(−31.206, 108.644)</td>
<td>MSE = 1.857 \cdot 10^3</td>
</tr>
<tr>
<td></td>
<td>99.9</td>
<td>182.084</td>
<td>(274.729)</td>
<td>487.721</td>
<td>(1.957 \cdot 10^3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(−305.835, 670.003)</td>
<td>MSE = 9.353 \cdot 10^4</td>
<td>(−2.988 \cdot 10^3)</td>
<td>MSE = 4.023 \cdot 10^6</td>
</tr>
</tbody>
</table>

Table 6.3 Tolerant 90% confidence intervals of estimates $x_p^c$ and $x_p^w$ for heavy-tailed distributions: 500 samples of $n = 1000$ observations each.

<table>
<thead>
<tr>
<th>PDF</th>
<th>$(1 - p) \cdot 100%$</th>
<th>$x_p^c$ mean($x_p^c$) (StDev($x_p^c$))</th>
<th>Confidence interval</th>
<th>$x_p^w$ mean($x_p^w$) (StDev($x_p^w$))</th>
<th>Confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pareto, $\gamma = 1$</td>
<td>99</td>
<td>80.452 (16.272)</td>
<td>(51.533, 109.351)</td>
<td>101.93 (22.841)</td>
<td>(61.364, 142.496)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MSE = 646.902</td>
<td></td>
<td>MSE = 525.436</td>
<td></td>
</tr>
<tr>
<td></td>
<td>99.9</td>
<td>791.071 (290.593)</td>
<td>(274.978, 1307)</td>
<td>1.051 \cdot 10^3</td>
<td>(321.879, 1780)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MSE = 1.281 \cdot 10^5</td>
<td></td>
<td>MSE = 1.711 \cdot 10^5</td>
<td></td>
</tr>
<tr>
<td>Pareto, $\gamma = 1/2$</td>
<td>99</td>
<td>8.879 (0.863)</td>
<td>(7.346, 10.412)</td>
<td>10.035 (1.046)</td>
<td>(8.177, 11.893)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MSE = 2.001</td>
<td></td>
<td>MSE = 1.095</td>
<td></td>
</tr>
<tr>
<td></td>
<td>99.9</td>
<td>27.733 (4.772)</td>
<td>(19.258, 36.208)</td>
<td>32.146 (6.111)</td>
<td>(21.293, 42.999)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MSE = 37.904</td>
<td></td>
<td>MSE = 37.618</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>MSE = 21.394</td>
<td></td>
<td>MSE = 16.410</td>
<td></td>
</tr>
<tr>
<td></td>
<td>99.9</td>
<td>75.566 (35.321)</td>
<td>(12.836, 138.296)</td>
<td>76.795 (37.626)</td>
<td>(9.971, 143.619)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MSE = 2.023 \cdot 10^3</td>
<td></td>
<td>MSE = 2.261 \cdot 10^3</td>
<td></td>
</tr>
</tbody>
</table>


Table 6.4 True values of high quantiles for different heavy-tailed distributions.

<table>
<thead>
<tr>
<th>PDF</th>
<th>$(1 - p) \cdot 100%$</th>
<th>$x_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pareto, $\gamma = 1$</td>
<td>99</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>99.9</td>
<td>1000</td>
</tr>
<tr>
<td>Pareto, $\gamma = 1/2$</td>
<td>99</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>99.9</td>
<td>31.623</td>
</tr>
<tr>
<td>Weibull, $\gamma = 2$</td>
<td>99</td>
<td>21.208</td>
</tr>
<tr>
<td></td>
<td>99.9</td>
<td>47.717</td>
</tr>
</tbody>
</table>

The first conclusion coincides with the conclusions of Theorem 14. The last two conclusions may be explained by the larger variance of \( x_p^w \), especially for smaller sizes.

### 6.5 Application to Web traffic data\(^6\)

High quantile estimators may be applied to determine the thresholds of traffic parameters, (Markovich, 2005b). For example, to optimize TCP one can estimate the quantiles of delays between the arrival times of packets and their acknowledgments. We now apply the estimators \( x_p^w \) and \( x_p^c \) to the real Web data described in Table 1.4.

Table 6.5 contains the values of the high quantile estimates \( x_p^w \) and \( x_p^c \) for the different characteristics of Web traffic. The EVI \( \gamma \) of both estimates was estimated by Hill’s estimator, where the parameter \( k \) was calculated by the bootstrap method. For this purpose, 150 bootstrap resamples were used.

In Table 6.6 the means, standard deviations, and bootstrap confidence intervals of both \( x_p^c \) and \( x_p^w \) for Web traffic characteristics are given. To construct the confidence intervals the procedure described in Section 6.4.2 was used. Here, all estimates were calculated by \( B = 50 \) bootstrap resamples from the sample \( X^n \) with replacement instead of \( N_R \) samples generated from the known distribution. From (1.34) we have \( k^* = 2.13 \).

<table>
<thead>
<tr>
<th>Quantile estimate r.v.</th>
<th>Quantile value ( \hat{x}_p \cdot 10^{-4} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_p^c )</td>
<td>d.s.s. 1.4005 5.812</td>
</tr>
<tr>
<td>s.s.s.</td>
<td>2.299 ( \cdot 10^3 ) 2.1 ( \cdot 10^4 )</td>
</tr>
<tr>
<td>s.r.</td>
<td>69.27 439.5</td>
</tr>
<tr>
<td>i.r.t.</td>
<td>0.1445 0.5493</td>
</tr>
<tr>
<td>( x_p^w )</td>
<td>d.s.s. 1.435 5.688</td>
</tr>
<tr>
<td>s.s.s.</td>
<td>2.407 ( \cdot 10^3 ) 2.02 ( \cdot 10^4 )</td>
</tr>
<tr>
<td>s.r.</td>
<td>56.67 431.6</td>
</tr>
<tr>
<td>i.r.t.</td>
<td>0.0954 0.5402</td>
</tr>
</tbody>
</table>


---

Tolerant 90% confidence intervals of estimates $x_p^w$ and $x_p^c$ for Web traffic data.

<table>
<thead>
<tr>
<th>r.v.</th>
<th>$(1 - p) \cdot 100%$</th>
<th>$x_p^c$</th>
<th>Confidence interval</th>
<th>$x_p^w$</th>
<th>Confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean($x_p^c$)</td>
<td>(StDev($x_p^c$))</td>
<td>Confidence interval</td>
<td>mean($x_p^w$)</td>
<td>(StDev($x_p^w$))</td>
</tr>
<tr>
<td>s.s.s.</td>
<td>99</td>
<td>$2.04 \cdot 10^7$</td>
<td>$(5.437 \cdot 10^6, 7.025 \cdot 10^6)$</td>
<td>$2.116 \cdot 10^7$</td>
<td>$(7.75 \cdot 10^6, 6.296 \cdot 10^6)$</td>
</tr>
<tr>
<td></td>
<td>99.9</td>
<td>$1.312 \cdot 10^8$</td>
<td>$(-7.503 \cdot 10^7, 9.682 \cdot 10^7)$</td>
<td>$2.136 \cdot 10^8$</td>
<td>$(-1.317 \cdot 10^8, 1.621 \cdot 10^8)$</td>
</tr>
<tr>
<td>d.s.s.</td>
<td>99</td>
<td>$1.466 \cdot 10^4$</td>
<td>$(5.354 \cdot 10^3, 3.374 \cdot 10^4)$</td>
<td>$1.424 \cdot 10^4$</td>
<td>$(6.587 \cdot 10^3, 6.587 \cdot 10^3)$</td>
</tr>
<tr>
<td></td>
<td>99.9</td>
<td>$5.69 \cdot 10^4$</td>
<td>$(-1.239 \cdot 10^4, 3.253 \cdot 10^4)$</td>
<td>$5.93 \cdot 10^4$</td>
<td>$(-4.131 \cdot 10^3, 2.978 \cdot 10^4)$</td>
</tr>
<tr>
<td>s.r.</td>
<td>99</td>
<td>$6.252 \cdot 10^5$</td>
<td>$(5.246 \cdot 10^5, 7.258 \cdot 10^5)$</td>
<td>$5.487 \cdot 10^5$</td>
<td>$(4.654 \cdot 10^5, 6.32 \cdot 10^5)$</td>
</tr>
<tr>
<td></td>
<td>99.9</td>
<td>$4.196 \cdot 10^6$</td>
<td>$(2.664 \cdot 10^6, 7.193 \cdot 10^6)$</td>
<td>$4.119 \cdot 10^6$</td>
<td>$(2.342 \cdot 10^6, 8.342 \cdot 10^6)$</td>
</tr>
<tr>
<td>i.r.t.</td>
<td>99</td>
<td>$1.281 \cdot 10^3$</td>
<td>$(1015, 124.765)$</td>
<td>$1.135 \cdot 10^3$</td>
<td>$(766.536, 173.505)$</td>
</tr>
<tr>
<td></td>
<td>99.9</td>
<td>$8.068 \cdot 10^3$</td>
<td>$(600.22, 3.506 \cdot 10^3)$</td>
<td>$8.026 \cdot 10^3$</td>
<td>$(730.75, 3.425 \cdot 10^3)$</td>
</tr>
</tbody>
</table>


From the latter study one may conclude that both estimates $x_p^w$ and $x_p^c$ give rather similar results since the sizes of the samples considered are large enough.

### 6.6 Exercises

1. High quantile estimation.
   Generate $X^n$, $n = 100$, according to some heavy-tailed distribution (e.g., Burr, Pareto or Fréchet) and determine its true 99% and 99.9% quantiles.
   Calculate the $(1 - p)$th quantiles ($p = 0.1, p = 0.01$) by the POT method (6.4). Estimate the parameters $\hat{\sigma}$ and $\hat{\gamma}$ by the ML method and method of moments. Take the order statistics $X_{(n-k)}$, $k \in \{10, 30, 50\}$, as threshold $u$.
   Compare the estimates with the true values of the quantiles. Draw conclusions regarding the sensitivity of the POT method to the parameter selection.

2. Calculate the $(1 - p)$th quantiles ($p = 0.1, p = 0.01$) using the estimators (6.5) and (6.8). Calculate $\hat{\gamma}$ for these estimators using the Hill and moment
estimators. Select the parameter $k$ of both estimators by the plot and bootstrap methods (Section 1.2.2).

Compare the values of high quantiles constructed by these estimates.

3. Calculate the bootstrap confidence intervals for estimates (6.4), (6.5), and (6.8) by formulas (6.15) and (1.34)–(1.36); see Section 6.4.2. For this purpose, calculate estimates $\hat{x}_p^1, \ldots, \hat{x}_p^B$ by $B = 50$ bootstrap resamples with replacement.
In this chapter the nonparametric estimation of a hazard rate function is considered for both light- and heavy-tailed distributions. For the heavy-tailed case a transformation approach to the light-tailed case is presented. In the light-tailed case the hazard rate is evaluated as the solution of an integral equation. Such tasks are ill-posed and, hence, the solution is obtained by a statistical analog of Tikhonov’s regularization method. The theoretical background of the latter method and the numerical solution of ill-posed problems using empirical data are presented. The regularized estimates are proved to converge in the uniform metric of space $C$ for a certain choice of the regularization parameter, as well as in the metric of space $L_2$ in the case of a bounded variation of the $k$th derivative of the hazard rate. Finally, the identification of semi-Markov models and their application in population analysis and teletraffic engineering are discussed. The estimate of the intensity of a nonhomogeneous Poisson process is given. A ratio of hazard rates is considered with regard to the application to the failure time detection in stochastic processes and hormesis detection in biological systems.
7.1 Definition of the hazard rate function

**Definition 16** Let $X$ be an r.v., for example the lifetime of some element, with continuous DF $F(x)$, and corresponding PDF $f(x)$. The hazard rate function (or, in population analysis, the mortality risk) of $X$ is defined by the function

$$ h(x) = \frac{f(x)}{1 - F(x)}. \quad (7.1) $$

The problem of the estimation of the hazard rate function relates to the different behavior of this function on the right-hand side of the real axis. In the case of light-tailed distributions $h(x) \to \infty$ as $x \to \infty$, while for the exponential distribution $h(x)$ is equal to the constant intensity $\lambda$ of this distribution, and for heavy-tailed distributions $h(x) \to 0$ as $x \to \infty$. This is illustrated in Figure 7.1 for the light-tailed normal and exponential distributions, as well as a Weibull distribution with shape parameter 0.3 and the Cauchy distribution which have heavy tails.

The von Mises conditions reflect the difference in this behavior for the three classes of a GEV distribution (1.2). The latter implies that $F(x)$ is in the domain of attraction of one of the three types. Let the second derivative $F''(x) = f'(x)$ of $F(x)$ exist. Then for $\gamma > 0$, if

$$ \lim_{x \to \omega(F)} \left( \frac{1 - F(x)}{f(x)} \right)' = \begin{cases} 
1/\gamma, & \text{then } F \text{ converges to Fréchet type}, \\
-1/\gamma, & \text{then } F \text{ converges to Weibull type}, \\
0, & \text{then } F \text{ converges to Gumbel type},
\end{cases} $$

where $\omega(F)$ is the supremum of the support of $F(x)$; see Reiss (1989, p. 159).

![Figure 7.1](image-url) Hazard rate function for standard exponential (horizontal solid line), standard normal (solid line), Weibull with the shape parameter 0.3 (dashed line), and Cauchy distributions (dotted line).
The hazard rate function may be defined in terms of the so-called survival function $F(x)$ by the equation

$$F(x) = 1 - F(x) = P\{X > x\} = \exp \left( - \int_0^x h(t) \, dt \right).$$

The survival function determines the probability of living not less than $x$ years. Hence, the hazard rate function determines the probability of the death of an individual in the time interval $[t, t + \Delta t)$, where $\Delta t$ is sufficiently small, under the condition of his survival until the age $t$, that is,

$$P\{t \leq T < t + \Delta t \mid T \geq t\} = h(t)\Delta t + o(\Delta t),$$

where $T$ is interpreted as a lifetime (one may rephrase this in terms of the stability of technical systems). In practice, one may use the approximation

$$h(t)\Delta t \simeq \frac{d(t, t + \Delta t)}{D(t)},$$

where $d(t, t + \Delta t)$ is the number of failed objects observed in the interval $[t, t + \Delta t)$ and $D(t)$ is the number of elements surviving until age $t$.

The function $h(t)$ may be interpreted as a rate of transition from the first state to the second in the simplest birth–death model. For instance, if $T$ is the duration of a chronic disease, that is, the time from onset of sickness until death, then the period spent in the first state corresponds to the illness of individuals and the second state corresponds to the death. In this case, $h(t)$ is the mortality rate among sick individuals (Figure 7.2).

Possible applications of the hazard rate function are as follows:

- the identification of Markov and semi-Markov models (the estimation of transition rates between different states);
- the mortality rate in population analysis;
- the ratio between hazard rates of two populations (groups) for failure time detection in stochastic processes, or hormesis detection in biological systems.

Many applied problems can be considered in terms of an inverse problem that establishes the relationship between the ‘result’ and the ‘source’ processes. Here, the researcher deals with a model and a set of experimental values of the process under study. The consideration of the ‘result–source’ processes enables

![Figure 7.2 Two-state survival model.](image-url)
one to integrate the mathematical descriptions of various applied problems with the methods to solve inverse problems. This integration is exemplified by the analysis of population processes and the solution of these problems by semi-Markov models, (Markovich, 1995; Markovich and Michalski, 1995). Formal relationships of the probabilities of being in different states at different times are represented as kernel integral equations whose right-hand side, and often the kernel itself, are known approximately from experimental data. The hazard rate function may be estimated by means of the approximate solution of the integral equations. A specific regularization technique for the solution of these stochastic integral equations, which constitutes an ill-posed problem, is required here.

7.2 Statistical regularization method

We begin with the state of the theoretical background of the statistical regularization method which will be required later. Let $U$ and $V$ be metric spaces with metrics $\rho_U$ and $\rho_V$, and $A$ be a continuous one-to-one operator from $U$ to $V$. We seek the solution $g$ of the operator equation

$$Ag = y, \quad g \in U, \quad y \in V,$$

(7.2)

for the case where one knows the operators $A_n$ and functions $y_n$, $n = 1, 2, \ldots$, instead of the precise data $(A, y)$. $A_n$ and $y_n$ are defined on a probability space $(\Omega, \mathcal{A}, P)$ and are close to $A$ and $y$ in some probabilistic sense; here, $y_n \in V$, and the operator $A_n$ is continuous for any $\omega \in \Omega$.

The solution $g_n = A_n^{-1}y_n$ cannot be used as an approximation of $g$ since it is unstable with respect to the variations in the empirical data. More precisely, small deviations of $y_n$ from $y$ may lead to large deviations of $g_n$, that is to say, the inverse operator $A_n^{-1}$ may be not continuous. Such a problem is ill-posed (see Definition 13, p. 67).

Example 10 Consider the estimation of the PDF $f(x)$ as a solution of Fredholm’s equation (2.8). For a continuous DF $F(x)$ we evidently have $f(x) = F'(x)$. If the unknown $F(x)$ is replaced by the stepwise empirical DF $F_n(x)$ the derivative $(F_n'(x))'$ does not exist at a finite number of points corresponding to the jumps in $F_n(x)$.

Example 11 (Tikhonov and Arsenin, 1977). Let $y$ be an $(n \times 1)$ vector. If $A$ is an $(n \times n)$ symmetric matrix and $\det A \neq 0$ (or $\text{rank } A = n$) then $A^{-1}$ exists. By an orthogonal transformation $g = Vg^*$, $y = Vy^*$ one can represent $A$ in diagonal form $[\lambda_1, \ldots, \lambda_n]$, where $\lambda_i$, $i = 1, \ldots, n$, are eigenvalues of the matrix $A$. Then a linear system $Ag = y$ is represented as $\lambda_i g_i^* = y_i^*$, $i = 1, \ldots, n$. If $\text{rank } A = r < n$, then $n - r$ eigenvalues of $A$ are equal to zero. Let $\lambda_i = 0$ for $i = 1, \ldots, r$ and $\lambda_i \neq 0$

1 For an individual it might be the states of health, disease and death, or for a technical system the states of good and bad work.
for $i = r + 1, \ldots, n$. For given approximations $y_n$ and $A_n$ such that $\|y_n - y\| \leq \delta$ and $\|A_n - A\| \leq \varepsilon$ ($\delta > 0$, $\varepsilon > 0$), the eigenvalues $\tilde{\lambda}_i$, $i = r + 1, \ldots, n$ may be close to zero for a sufficiently small $\varepsilon$. Then $\tilde{g}_i = \gamma_i / \lambda_i$ may be large for small perturbations of $A_n$ and $y_n$. This implies that the solution of the system of linear equations $Ag = y$ is unstable.

The regularization method proposed in Tikhonov and Arsenin (1977) involves the stabilization of solutions using the reduction of the set of possible solutions $D \subseteq U$ to a compact set $D^*$ due to the following lemma.

**Lemma 3** The inverse operator $A^{-1}$ is continuous on the set $N^* = AD^*$ if the continuous one-to-one operator $A$ is defined on the compact $D^* \in D \subseteq U$.

This reduction is provided by the stabilizing functional, which is defined on $D$. The regularization method is similar to the Lagrange method in the sense that we want to find a solution $g_n$ which minimizes a functional $\Omega(g_n)$ such that $\|A_n g_n - y_n\| \leq \varepsilon$, $\varepsilon > 0$.

To find the solution of (7.2), we extend the method of regularization from a deterministic operator equation to the case of stochastic ill-posed problems. The function that minimizes the functional

$$R_\gamma(y_n, g) = \|A_n g - y_n\|_V^2 + \gamma \Omega[g]$$

in a set $D$ of functions $g \in U$ is taken as an approximate solution of (7.2). Here, $\gamma > 0$ is the regularization parameter and $\Omega[g]$ is a stabilizing functional that satisfies the standard conditions (see p. 67).

Theorems 15 and 16 provide the theoretical background of the statistical regularization method for the case of an accurately given operator $A$ (Vapnik and Stephanyuk, 1979; Vapnik, 1982), and Theorem 17 for the case of an inaccurately given operator $A$ (Stefanyuk, 1986).

**Theorem 15** If, for each $n$, a positive $\gamma = \gamma(n)$ is chosen such that $\gamma \to 0$ as $n \to \infty$, then for any positive $\nu$ and $\mu$ there will be a number $N = N(\nu, \mu)$ such that, for all $n > N$, the elements $g_n^\gamma(x)$ that minimize the functional (7.3) satisfy the inequality

$$P\{|\rho V (g_n^\gamma, g) > \nu\} \leq P\{|\rho V (y_n, y) > \mu \gamma\},$$

where $g$ is the precise solution of (7.2) with the right-hand side $y$, and $\rho(f, g) = \|f - g\|$.

For the estimation of a PDF $f(x)$ by the regularization method (see pp. 67, 68) one can find the conditions on the regularization parameter $\gamma(n)$ which provide consistent estimates.

Let $y$ be the DF $F(x)$, $y_n$ be the empirical DF $F_n(x)$, $V = C[a, b]$ and $\rho V^2(F_n(x), F(x)) = \sup_x |F_n(x) - F(x)|$. Taking into account the inequality

$$P\{\sup_x |F_n(x) - F(x)| > \eta\} \leq 2 \exp \left(-2n \eta^2\right),$$

(7.5)
(Prakasa Rao, 1983), we get from (7.4) that

\[ P\{\rho_U(f_n^\gamma, f) > \nu\} \leq 2 \exp(-n \mu \gamma). \]  

(7.6)

Hence, if \( \gamma(n) \to 0 \) and \( n \gamma(n) \to \infty \) as \( n \to \infty \) then the sequence \( f_n^\gamma(x) \) converges in probability to the PDF \( f(x) \) in metric space \( U \). If \( \sum_{n=1}^{\infty} \exp(-n \mu \gamma) < \infty \) then the sequence converges with probability one by the Borel–Cantelli lemma.

**Theorem 16**  Let \( U \) be a Hilbert space, \( A \) be a linear operator, and \( \Omega(g) = ||g||_U^2 \). Then, for any \( \varepsilon > 0 \) and any constants \( c_1, c_2 > 0 \), there exists a number \( n(\varepsilon) \) such that, for all \( n > n(\varepsilon) \),

\[ P\{ ||g_n^\gamma - g||_U^2 > \varepsilon \} \leq 2P\{ \rho_V^2(y_n, y) > (\varepsilon/2)\gamma \}. \]

**Theorem 17**  Let \( U \) and \( V \) be normed spaces. For any \( \varepsilon > 0 \) and any constants \( c_1, c_2 > 0 \), there exists a number \( n(\varepsilon) \) such that, for all \( \gamma \leq \gamma_0 \),

\[ P\{ \omega : ||g_n^\gamma - g||_U > \varepsilon \} \leq P\{ \omega : \frac{||y_n - y||_V}{\sqrt{\gamma}} > c_1 \} + P\{ \omega : \frac{||A_n - A||}{\sqrt{\gamma}} > c_2 \}, \]

(7.7)

where

\[ ||A_n - A|| = \sup_{g \in D} \frac{||A_n g - Ag||_V}{\Omega^{1/2}(g)}. \]

(7.8)

These theorems imply that the minimization of (7.3) is a stable problem, i.e. close functions \( y_n \) and \( y \) (and close operators \( A_n \) and \( A \)) correspond to close (in probabilistic sense) regularized solutions \( g_n^\gamma \) and \( g \) that minimize the functionals \( R_\gamma(y_n, g) \) and \( R_\gamma(y, g) \), respectively.

For Hilbert spaces \( U \) and \( V \), the solution of (7.2) with \( \Omega(g) = ||g||_U^2 \) has a simple form,

\[ g_n^\gamma = (\gamma I + A_n^* A_n)^{-1} A_n^* y_n, \]

(7.9)

where \( I \) is a unit operator and \( A_n^* \) is the adjoint operator of \( A_n \). The stability of the approximation \( g_n^\gamma \) to \( g \) is ensured by an appropriate choice of \( \gamma \). Various methods for choosing the regularization parameter were developed, for example, in Morozov (1984), Engl and Gfrerer (1988) and Vapnik et al. (1992). The mismatch method determines \( \gamma \) from the equality

\[ ||A_n g_n^\gamma - y_n||_V = \varepsilon(n) + \zeta(n, g), \]

(7.10)

where \( \varepsilon(n) \) and \( \zeta(n, g) \) are known estimates of the data error, \( ||y_n - y||_V \leq \varepsilon(n) \), \( ||A_n g - Ag||_V \leq \zeta(n, g) \); see Morozov (1984). The stochastic analog of the mismatch method is the discrepancy method (2.37).

If the operator is defined precisely (\( \zeta(n, g) = 0 \)), then the choice of \( \gamma \) from (7.10) provides a rate of convergence of the regularized estimate \( g_n^\gamma \) to \( g \) that is no better than \( O(\varepsilon^{1/2}) \); see Engl and Gfrerer (1988).
7.3 Numerical solution of ill-posed problems

Many integral equations such as (7.20), (7.21), (7.37), arising particularly in teletraffic and population modeling, are related to the hazard rate. Hazard rate functions can often be formulated in terms of a Volterra integral equation of the first kind

\[
\int_0^x K(x, t) g(t) \, dt = y(x)
\]

(7.11)

or the second kind

\[
g(x) - \int_0^x K(x, t) g(t) \, dt = y(x),
\]

(7.12)

where \( g \in U \), \( y \in V \), and \( K(x, t) \) is a real-valued kernel function. Let \( U \) and \( V \) be Hilbert spaces. We suppose that these equations can be represented by systems of linear equations. For this purpose, we represent the unknown function \( g(t) \) by a linear combination

\[
\hat{g}(t) = \sum_{j=1}^{N} \alpha_j \varphi_j(t)
\]

(7.13)

of \( N \) known normalized orthogonal functions \( \varphi_j(t) \) of a basis in \( U \), for example, in \( L_2 \). Laguerre or trigonometric polynomials provide examples of such functions. Substituting (7.13) into (7.11) or (7.12), we get, for \( i = 1, \ldots, n \),

\[
\sum_{j=1}^{N} \alpha_j \int_0^{X_i} K(X_i, t) \varphi_j(t) \, dt = y(X_i)
\]

and

\[
\sum_{j=1}^{N} \alpha_j \left( \varphi_j(X_i) - \int_0^{X_i} K(X_i, t) \varphi_j(t) \, dt \right) = y(X_i),
\]

respectively. Generally, we obtain a system of linear equations

\[
A_n \alpha = Y_n,
\]

(7.14)

where the elements of the \( n \times N \) matrix \( A_n \) are given, for \( i = 1, \ldots, n \) and \( j = 1, \ldots, N \), by

\[
a_{i,j} = \int_0^{X_i} K(X_i, t) \varphi_j(t) \, dt
\]

and

\[
a_{i,j} = \varphi_j(X_i) - \int_0^{X_i} K(X_i, t) \varphi_j(t) \, dt
\]

in the case of the equations (7.11) and (7.12), respectively. Here \( A_n \) and \( Y_n \) are random, since they are obtained from a sample \( X^n = (X_1, \ldots, X_n) \). Here
\[ Y_n = (Y^1, \ldots, Y^n)^T \] is a random \( n \times 1 \) vector of observations in the sample points \( X_1, \ldots, X_n \) in the presence of stochastic errors, and \( \alpha = (\alpha_1, \ldots, \alpha_N)^T \) is an \( N \times 1 \) vector of unknown parameters we wish to estimate.

For equation (7.20) the vector \( Y_n \) can be defined as the random vector \( (F_n(X_1), \ldots, F_n(X_n))^T \) and the elements of the matrix \( A_n \) have the form

\[ a_{i,j} = \int_0^{X_i} (1 - F_n(t)) \varphi_j(t) dt. \]

Here, the unknown DF \( F(t) \) is replaced by its empirical estimate \( F_n(t) \).

The matrix \( A_n \) may possess eigenvalues equal to or close to zero. In the first case, the system has no solution, whereas in the second case this solution is a very poor approximation to the real \( \alpha \). Roughly speaking, the regularization procedure increases the eigenvalues by adding the regularization parameter \( \gamma \) and helps to solve ill-posed and ill-conditioned problems.

According to Tikhonov’s regularization method (Tikhonov and Arsenin, 1977), one constructs a regularized approximate solution \( \alpha^\gamma \in U \) as the global minimum in \( U \) of the smoothing functional

\[ R_{\gamma}(Y_n, \alpha) = \|A_n \alpha - Y_n\|_V^2 + \gamma \|\alpha\|_U^2 \] (7.15)

for a given value \( \gamma > 0 \) of the regularization parameter, where \( \|\alpha\|_U^2 \) satisfies all conditions of the stabilization functional (see p. 67).\(^2\)

By the common theory proved in Vapnik (1982) for an accurately given matrix \( A \) if the function \( K(x, t) \) is precisely known and in Stefanyuk (1986) for an inaccurately given matrix \( A_n \), the convergence of the regularized estimates to the exact functions in the metric space \( U \) is satisfied if \( \gamma \to 0 \) for increasing sample size \( n \to \infty \) (see Theorems 15–17).

For Hilbert spaces \( U \) and \( V \) the minimum of \( R_{\gamma}(Y_N, \alpha) \) is achieved (analogously to (7.9) when \( A_n \) is a matrix) by

\[ \alpha^\gamma = \alpha^\gamma(A_n, Y_n) = (\gamma I + A_n^T A_n)^{-1} A_n^T Y_n, \]

(7.16)

where \( I \) is the identity matrix.

The parameters \( N \) and \( \gamma \) may be selected from a sample from the minimum of the criterion (Michalski, 1987)

\[ I_{\gamma,N} = \frac{\|Y_n - C_\gamma Y_n\|_2^2}{1 - \frac{2}{n} \text{tr}(C_\gamma)}, \]

(7.17)

which is a variant of the cross-validation method (Golub et al., 1979). Here \( C_\gamma = A_n (\gamma I + A_n^T A_n)^{-1} A_n^T \), \( \text{tr}(C_\gamma) = N - \sum_{i=1}^N (1 + \lambda_i/\gamma)^{-1} \) is the trace of \( C_\gamma \) and \( \lambda_i \) are the eigenvalues of the matrix \( A_n^T A_n \). The minimization is performed in the region \( 0 < 2 \text{tr}(C_\gamma) < n \), where the denominator of the expression (7.17) is positive.

\( ^2 \|\alpha\|_2^2 = \sum_{i=1}^N \alpha_i^2 \) is an example of \( \|\alpha\|_U^2 \).
Finally, the regularized nonparametric estimate is obtained from
\[ g^\gamma(t) = \sum_{j=1}^{N} \alpha^\gamma_j \varphi_j(t), \]
where \( \alpha^\gamma \) estimates \( \alpha \).

### 7.4 Estimation of the hazard rate function of heavy-tailed distributions

Without loss of generality we shall now restrict ourselves to nonnegative r.v.s \( X \). For the estimation of the hazard rate \( h(x) \) in the case of heavy-tailed distributions, it is natural to transform an underlying r.v. \( X \) into a new r.v. \( Y = T(X) \) with known behavior of \( h(x) \). It is convenient to consider the transformations \( T: [0, \infty) \to [0, 1] \) to a finite interval. The transformations considered in Sections 4.2 and 4.3 may be used as \( T(x) \).

By (3.26) and (7.1) we have
\[ h(x) = \frac{g(T(x))T'(x)}{1 - G(T(x))} = h^\delta(T(x))T'(x), \]
where \( G(x) = \int_0^x g(u)du \), and \( g(x) \) and \( h^\delta(x) \) are the PDF and hazard rate of the transformed r.v. \( Y \). The latter formula is a full analog of (3.26) for PDFs.

Like the PDF, the function \( h(x) \) is invariant with respect to monotone transformation in the metric of the space \( L_1 \), that is,
\[ \int_0^\infty |\hat{h}(x) - h(x)|dx = \int_0^1 |\hat{h}^\delta(x) - h^\delta(x)|dx. \]

Here, \( h(x) \) is the hazard rate of the r.v. \( X \), \( \hat{h}(x) \) is the estimate of \( h(x) \), \( \hat{h}^\delta(x) = \hat{g}(x)/(1 - \hat{G}(x)) \) is the estimate of \( h^\delta(x) \), \( \hat{g}(x) \) and \( \hat{G}(x) \) are the estimates of \( g(x) \) and \( G(x) \) on \([0,1]\). This invariance is not valid in the spaces \( L_2 \) and \( C \). In \( L_2 \) we get
\[ \int_0^\infty \left( \hat{h}(x) - h(x) \right)^2 dx = \int_0^1 \left( \hat{h}^\delta(u) - h^\delta(u) \right)^2 T'(T^{-1}(u))du \]
\[ \leq c \int_0^1 \left( \hat{h}^\delta(u) - h^\delta(u) \right)^2 du, \]
if, for any \( u \in [0, 1] \),
\[ 0 < T'(T^{-1}(u)) \leq c. \]  

(7.19)

Obviously,
\[ \int_0^\infty \left( \hat{h}(x) - h(x) \right)^2 dx \leq \sup_{x \in [0,1]} |\hat{h}^\delta(x) - h^\delta(x)| \int_0^1 T'(T^{-1}(u))du. \]
This implies that the accuracy of the estimation of the hazard rate function on a finite interval determines the accuracy of the estimation on \([0, \infty)\). Evidently, the transformations \(2/\pi \arctan x\) and \((4.11)\) obey the property \((7.19)\).

We note that the estimation of the hazard rate \(h^g(x)\) of the r.v. \(Y\) has the same problems as the estimation of \(h(x)\) for distributions defined on \([0,1]\). That is, \(h^g(x) \to \infty\) holds as \(x \to 1\). The proof of this property is provided in Stefanyuk (1992). For example, for the Pareto distribution with DF \(F(x) = 1 - (1 + \gamma x)^{-1/\gamma}\), \(x \geq 0\), and transformation \((4.11)\) we have by \((3.25)\) that

\[
h^g(x) = 2 \left( (1 - x)^{2\gamma + 1} + \gamma \hat{\gamma} \left( 1 - x - (1 - x)^{2\gamma + 1} \right) \right)^{-1}.
\]

This implies that \(h^g(x) \to \infty\) as \(x \to 1\). For the Weibull-type distribution with DF \(F(x) = 1 - \exp(-x^\alpha), \alpha > 0, x > 0\), we get

\[
h^g(x) = \alpha \left( \frac{(1 - x)^{-2\hat{\gamma}} - 1}{\hat{\gamma}} \right)^{\alpha - 1} (1 - x)^{-2\hat{\gamma} - 1} \sim (1 - x)^{-1 - 2\hat{\gamma}} 
\]

that is, \(h^g(x) \to \infty\) as \(x \to 1\) for \(\alpha > 0\).

In populations of both living individuals and inanimate objects such as automobile motors a common tendency has been discovered: the mortality risk or hazard rate decreases at infinity, which corresponds to heavy-tailed distributions (Yashin et al., 1996).

Below, the accuracy of estimates of \(h(x)\) in the metrics of the spaces \(L_2\) and \(C\) (Section 7.5) and the ratio between hazard rates of two populations (Section 7.6) are considered in the finite case (for compactly supported distributions), when PDF \(f(x)\) is equal to zero outside a bounded interval.

### 7.5 Hazard rate estimation for compactly supported distributions

#### 7.5.1 Estimation of the hazard rate from the simplest equations

Let us assume that there exists a sample \(X^n = (X_1, \ldots, X_n)\) of independent observations of a r.v. (say, the lifetime of an individual) that takes values in a limited interval \([0, d]\) and is distributed with PDF \(f(x)\) and DF \(F(x)\), where \(F(x) \neq 1\) for \(x \in [0, d]\). By definition, the hazard rate \(h(x)\) obeys \((7.1)\).

The estimation of this function is hindered primarily by the fact that \(h(x)\) tends to infinity as \(x \to d\). Let us represent \(h(x)\) as the solution of the equations

\[
\int_0^t h(x) (1 - F(x)) \, dx = F(t) \tag{7.20}
\]
or

\[
\int_0^t h(x) \, dx = -\ln (1 - F(t)) \tag{7.21}
\]
Let us represent these equations in the operator form

$$Ah = y, \quad h \in U, \ y \in V,$$

where $U$ and $V$ are normed spaces. The form of the operator $A$ depends on the form of the kernel function of the integral equations.

To solve (7.20) and (7.21) approximately, the unknown DF $F(x)$ is replaced by its empirical estimate $F_n(x)$ constructed from the sample $X^n$. This means that in the case of (7.20) both the right-hand side and the operator are defined imprecisely, and in the case of (7.21) only the right-hand side is defined imprecisely.

**Solution of equation (7.20)**

The estimates of the hazard rate arising from (7.20) were proposed in Stefanyuk (1992). The minimum of the functional

$$R_\gamma(F_n, h) = \sum_i \left( \int_{\Delta_i} h(x) (1 - F(x)) \, dx - \int_{\Delta_i} dF_n(x) \right)^2 + \gamma \int_0^d (d - x)^2 h^2(x) \, dx$$

with respect to $h(t)$ for a fixed $\gamma = \gamma(n) > 0$ was proposed as an estimate of $h(t)$. Here, $\Delta_i$ are disjoint subintervals covering the interval $[0, d]$. Since $h(x) \to \infty$ as $x \to d$ the weight $(d - x)^2$ is implemented such that the integral exists.

The minimum of the latter functional is reached on the function

$$h^*(x) = \frac{1 - F(x)}{(d - x)^2} \sum_i c_i^* \mathbf{1}(x \in \Delta_i),$$

where

$$c_i^* = \frac{\int_{\Delta_i} dF_n(x)}{\int_{\Delta_i} \left( \frac{1 - F(x)}{d - x} \right)^2 \, dx + \gamma}.$$

Since $F(x)$ is unknown, one may replace it by the empirical DF $F_n(x)$ or by its piecewise linear smoothing (a polygon).

**Solution of equation (7.21)**

The theory for solving (7.21) was developed in Markovich (1998). Note that $F(x)$ can be replaced on $[0, d)$ by a close function – for example, by the empirical DF $F_n(x)$ or by a polygon. However, if $F_n(x)$ is used, then the right-hand side of $\gamma_n(x) = -\ln(1 - F_n(x))$ may be unlimited on $[0, d)$, provided that the sample occupies an interval smaller than $[0, d)$. The polygon is close to $F(x)$ on $[0, d)$ in a linear sense.

Suppose it is known in advance that the required function $h(t)$ is defined on $[0, x_a]$, where $0 \leq x_a < d$, and $F(x_a) = a$, $0 \leq a < 1$. Let $a$ be known from the conditions of the problem, in population analysis for example, $0 < a < 1$. 


is a proportion of the individuals that died before the age of \( x_a \). Then one can replace \( -\ln(1 - F(t)) \) on \( t \in [0, x_a] \) by \( -\ln(1 - F_n(t) + (1 - \eta(t)) \cdot 1(F_n(t) = 1)) \), where the function \( \eta(t) \) determines the interval of the line connecting the points \((X_{(n)}, F_n(X_{(n-1)}))\) and \((x_a, a)\), \( X_{(1)} \leq \ldots \leq X_{(n-1)} \leq X_{(n)} \) are the order statistics of the sample \( X^n \), and \( 1(A) \) is the indicator function of the event \( A \), that is,

\[
\eta(x) = \frac{1/n + a - 1}{x_a - X_{(n)}} (x - x_a) + a
\]

holds as \( x \in [X_{(n)}, x_a] \). Obviously, \( \eta(x) \leq \max(1 - 1/n, a) \) holds, since \( \eta(X_{(n)}) = 1 - 1/n \) and \( \eta(x_a) = a \). Let

\[
\eta^*(t) = (1 - \eta(t)) \cdot 1(F_n(t) = 1).
\]  

(7.22)

We now give an example of the regularized estimate of the hazard rate that converges to \( h(t) \) in the uniform metric. Let \( h(t) \) satisfy the Hölder (or Lipschitz) condition with the \( \mu \quad (0 < \mu \leq 1) \),

\[
\sup \{|h(t_1) - h(t_2)|/|t_1 - t_2|^\mu : t_1, t_2 \in [0, x_a], t_1 \neq t_2\} < \infty,
\]

which means it belongs to the Hölder space \( H^\mu[0, x_a] \) with the norm

\[
\|h\|_H = \sup_{x \in [0, x_a]} |h(x)| + \sup_{t_1, t_2 \in [0, x_a], t_1 \neq t_2} \left\{ |h(t_1) - h(t_2)|/|t_1 - t_2|^\mu \right\},
\]  

(7.23)

and let \( V \) be the space \( C[0, x_a] \) of functions that are continuous on \([0, x_a]\) and have the norm

\[
\|y\|_C = \sup_{x \in [0, x_a]} |y(x)|.
\]

The functional

\[
\Omega(h) = \|h\|_H^2,
\]

which satisfies all necessary conditions (see p. 67) is taken as the stabilizing functional. The regularized estimate \( h^\gamma(x) \) may be determined by minimizing the functional

\[
R_\gamma(y_n, h) = \left( \sup_{x \in [0, x_a]} |y_n(x) - y(x)| \right)^2 + \gamma \Omega(h),
\]

where \( y(x) = -\ln(1 - F(x)) \), \( y_n(x) = -\ln(1 - F_n(x) + \eta^*(x)) \). By Theorem 15, the estimate \( h^\gamma(x) \) converges to the required function \( h(x) \) in the metric \( H^\mu[0, x_a] \) and, consequently, in the metric \( C[0, x_a] \).

The following theorem, proved in Appendix E, concerns the uniform convergence of the regularized estimates \( h^\gamma(x) \) to the solution \( h(x) \) of (7.21). Let

\[
U = V = C[0, x_a].
\]
Theorem 18 If \( x \in [0, x_a] \), where \( F(x_a) = a, ~0 \leq a < 1 \), \( h^\gamma(x) \) is the regularized estimate of the function \( h(x) \), and the regularization parameter \( \gamma \) obeys
\[
\gamma = \gamma(n) \to 0,
\]
then
\[
P \left\{ \omega : \lim_{n \to \infty} \sup_{x \in [0, x_a]} |h^\gamma(x) - h(x)| = 0 \right\} = 1.
\]

The following lemma, also proved in Appendix E, is required to prove the theorem.

Lemma 4 If \( x \in [0, x_a] \), where \( F(x_a) = a, ~0 \leq a < 1 \), then
\[
\left\| y - y_n \right\|_{C[0, x_a]} \leq -\ln \left( 1 - \sup_{x : F(x) \leq a} \frac{|F_n(x) - \eta^*(x) - F(x)|}{1 - F(x)} \right) = \epsilon(n).
\]

The uncertainty of the function \( y(x) \) can be estimated by the Rényi statistic (Bolshev and Smirnov, 1965):
\[
R_n(0, a) = \sup_{F(x) \leq a} \frac{|F_n(x) - F(x)|}{1 - F(x)}.
\]

The value corresponding to the maximum of the PDF of the statistic \( R_n(0, a) \) can be taken into account in the estimate of the inaccuracy of \( \epsilon(n) \). According to Bolshev and Smirnov (1965), the value of \( \sqrt{\frac{na}{1-a}} R_n(0, 1 - a), \quad 0 < a \leq 1 \), that corresponds to the greatest value of the PDF of the distribution is 0.9. Then, for \( a^* = 1 - a \), we have
\[
\epsilon(n) = -\ln \left( 1 - 0.9 \sqrt{\frac{a^*}{n(1-a^*)}} \right). \tag{7.24}
\]

We now turn to the problem of the optimality of the regularization method when we solve (7.21). Let \( U = V = L_2[0, x_a] \), and let \( h(x) \) be approximated by \( h^\gamma(x; A, y_n) \), that is, the global minimum of the functional
\[
R_\gamma(y_n, h) = \|Ah - y_n\|^2 + \gamma\|h\|^2
\]
in \( U \) for a given value \( \gamma > 0 \) of the regularization parameter. Here and henceforth, \( \|\cdot\| \) is the norm of the space \( L_2[0, x_a] \). Here, \( Ah(t) \equiv \int_0^t \theta(t - \tau) h(\tau) d\tau. \) Since \( U \) and \( V \) are Hilbert spaces the regularized solution is
\[
h^\gamma(A, y_n) = (\gamma I + A^*A)^{-1} A^*y_n,
\]

\footnote{The limiting distribution of the Rényi statistic for \( 0 < a \leq 1 \) is
\[
\lim_{n \to \infty} P \left\{ \sqrt{\frac{na}{1-a}} R_n(0, 1 - a) < x \right\} = 2\Phi(x) - 1, \quad x > 0,
\]
where \( \Phi(x) \) is the DF of the standard normal distribution (Rényi, 1953).}
where $A^*$ is the adjoint operator of $A$. Let us characterize the accuracy of the regularization method with a fixed choice of the parameter $\gamma = \gamma(n)$ ($n$ is the sample size) by

$$\Delta(\epsilon, \gamma) = \|h(x) - h^\gamma(x; A, y_n)\|.$$ 

We note that the operator $B = A^*A$ is self-adjoint (Hermitian) with kernel $K(t, s) = \int_0^{x_a} \theta(t - \tau) \theta(t - s) dt = x_a - \max(\tau, s), 0 \leq \tau, s \leq x_a$. We denote by $0 < \lambda_1^2 < \lambda_2^2 < \ldots$ the characteristic numbers of the positive symmetric kernels that are defined by the operators $AA^*$ and $A^*A$. By \{\varphi_k(x), k = 1, 2, \ldots\} and \{\psi_k(x), k = 1, 2, \ldots\} we denote the corresponding systems of eigenfunctions that are orthonormalized in $L_2[0, x_a]$. Let $\varphi_i(x) = \lambda_i \int_0^{x_a} 1(x \leq t) \psi_i(t) dt$. The eigenvalues of the Hermitian operator $B$ are found from

$$s_k = (\psi, B\psi) = \int_0^{x_a} \int_0^{x_a} \psi_k(x) K(x, \xi) \psi_k(\xi) dx d\xi,$$

provided that $(\psi, \psi) = \int_0^{x_a} |\psi(\xi)|^2 d\xi = 1$ (Kirillov and Gvishiani, 1982). We choose systems of functions

$$\left\{ \psi_k(x) = \sqrt{\frac{2}{x_a}} \cos \left( \frac{\pi k x}{x_a} \right) \right\}, \quad \left\{ \varphi_k(x) = \sqrt{\frac{2}{x_a}} \sin \left( \frac{\pi k x}{x_a} \right) \right\}, \quad k = 1, 2, \ldots,$$

that are orthonormalized in $L_2[0, x_a]$. Then

$$s_k = \frac{1}{\lambda_k^2} = \frac{2}{x_a} \int_0^{x_a} \int_0^{x_a} \cos \left( \frac{\pi k x}{x_a} \right) (x_a - \max(x, \xi)) \cos \left( \frac{\pi k \xi}{x_a} \right) dxd\xi = \left( \frac{x_a}{\pi k} \right)^2.$$

(7.25)

Let us assume that the $k$th ($k \geq 1$) derivative of $h(x)$ exists and has a limited variation on $[0, x_a]$. Then function $h(x)$ can be extended to $[-x_a, 0]$ by means of a polynomial $r(x)$ of order $(2k - 1)$, which is defined by the conditions $r(0) = 0, r'(0) = 0, \ldots, r^{(k-1)}(0) = 0$ and $r(-x_a) = h(x_a), r'(-x_a) = h'(x_a), \ldots, r^{(k-1)}(-x_a) = h^{(k-1)}(x_a)$. The polynomial $r(x)$ can be extended further to the whole real axis. The set of functions meeting this conditions will be denoted by $\varphi_k$.

Let us now consider a Fourier series. We note that any $y(x) \in L_2[0, x_a]$ can be represented by a Fourier series in the orthonormalized basis $\{\varphi_j\}$ in $L_2[0, x_a]$:

$$y(x) = \sum_{i=1}^{\infty} c_i \varphi_i(x).$$

(7.26)
The function \( h(x) \in \varphi_k \) is represented by a series in the orthonormalized basis \( \{\psi_k(x), k = 1, 2, \ldots\} \):

\[
 h(x) = \sum_{i=1}^{\infty} a_i \psi_i(x). \tag{7.27}
\]

Under the above conditions where \( y(x) \) and \( h(x) \) belong to \( L_2[0, x_a] \), the series (7.26) and (7.27) converge in the metric of \( L_2[0, x_a] \) and \( a_i = \lambda_i c_i \). Since \( h(x) \in \varphi_k \), the inequality

\[
 |a_i| \leq \frac{V_k}{x_a^{i+1}}, \quad i = 1, 2, \ldots, \tag{7.28}
\]

where \( V_k \) is the variation of \( h^{(k)}(x) \), is valid for its Fourier coefficients (Fikhtengol’ts, 1965).

**Theorem 19** Let \( X^n = (X_1, \ldots, X_n) \) be a sample of i.i.d. r.v.s with PDF \( f(x) \) and DF \( F(x) \) that are concentrated on \([0, d]\). Let \( x \in [0, x_a], F(x_a) = a, 0 \leq a < 1, h(x) \in \varphi_k \) and the characteristic numbers of the operators \( AA^* \) and \( A^*A \) satisfy (7.25). If, in the regularized estimate \( h^\gamma(x; A, y_n) \) of the solution of (7.21), \( \gamma = n^{-\alpha} \) holds, where \( 4\beta/(2k + 1) \leq \alpha < 1 - 2\beta, 0 < \beta < (k + 1/2)/(2k + 3) \) as \( k \in \{0, 1\} \) and \( \beta \leq \alpha < 1 - 2\beta, 0 < \beta < 1/3 \) as \( k \geq 2 \), then the asymptotic rate of convergence of the estimate \( h^\gamma(x; A, y_n) \) to \( h(x) \) obeys the expression

\[
 P\{ \omega : \lim_{n \to \infty} n^\beta \| h^\gamma(x; A, y_n) - h(x) \| \leq c \} = 1.
\]

Here \( c \) is constant that is independent on \( n \), and \( \| \cdot \| \) is a norm in \( L_2[0, x_a] \).

Theorem 19 is proved in Appendix E.

It follows from the findings of Ivanov et al. (1978) that there exists no hazard rate estimate with bounded variation of the \( k \)th derivative that converges in \( L_2 \) with a rate better than \( (\varepsilon(n))^{(k+0.5)/(k+1.5)} \).

**Remark 11** Let us select \( \varepsilon(n) \) from (7.24), i.e., \( \varepsilon(n) \sim n^{-1/2} \). Then one may observe that the rate of convergence declared in Theorem 19 is optimal in the class \( \varphi_k \), that is, \( n^{-k(0.5)/(2(k+1.5))} \) for \( k = 0 \) (that is the case for the function \( h(x) \) with a limited variation) and for \( k = 1 \).

### 7.5.2 Estimation of the hazard rate from a special kernel equation

We consider the following problem from population analysis. The cause-specific mortality rate among sick people is the object of the interest. The dynamic of the specific disease (and generally any stochastic changes) in the population may be described by compartmental semi-Markov models of different complexity (or number of states). These models allow us to estimate the cause-specific mortality rate among sick people from mortality and incidence data obtained in the whole
population or in groups of interest. Similarly, we can estimate the destruction rate from a specific cause among technical systems in a ‘degraded operating system’ state. The semi-Markov property of the models is important because it reflects the natural changes in the properties of the objects due to the fact of being for a particular time in the specific states.

Such methodology is important for the investigation of the influence of risk factors such as radiation on the health of a population (Markovich, 1995). Similarly, one can estimate different indices in populations such as the rate of revealed morbidity (Markovich and Michalski, 1995).

The two-state model (Figure 7.2) requires the precise monitoring of the lifetimes of sick people, which is expensive. Furthermore, there is a latent morbidity in the population which is not available. Hence, one can consider a three-state model, where there are two life states (‘healthy but at risk’ and ‘sick’) and a death state (Figure 7.3, based on Figure 2 in Markovich et al., 1996). A risk group includes individuals who are in the life states. A transition to the ‘sick’ state is made from the ‘healthy but at risk’ state with rate $\lambda(t)$. A transition to the ‘death’ state is made from ‘healthy but at risk’ and ‘sick’ states with mortality rate $\mu_1(t)$ from causes other than the specific disease – from the ‘sick’ state with the cause-specific mortality rate $\delta(t)$.

We denote by $S(t) = \exp(-\int_0^t (\gamma(u) + \mu_1(u)) du)$ the survivor function corresponding to the total mortality rate of the risk group, that is, the probability of a member of the risk group aged $x$ being alive. We denote by $P_1(t)$ the probability of being in the ‘healthy but at risk’ state. We denote by $g(t) = \delta(t) \exp(-\int_0^t \delta(\tau) d\tau)$ the PDF of the disease duration for sick individuals before death. The relation between the cause-specific mortality in the risk group $\gamma(t)$ and cause-specific mortality among sick people $\delta(t)$ is given by Fredholm’s integral equation

$$
\int_0^x \lambda(y)P_1(y) \exp\left(\int_0^y \mu_1(u) du\right) g(x-y) dy = \gamma(x) \exp\left(-\int_0^x \gamma(u) du\right). \tag{7.29}
$$

We suppose that the incidence rate in the risk group $I(t)$ is available. Note that

$$
\lambda(y)P_1(y) = I(y)S(y).
$$

![Figure 7.3 Three-state model of survival.](image-url)
on substituting the latter expression in (7.29) we have
\[
\int_0^x K(y)g(x - y) \, dy = \gamma(x) \exp \left( - \int_0^x \gamma(u) \, du \right),
\]
where
\[
K(y) = I(y)S(y) \exp \left( \int_0^y \mu_1(u) \, du \right) = I(y) \exp \left( - \int_0^y \gamma(u) \, du \right).
\]
One can estimate \( \delta(t) \), using the solution \( g(x) \) of (7.30), by
\[
\delta(z) = \frac{g(z)}{1 - \int_0^z g(y) \, dy}.
\]

An equation similar to (7.30) can be obtained in a more complicated model, where there is additionally a ‘healthy and not at risk’ state (Figure 7.4, based on Figure 3 in Markovich et al., 1996). The difference is that \( \gamma(t) \) is the cause-specific mortality for the whole population and \( r(t) \) is the rate of transition from ‘healthy and not at risk’ to ‘healthy but at risk’. The three-state model requires the incidence rate in the special contingents with regard to some risk factors, which is expensive and usually limited in scale. The four-state model uses mortality and incidence data for the whole population from official statistics.

We now aim to prove the uniform convergence of regularized solutions of equations such as (7.30). Let \( X_{n_1} = (X_1, \ldots, X_{n_1}) \) be a sample of i.i.d. observations of an r.v. \( X \) that takes values in a bounded interval \([0, d]\) and is distributed with continuous PDF \( y(x) \) and DF \( H(x) \). For example, this r.v. could be the time to the death after onset of the cause-specific disease in the risk group (three-state model) or in the whole population (four-state model).

Let \( Y_{n_2} = (Y_1, \ldots, Y_{n_2}) \) be a sample of i.i.d. observations of the second r.v. that assumes values in the bounded interval \([0, d]\) and has continuous PDF \( f(y) \), the DF \( F(y) \), and the hazard rate \( I(y) \). Note that \( I(y) = f(y) / (1 - F(y)) \). An example
of this r.v. is the time to onset of the disease among people at risk (or in the whole population).

Let us assume that \( H(x) \neq 1 \) and \( F(x) \neq 1 \) for \( x \in [0, d] \).

Let \( Z \) be an unobservable r.v. that assumes values in the bounded interval \([0, d]\) and has continuous PDF \( g(z) \) and hazard rate \( \delta(z) \). The latter function is to be estimated. The time to death after onset of the cause-specific disease among sick people is an example of the r.v. \( Z \).

Let us consider an integral equation

\[
\int_{0}^{x} K(y)g(x-y)dy = y(x), \quad x \in [0, d],
\]

where \( K(y) = I(y)(1-H(y)) \), relative to the PDF of the time of death \( g(z) = h(z)e^{-\int_{0}^{z} h(\tau)d\tau} \) and find \( h(z) \) from the equation

\[
h(z) = \frac{g(z)}{1-\int_{0}^{z} g(y)dy},
\]

Let the PDFs \( y(x) \), \( f(x) \) and \( g(x) \) belong to the space \( C[0, d] \). We assume that the right-hand side and the kernel of equation (7.31) are unknown and are estimated from the empirical data. For example, some nonparametric estimate (histogram, kernel estimate or, generally, a regularized estimate) \( y_{n_1}(x) \) is given instead of \( y(x) \).

The kernel \( K(y) \) is also replaced by its estimate, for example,

\[
\hat{K}(y) = I_{n_2}(y)(1-H_{n_1}(y)) = \frac{f_{n_2}(y)}{1-F_{n_2}(y) + \eta^*(y)(1-H_{n_1}(y))},
\]

where \( \eta^*(t) \) is determined by (7.22), \( f_{n_2}(x) \) is a nonparametric estimate of \( f(x) \) from the sample \( Y^{n_2} \), and \( H_{n_1}(x) \) and \( F_{n_2}(x) \) are the empirical DFs constructed from samples \( X^{n_1} \) and \( Y^{n_2} \).

The uniform convergence of the regularized estimates \( g^\gamma(x) \) to \( g(x) \) in \([0, x_a]\), for example, \( 0 \leq x_a < d \) and \( F(x_a) = a, 0 \leq a < 1 \), is the object of interest.

**Theorem 20** Let \( \sup_{x \in [0, x_a]} f(x) \leq \varepsilon \) for a fixed \( \varepsilon > 0, g^\gamma(x) \) be a regularized estimate of \( g(x) \) obtained by the regularization method, where the stabilizing functional \( \Omega(g) \) satisfies the condition \( \Omega_{\min} = \inf_{g \in D} \Omega(g) > 0 \), and the regularization parameter \( \gamma = \gamma(n) \) be

\[
\gamma(n) \rightarrow 0, \quad \sum_{n=1}^{\infty} \exp(-\mu n \gamma(n)) < \infty, \quad (7.32)
\]

4 One can take the square of the norm \( (7.23) \) as \( \Omega(g) \) satisfying this condition. Note that \( \Omega(g) = \|g\|_C^2 \) does not satisfy the third condition of the stabilizing functional (page 67). In fact, the sequence \( g_n(t) = c \sin nt \) belongs to the sphere \( \sup_{t} |g_n(t)| \leq c \) but does not contain all its limit points, i.e. the set \( M_c = \{g_n : \Omega(g_n) \leq c\} \) is not compact in \( C \) (Vapnik and Stefanyuk, 1979).
at least for one $\mu > 0$, where $n = \min(n_1, n_2)$. Let $g_{n_1}^\gamma(x)$ and $f_{n_2}^\gamma(x)$ be some estimates of $g(x)$ and $f(x)$, respectively, obtained by the regularization method. Then,

$$P\{\omega : \lim_{n \to \infty} \|g^\gamma(x) - g(x)\|_{C[0, x_n]} \leq c\} = 1,$$

where $0 < c < \infty$ is a constant.

The following lemma is required to prove the theorem. Let $A_g = \int_0^x I(y)(1 - H(y))g(x - y)dy$, $A_n g = \int_0^x I_n(y)(1 - H_{n_1}(y))g(x - y)dy$. We denote

$$\inf_{x \in [0, x_n]} \{1 - F_{n_2}(x) + \eta^*(x)\} = \min\{1 - F_{n_2}(x), 1 - a, 1/n\} = C^*.$$

**Lemma 5** If $x \in [0, x_n]$, where $F(x_n) = a, 0 \leq a < 1$, and $f_{n_2}^\gamma(x)$ is an estimate of the PDF $f(x)$, then

$$\|A_n g - Ag\|_C = \sup_{x \in [0, x_n]} |A_n g - Ag| \leq 2 \sup_{x \in [0, x_n]} |I_{n_2}(x) - I(x)| \leq \sup_{x \in [0, x_n]} f(x) \sup_{x} \left|H_{n_1}(x) - H(x)\right| \frac{1 - a}{C^*(1 - a)},$$

where

$$\left|I_{n_2}(x) - I(x)\right| \leq \frac{|f_{n_2}(x) - f(x)| \left(1 + |F_{n_2}(x) - F(x)|\right) + f(x) |F(x) - F_{n_2}(x) + \eta^*(x)|}{C^*(1 - a)}.$$

Let $G(x) = \int_0^x g(\tau)d\tau, \quad G^\gamma(x) = \int_0^x g^\gamma(\tau)d\tau$.

**Theorem 21** Let the assumptions of Theorem 20 hold, $g^\gamma(x)$ be an estimate of the PDF $g(x)$, $x \in [0, x_n]$ such that $G^\gamma(x) \leq C < 1$, $\sup_{x \in [0, x_n]} g(x) \leq \epsilon$ for a fixed $\epsilon > 0$ and $G(x_n) = b, 0 \leq b < 1$. Then

$$P\{\omega : \lim_{n \to \infty} \|h^\gamma(x) - h(x)\|_{C[0, x_n]} \leq c\} = 1,$$

where $0 < c < \infty$ is a constant.

### 7.6 Estimation of the ratio of hazard rates

We consider two r.v.s $X, Y$ (e.g., these might be the survival times in two populations of objects) distributed with PDFs $f(x), g(x)$ and DFs $F(x), G(x)$, respectively. Let $X^n = (X_1, \ldots, X_n), Y^n = (Y_1, \ldots, Y_n)$ be samples of independent
observations of these r.v.s on an interval $[0, d]$ (in population analysis, $d$ is a maximal survival time); $n$ is the sample size. Generally, the sample sizes of $X$ and $Y$ may be different.

Let us suppose that the distributions are compactly supported, that is, $f(x) = g(x) = 0$ if $x \not\in [0, d]$. Let us assume that $g(x) \neq 0$, $x \in [0, d]$ and $F(x) \neq 1$, $G(x) \neq 1$ hold for $x \in [0, d)$. The hazard rate is defined by

$$
\mu_1(x) = \frac{f(x)}{1 - F(x)}
$$

for the first cohort, and by

$$
\mu_2(x) = \frac{g(x)}{1 - G(x)}
$$

for the second cohort. In some applications such as hormesis (Markovich, 2000) or failure time detection (Stefanyuk, 1986), it is useful to consider the ratio of the hazard rates

$$
r(x) = \frac{\mu_1(x)}{\mu_2(x)}
$$

(7.35)

or the ratio between two PDFs, the so-called likelihood ratio (see Section 7.6.1):

$$
q(x) = \frac{f(x)}{g(x)}.
$$

The problem of estimation results from the fact that $r(x)$ can tend to infinity as $x \to d$. We can consider the function $r(x)$ as a solution of Volterra’s integral equation

$$
\int_0^x r(u) \frac{1 - F(u)}{1 - G(u)} dG(u) = F(x)
$$

(7.36)

for $x \in [0, d)$. The estimation of $r(x)$ is an ill-posed problem. In (7.36) the DFs $F(x)$ and $G(x)$ are unknown. However, they can be replaced by close approximations, for example, by the empirical DFs

$$
F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \theta(x - X_i), \quad G_n(y) = \frac{1}{n} \sum_{i=1}^{n} \theta(y - Y_i)
$$

constructed by the samples $X^n$ and $Y^n$. According to the Glivenko–Cantelli theorem the empirical DF is a good approximation of the corresponding DF for sufficiently large $n$ with probability close to one.

One can take $K_n(x, u) = \frac{1 - F_n(u)}{1 - G_n(u) + \eta^*(u)} g_n(u)$ as a kernel function $K(x, u) = \frac{1 - F(u)}{1 - G(u)} g(u)$ in (7.36), where $g_n(u)$ is some estimate of $g(x)$, and $\eta^*(u)$ is determined
NONPARAMETRIC ESTIMATION OF THE HAZARD RATE FUNCTION

by (7.22). The term \( \eta^*(u) \) is used to prevent a zero denominator of the \( K_n(x, u) \), since \( G_n(x) \) may be equal to one at \((a, b)\), when the sample \( Y^n \) occupies an interval less than \([0, d)\).

7.6.1 Failure time detection

Let \( z_t, t = 1, 2, \ldots \), be an observed stochastic process arising from i.i.d. r.v.s with PDF \( f(x) \). We suppose that at time \( \theta \) a PDF \( f(x) \) changes to \( g(x) \) and \( f(x) \neq g(x) \). We need to estimate this moment \( \theta \), called the failure time (Figure 7.5).

Taking into account the independence of \( z_t \), the likelihood function of \( \theta \) is given by

\[
\ell_\theta(z_1, \ldots, z_T) = \left( \prod_{t=1}^{\theta-1} f(z_t) \right) \left( \prod_{t=\theta}^T g(z_t) \right).
\]

Taking logarithms, we get

\[
\ln(\ell_\theta(z_1, \ldots, z_T)) = \sum_{t=1}^{\theta-1} \ln \left( \frac{f(z_t)}{g(z_t)} \right) + \sum_{t=\theta}^T \ln g(z_t).
\]

The maximum likelihood estimate of \( \theta \) is a value that provides the maximum of the first term on the right-hand side of the latter equation:

\[
\theta^* \colon S_\theta = \sum_{t=1}^{\theta-1} \ln \left( \frac{f(z_t)}{g(z_t)} \right) \to \max_{\theta}.
\]

To find \( \theta^* \) we first have to estimate \( q(x) = f(x)/g(x) \) from the equation

\[
\int_a^y q(x) dG(x) = F(y), \quad y \in (a, b), \quad (7.37)
\]

Figure 7.5 Example of a process with a failure time: \( f(x) \in N(0, 1) \) (solid line), \( g(x) \in N(3, 1) \) (dotted line), \( \theta = 50 \).
assuming \( f(x) \) and \( g(x) \) are defined at the finite interval \((a, b)\), \( g(x) \neq 0, x \in (a, b)\). The DFs \( F(x) \) and \( G(x) \) are not known exactly and can be replaced by their empirical estimates. The solution of \( q(x) \) from (7.37) is an ill-posed problem and can be solved by the statistical regularization method (Stefanyuk, 1986).

7.6.2 Hormesis detection

The term ‘hormesis’ is accepted in the modern literature to mean the positive effect on living organisms and constitutions of small doses of toxic substances or stressors, which in larger doses may be unhealthy and destructive. For example, animals preexposed to low-dose radiation may have higher resistance to lethal or sublethal doses (Luckey, 1980). It is well known that high radiation doses may cause illness, but this relationship is not clear for low doses. Of special interest are late effects such as cancer, genetic changes, and congenital malformations. The idea of radiation hormesis arose from the necessity of a natural radiation background for the normal development of organisms, which is ascertained by Planel et al. (1967) from his experiments on fruit flies (\( Drosophila \)). This means that if the dose is zero then the mortality rate is not zero. It is incorrect to extend the case of the mortality risk at large doses to small doses. These are commonly considered to be possible hormetic outcomes of low-level radiation exposure: increased longevity, growth and fertility, a reduction in cancer frequency (Sagan, 1987). The term hormesis may well be applied to any physiological effect which occurs at low doses.

Studies of the longevity and tolerance of large populations of fruit flies after brief exposure to some heat treatment have shown that during the stress time interval the mortality rate of flies in the stress group was the same as in the control group without the stress treatment. Then, after stress, the mortality rate in the stress group was less than in the control group (Khazaeli et al., 1997). Such an effect may not be explained simply by a selection process, that is, by the presence of the heterogeneity in cohorts.

Here, we focus on the problem of hormesis detection in relation to longevity in the case of the possible presence of selection. The lifetimes in both the stress (experimental) group affected by different doses of some stress factor and the control group experiencing standard living conditions (no stress) are observed. To detect the presence of hormesis the ratio between mortalities in the stress and control groups may be useful.

Hitherto, most analysts have used the so-called parametric approach to the analysis of stress data. The approach considered in Sachs et al. (1990) includes Markov cell survival models. Continuous-time Markov chain models provide a powerful formulation for incorporating stochastic effects (such as stochastic fluctuations in the number of cell lesions added by an event of given specific energy) into time-dependent radiation cell survival. To detect hormesis according
to the model, (Yakovlev et al., 1993) the mortality rate is estimated on lifetimes by modeling repair and misrepair processes at the cell level. The capacity of the repair system is estimated as a parameter of the model. It is assumed that the capacity of the repair system of the organism depends on the dose rate. A stepwise reduction of the repair intensity at some nonrandom time instant is suggested. The model is demonstrated on real data, where the decline of the mortality rate for low doses is observed. In the frailty model (Yashin et al., 1996), based on the model of proportional hazards (Cox and Oakes, 1984), a heterogeneity (frailty) stochastic variable is introduced to explain the decline in the mortality rate for low doses of the stress as different reactions of individuals to the same exposure. The problem is that hormesis as well as selection effects may lead to a decline in the mortality rate. The question is how to detect the presence of the hormesis. Yakovlev et al. (1993) indicate the values of parameters showing a strong hormesis effect.

The obvious disadvantage of a parametric approach is that the reliability of the results depends on the correspondence of the model to the empirical data. Another problem is that simple models usually do not fit empirical data well, and when one tries to estimate the parameters for more complex models, in most cases the estimates have large variances because of the lack of data. Sometimes it is more realistic to admit some theoretical properties of an unknown function than to define its parametric form. To detect hormesis in the data, we examine nonparametric estimates of the ratio between the mortality in the group under the stress and the mortality in the control group without the stress, (Markovich, 2000). This ratio may be considered as a dose–effect relationship. Formally, the estimates of this function are obtained as solutions of operator equations. A specific regularization technique for the solution of these equations, which are ill-posed, is described. The estimates are considered for a homogeneous and a heterogeneous population. Models are illustrated on simulated data.

**Evaluation of the mortalities ratio function for homogeneous populations**

Suppose that the individuals in the stress and control cohorts are homogeneous in the sense that they have similar reactions on the same stress factor. Let us assume that we observe two r.v.s $X$ and $Y$. These are individual lifetimes in the stress and control cohorts with PDFs $f(x)$ and $g(x)$ and DFs $F(x)$ and $G(x)$, respectively. $X^n = (X_1, \ldots, X_n)$, $Y^n = (Y_1, \ldots, Y_n)$ are samples of independent observations of these r.v.s on the closed interval $[0, d]$, and $n$ is a sample size. Let us suppose that the distributions are finite or compactly supported. This means that $f(x) = g(x) = 0$, when $x \not\in [0, d]$. Suppose, that $g(x) \neq 0, x \in [0, d]$. Furthermore, suppose that $F(x) \neq 1$ and $G(x) \neq 1$ for $x \in [0, d)$. As in Section 7.6, p. 198, we can consider the mortality ratio function $r(x)$ (7.35) as a solution of the integral equation (7.36).
Evaluation of the mortalities ratio function for heterogeneous populations

It is sometimes assumed that all individuals in the population are heterogeneous, that is, every individual in a population is supposed to have an individual frailty. Several recent studies show that the presence of heterogeneity in the mortality should be taken into account in order to get a better fit to mortality data (Vaupel et al., 1979). Then the function of the ratio of conditional mortality risks \( \mu(x|z) \) for those individuals who have the frailty \( z \) is given by

\[
\frac{r(x|z)}{\mu_2(x|z)} = \frac{\mu_1(x|z)}{\mu_2(x|z)},
\]

(7.38)

Here, we assume that the frailties in both cohorts are identical. Let us suppose that the conditional PDF \( g(x|z) \neq 0 \) for \( x \in [0, d] \) and \( F(x|z) \neq 1 \), \( G(x|z) \neq 1 \) for \( x \in [0, d] \). Let us treat the conditional PDF \( f(x|z) \) as a solution of the equation

\[
\int_0^x \int_0^\infty f(u|z) \varphi(z) dz du = F(x),
\]

(7.39)

where \( \varphi(z) \) is some bounded PDF of the frailty \( z, x \in [0, d] \), (for the degenerate distribution, we are led to the case of the homogeneous population), and \( F(x) \) is unknown. We assume that \( \varphi(z) \) is known (e.g., it is a gamma PDF). We can replace \( F(x) \) by the empirical DF \( F_n(x) \) and determine \( f(x|z) \).

After the estimating \( f(x|z) \) and \( g(x|z) \) from the samples \( X^u \) and \( Y^n \), respectively, one can evaluate the mortality risks for both cohorts using the formulas

\[
\mu_1(x|z) = \frac{f(x|z)}{1 - \int_0^x f(u|z) du}, \quad \mu_2(x|z) = \frac{g(x|z)}{1 - \int_0^x g(u|z) du},
\]

(7.40)

and then determine the ratio of conditional risks from formula (7.38).

Sometimes it is more convenient to estimate the function of the conditional likelihood ratio

\[
g(x|z) = \frac{f(x|z)}{g(x|z)}.
\]

Numerical solution

The integral equations (7.36) and (7.39) can be represented in the operator form (7.2) for the corresponded functions \( g \in U \) and \( y \in V \) (Section 7.2). Let \( U \) and \( V \) be real Hilbert spaces. Then the operator equation (7.2) can be presented as a system of the linear equations (7.14). This system is unstable in view of the inaccuracy of the assignment of empirical data, and so its solution is an ill-posed problem. The solution must be stabilized. The regularized solution \( \alpha \) can be found as a global minimum in \( U \) of the regularizing functional (7.15) for a fixed value \( \gamma > 0 \) of the regularization parameter. For the Hilbert spaces \( U \) and \( V \) the estimate of the vector
\( \alpha \) is determined by formula (7.16). We will define the matrix \( A_n \) for (7.36) and (7.39).

Let us consider (7.36). We will represent the unknown function \( r(t) \) as a linear combination of \( N \) known functions,

\[
r(t) = \sum_{k=1}^{N} \alpha_k \psi_k(t),
\]

where \( \{\psi_k(t), k = 1, \ldots, N\} \) are (e.g., trigonometric or Laguerre) polynomials, and \( \{\alpha_k, k = 1, \ldots, N\} \) are unknown parameters. Substituting (7.41) into (7.36), we can obtain the elements of the matrix \( A_n \) from the samples \( X^n \) and \( Y^n \):

\[
a^n_{i,k} = \int_{0}^{x_i} \frac{1 - F_n(u)}{1 - G_n(u) + \eta^*(u)} \psi_k(u) g_n(u) \, du, \quad i = 1, \ldots, n; \quad k = 1, \ldots, N,
\]

where \( g_n(x) \) is an estimate of the PDF \( g(x) \). We define \( Y_n \) as an \( n \times 1 \) random vector \( (F_n(X_1), \ldots, F_n(X_n))^T \).

Turning to (7.39), by analogy with the proportional hazards model of Cox and Oakes (1984), we represent the unknown PDF \( f(t \mid z) \) as a linear combination

\[
f(t \mid z) = z \sum_{k=1}^{N} \alpha_k \psi_k(t).
\]

Substituting (7.42) into (7.39), we can obtain elements of the matrix \( A_n \) from the sample \( X^n \):

\[
a_{i,k} = \int_{0}^{x_i} \psi_k(u) \int_{0}^{\infty} z \varphi(z) \, dz \, du.
\]

In this case, too, \( Y_n = (F_n(X_1), \ldots, F_n(X_n))^T \). From (7.40) and (7.42) we have

\[
\mu_1(x \mid z) = \frac{z \sum_{k=1}^{N} \alpha_k \psi_k(x)}{1 - z \sum_{k=1}^{N} \alpha_k \int_{0}^{x} \psi_k(u) \, du},
\]

In a similar manner,

\[
\mu_2(x \mid z) = \frac{z \sum_{k=1}^{N} \beta_k \psi_k(x)}{1 - z \sum_{k=1}^{N} \beta_k \int_{0}^{x} \psi_k(u) \, du},
\]

where \( \{\beta_k, k = 1 \ldots N\} \) are built up from the sample \( Y^n \). Then,

\[
r(x \mid z) = \frac{(1 - z \sum_{k=1}^{N} \beta_k \int_{0}^{x} \psi_k(u) \, du) \sum_{k=1}^{N} \alpha_k \psi_k(x)}{(1 - z \sum_{k=1}^{N} \alpha_k \int_{0}^{x} \psi_k(u) \, du) \sum_{k=1}^{N} \beta_k \psi_k(x)}.
\]
Application

The estimate of the mortality ratio function is applied to detect hormesis in the possible presence of selection, which is caused by heterogeneity in the population. Two examples illustrate the application of the regularization method to simulated data. One of these is the estimation of the mortality ratio

$$r(x) = \frac{\mu_s(x)}{\mu_c(x)} \quad \text{based on (7.36).}$$

Here, $\mu_s(x)$ and $\mu_c(x)$ are mortalities observed in the stress and control groups, respectively. The other is the estimation of the conditional mortality in the stress group $\mu_s(x|z)$ for a fixed frailty $z$, based on (7.39) and (7.40). Using estimates of conditional mortalities in the stress and control groups $\mu_s(x|z)$ and $\mu_c(x|z)$, one can obtain the conditional ratio $r(x|z) = \frac{\mu_s(x|z)}{\mu_c(x|z)}$. All estimates for the stress group are provided for some fixed stress dose rate $m$. The dose may be interpreted as an exposure. In both examples the unknown functions $r(x)$ and $\mu_s(x|z)$ were obtained as linear combinations of Laguerre polynomials (see (7.41), (7.43)) and the unknown coefficients of the expansions were estimated by means of the regularization method. The regularization parameter was selected by the discrepancy method (2.42), based on the Kolmogorov statistic $D_n$.

A question arises: How useful is the mortality ratio function in helping to detect the hormesis effect caused by a certain agent that is small in some sense? Having a set of samples of individual longevities in the stress group $\{X^m_n\}$ for different dose rates $m = \{m_1, \ldots, m_p\}$ and a sample $Y^n$ for the control group, one can obtain $r(x)$ as a solution of (7.36) for every fixed dose. As a result we find a dose–effect dependence $r(x, m)$. In accordance with the amount of available information, one can operate with the conditional mortality risk ratio $r(x|z)$ derived from (7.38)–(7.40) for various doses and obtain the conditional dose–effect dependence $r(x, m|z)$. The likelihood ratios $q(x)$ and $q(x|z)$ can also be used to detect the hormesis.

Figure 7.6 illustrates roughly the behavior of the mortalities in the stress and control groups actually observed at given stress doses (Khazaeli et al., 1997; Yashin et al., 1996). Let the stress age interval be $[x_1, x_2] = [5, 10]$. To obtain these dependencies some functions of the dose rate and age were used. We are interested in relatively small stress doses when the debilitation effect is not observed, that is, the mortality risk does not increase, (Yashin et al., 1996).

One should distinguish the behaviors on the stress interval $[x_1, x_2]$ and after this interval, (Yakovlev et al., 1993; Sachs et al., 1990). In view of the interrelation between the biochemical mechanisms of hormesis and selection (Feinendegen et al., 1988), it seems more reasonable to try to establish exact boundaries of the dose interval beyond which hormesis cannot be observed. It is more realistic to point out an approximate dose interval that corresponds to the presence of hormesis and also to determine ‘how far’ the observable process is from ‘pure’ hormesis (without the presence of selection). Figure 7.7 plots the ratio $r(x, m)$ of mortality risks during and after stress. Obviously, if the dose $m = 0$, then the mortalities in the stress and control groups are the same and $r(x, m) = 1$. The next dose interval $[0, m_{ph}] = [0, 1]$ corresponds to ‘pure hormesis’ without selection. This means that during a stress
Figure 7.6  Model mortality risks (left) and requisite PDFs (right) in the stress group and control group at various stresses: control group (solid line with crosses); stress group at doses 0.5 (dotted line), 2.5 (dashed line), and 4.2 (solid line). Reprinted from Automation and Remote Control, 61(1), Part 2, pp. 133–143, Detection of hormesis by empirical data as an ill-posed problem, Markovich NM, Figure 1. © 2000. With permission from Pleiades Publishing Inc.

Figure 7.7  Model ratio \( r(x, m) = \mu_\omega(x, m)/\mu_\omega(x, m) \) against stress dose \( m \): (left) during stress at fixed age \( x = 7 \); (right) after stress at fixed age \( x = 12 \). Reprinted from Automation and Remote Control, 61(1), Part 2, pp. 133–143, Detection of hormesis by empirical data as an ill-posed problem, Markovich NM, Figure 2. © 2000. With permission from Pleiades Publishing Inc.

time interval the mortality in the stress group will be the same as that in the control group and \( r(x, m) = 1 \), but after the stress, the mortality risk in the stress group will be less than that in the control group, and so \( r(x, m) < 1 \) (see the curve in Figure 7.6 that corresponds to the dose 0.5). The dose interval \([m_{ph}, m_{sh}] = [1, 4]\) corresponds to the case where selection and hormesis are present simultaneously (see the curve in Figure 7.6 that corresponds to the dose 2.5). The last dose interval,
[msh, mi] = [4, 5], corresponds to the increasing selection process and the lack of hormesis (see the curve in Figure 7.6 that corresponds to the dose 4.2). In this case more frail individuals die in the stress interval and \( r(x, m) > 1 \) for \( x \in [x_1, x_2] \). As a result, the mortality risk of the surviving individuals in the stress group decreases, but still remains higher than that in the case of the presence of hormesis and selection simultaneously, and then \( r(x, m) \) increases for \( x > x_2 \). Thus, \( r(x, m) \) increases during and after the stress in view of the absence of advantages gained from hormesis. Using \( r(x, m) \), we can determine the dose bounds of the presence of hormesis. If \( r(x, m) \approx 1 \) during the stress and \( r(x, m) < 1 \) after the stress, then we may conclude that the empirical data contain pure hormesis. If \( r(x, m) > 1 \) during the stress and \( r(x, m) \) decreases after the stress \( (r(x, m) < 1) \), then hormesis and selection take place simultaneously. If \( r(x, m) > 1 \) during the stress and \( r(x, m) \) grows after the stress for some \( m > m_{sh} \), then hormesis is absent and only selection occurs.

A similar analysis can be carried out using the likelihood ratio function. Furthermore, estimates of \( r(x) \) and \( \mu_s(x|z) \) on simulated data can be obtained. To demonstrate this we generated the samples \( X^n \) and \( Y^n \) of lifetimes in stress and control groups, respectively, with sample size \( n = 10\,000 \), corresponding to the mortality rates in Figure 7.6. The number of basic functions of the approximation (7.41) is \( N = 8 \). Figure 7.8 illustrates the simulated mortality risk \( \mu_s(x|z) \) and ratio \( r(x) \), corresponding to the stress group given dose \( m = 2.5 \); as well as the corresponding mortalities and ratio, obtained on generated data using the Kaplan–Meier estimator (Cox and Oakes, 1984) and regularized estimates of \( r(x) \) and \( \mu_s(x|z) \), obtained for two frailties. To increase the accuracy, the estimates during the stress interval and after this interval are obtained separately.

![Figure 7.8](image-url)  
**Figure 7.8** Left: Estimates of \( \mu_s(x/z) \) for frailties \( z \in \{1.25; 1.5\} \) (dotted line and dashed line, respectively), obtained from formula (7.43); model (solid line) and generated mortality risk \( \mu_s(x) \) in the stress group at stress \( m = 2.5 \). Right: Estimated (solid line), generated (dotted line), and model ratio \( r(x) = \mu_s(x)/\mu_c(x) \) of mortality risks for the stress group at stress \( m = 2.5 \). Reprinted from *Automation and Remote Control*, 61(1), Part 2, pp. 133–143, Detection of hormesis by empirical data as an ill-posed problem, Markovich NM, Figures 3 and 4. © 2000. With permission from Pleiades Publishing Inc.
7.7 Hazard rate estimation in teletraffic theory

7.7.1 Teletraffic processes at the packet level

Figure 7.9 shows the packet dynamics in a packet-switched communication network with routers as switching entities. Here the packet arrival process at a buffer in front of a transmission link in a router that is modeled as server can most simply be described by a Poisson process. Its intensity $\lambda(t)$ varies randomly over time, or can be taken to be constant if the arrival process does not vary strongly. It can be modeled as Markov-modulated Poisson process if the intensity is a stepwise random function, or as a more general semi-Markov process. $N$ is the size of the data buffer. The buffer is described as ‘free’ if there are spare places in the buffer. If the buffer is full a new arriving packet is rejected with intensity $\mu(t)$, otherwise the packet is accepted and placed in the buffer. The server picks up packets from the buffer at rate $\eta(t)$ and then clears the buffer for new packets.

Our problem is to estimate the parameters of packet dynamics from empirical data. The following sources of information may be available:

- the time intervals between consecutive arrivals of packets, called ‘inter-arrivals’ (Figure 7.10);
- the number of packets arriving in intervals of fixed time length;
- on- and off-periods of traffic generation (Figure 7.10);
- durations of full and free buffer regimes;
- the number of lost packets in intervals of fixed time length.

Here and in Chapter 8 we use empirical data to estimate the parameters of an arrival process (the intensity of a nonhomogeneous Poisson process, the renewal function of a renewal arrival process, and the heavy-tailed distribution of inter-arrivals between packets or of on-and off-periods), the capacity of a buffer (the buffer overload rate, the mean duration of a full buffer), the loss process (the risk

![Figure 7.9 Integrated system of packet dynamics.](image)
of packet losses, the mean duration of operation without losses) and the capacity of the server.

7.7.2 Estimation of the intensity of a nonhomogeneous Poisson process

First we consider a nonhomogeneous Poisson process \( \{ N(t), t \geq 0 \} \) with intensity function \( \lambda(t), t \geq 0 \), as a model of an arrival process of events, such as calls in a BISDN or packets in an IP network. The stochastic process \( N(t) \) counts the random number of arrivals before time \( t \) provided that the procedure was started at \( t = 0 \) with \( N(0) = 0 \).

Since the arrival stream is nonstationary, the probability

\[
p_i(s, t) = P\{ N(t+s) - N(t) = i \}
\]

of obtaining \( i \) arrivals depends on the length \( s \) of the observation interval as well as on its starting point \( t \). It is defined by the Poisson distribution

\[
p_i(s, t) = \frac{(\Lambda(s, t))^i}{i!} \exp(-\Lambda(s, t)),
\]

for any \( i \geq 0 \), where

\[
\Lambda(s, t) = E \{ N(t+s) - N(t) \} = \sum_{i=1}^{\infty} i p_i(s, t)
\]

is the mean number of arrivals occurring in an observation period of length \( s \) starting at \( t \). By Gnedenko and Kowalenko (1971, p. 85) we have

\[
\int_t^{t+s} \lambda(u) \, du = \Lambda(s, t).
\]
By definition, the intensity function $\lambda(t)$ is equal to the mean number of arrivals per time unit in an infinitesimally small interval $[t, t+s)$, $s \searrow 0$, if the derivative of $\Lambda(s, t)$ exists. In practice $\lambda(t)$ often changes periodically or may even strongly increase. However, one just counts the number of arrivals in fixed time intervals. In this case we are interested in estimating $\lambda(t)$ within these intervals using the available empirical data. For this purpose, we subsequently discuss different estimation procedures.

First, let us note that the parametric estimates could be useful here to forecast the behavior of $\lambda(t)$. To estimate the parameters the maximum likelihood method may be applied. However, since the parametric form of $\lambda(t)$ is usually not available, one has to use a nonparametric approach.

Let $\Delta_i = [t + (i-1)\Delta, t+i\Delta), i = 1, \ldots, m,$ denote disjoint subintervals of equal length $\Delta$ covering the finite observation interval $[t, t+s)$. Körner and Nyberg (1993) estimated the arrival rate from the data of one experiment by

$$\hat{\lambda}(t) = \frac{\sum_{i=1}^{m} k_i}{\Delta_i} I(t \in \Delta_i), \quad (7.44)$$

where $k_i$ is the number of arrivals in an interval $\Delta_i$ and $I(t \in \Delta_i)$ is the indicator function of the $i$th subinterval $\Delta_i$. However, the error of this estimate may be very large, since the estimate is constructed by only one random experiment.

Let us consider $l$ independent observations of the process in time intervals of fixed length $s$ with the same starting point $t$, on which the behavior of the observed process is similar. For example, due to the periodicity of the call or session arrival processes in a BISDN or IP network, one may count the number of arrivals over a period of $l$ days within the same fixed time interval $[t, t+s)$. Let $X_i, i = 1, \ldots, m$, denote the generic discrete r.v. counting the observed number of arrivals in each interval $\Delta_i$, $X_i = 0, 1, 2, \ldots$. Let $X^l_i = \{X_{i,1}, \ldots, X_{i,l}\}$ be the sample of independent observations of this r.v. and $l$ be the sample size. The ML estimate is given by

$$\hat{\lambda}(t) = \sum_{i=1}^{m} \frac{X_i}{\Delta_i} I(t \in \Delta_i), \quad (7.45)$$

where the mean values

$$\bar{X}_i = \frac{1}{l} \sum_{k=1}^{l} X_{i,k}$$

are obtained by the observations $\{X_{i,k} : k = 1, \ldots, l\}$ within the interval of the length $s = \Delta_i$, (Markovitch and Krieger, 1999). The latter estimate is unbiased, consistent and asymptotically normal. The accuracy of the estimate depends strongly on the choice of the subintervals $\Delta_i$. 

7.8 Semi-Markov modeling in teletraffic engineering

7.8.1 The Gilbert–Elliott model

Let us describe the packet arrival and buffer overload process by a two-state Gilbert–Elliott (GE) semi-Markov model (Figure 7.11), where the states are denoted by $G$ and $B$, and $g(t)$ and $b(t)$ are respectively the rates of transition from state $G$ to state $B$ and from state $B$ to state $G$. The states $G$ and $B$ correspond to ‘good’ and ‘bad’ regimes with a low and high error probability, respectively. That is to say, $G$ denotes a buffer with free spaces and $B$ denotes a full buffer with no free spaces for new incoming packets. If the buffer is full then all new packets are lost and we will consider them as system errors. Consequently, $g(t)$ is the intensity with which the buffer is filled and reflects the stream of requests, and $b(t)$ the intensity with which the buffer content is cleared or restored and reflects the capacity of the server.

Other interpretations of the GE model are possible. We may consider the generation of packets by a source in a high-speed network. We suppose that $G$ corresponds to the on-period and $B$ to the off-period. Consequently, $g(t)$ and $b(t)$ reflect the rates of switching between these two regimes.

Usually, the transition rates between states in a GE model are assumed to be constant (Ohta and Kitani, 1990; Bratt, 1994). In our study, however, we consider an inhomogeneous Markov model, where the transition rates depend on time. The model has a semi-Markov property since these rates depend on the time spent in states $G$ and $B$. For example, the longer the buffer is free, the greater the probability of filling it.

Let us describe different approaches to identifying a two-state semi-Markov model from empirical data. Suppose that we observe two r.v.s $T_G$ and $T_B$. Commonly these are the durations of the ‘good’ and ‘bad’ regimes of the model – say, of a transmission channel – or the times of the transitions from $G$ to $B$ or from $B$ to $G$. The r.v.s have PDFs $f_G(x)$ and $f_B(x)$ and DFs $P\{T_G < x\}$ and $P\{T_B < x\}$, respectively. $T^n_G = (T^n_G, T^n_G, \ldots, T^n_G)$ and $T^n_B = (T^n_B, T^n_B, \ldots, T^n_B)$ are samples of independent observations of these r.v.s.

The probability of transition from $G$ to $B$ in the time interval $[\tau, \tau + d\tau]$ for small values of $d\tau$ is equal to $g(\tau) P_G(\tau) d\tau$. Then the transition probability in the time interval $[0, x]$ is

$$\int_0^x g(\tau) P_G(\tau) d\tau = P_{GB}(x), \quad (7.46)$$

![Figure 7.11 Gilbert–Elliott semi-Markov model.](image-url)
where \( P_{GB}(x) = P \{ T_G < x \} \) and \( T_G \) is the time of the transition from \( G \) to \( B \) or the length of time spent in state \( G \). Similarly, the probability of the transition from \( B \) to \( G \) in the time interval \([0, x]\) is

\[
\int_0^x b(\tau) P_B(\tau) \, d\tau = P_{BG}(x), \quad (7.47)
\]

where \( P_{BG}(x) = P \{ T_B < x \} \) and \( T_B \) is the time of the transition from \( B \) to \( G \) or the duration of being in state \( B \). Note that \( P_G(x) = P \{ T_G \geq x \} = 1 - P_{GB}(x) \), \( P_B(x) = P \{ T_B \geq x \} = 1 - P_{BG}(x) \) are the probabilities of being in the ‘good’ and ‘bad’ states, respectively.

Let us treat \( g(t) \) and \( b(t) \) as solutions of Volterra’s integral equations (7.46) and (7.47). The problem is that the DFs \( P_{GB}(x) \) and \( P_{BG}(x) \) as well as the probabilities \( P_G(x) \) and \( P_B(x) \) are unknown. But there are samples of independent observations of the r.v.s \( T^n_G \) and \( T^n_B \). Therefore, one may replace \( P_{GB}(x) \) and \( P_{BG}(x) \) by the empirical DFs

\[
P^n_{GB}(x) = \frac{1}{n} \sum_{i=1}^{n} \theta(x - T^n_G), \quad P^n_{BG}(x) = \frac{1}{n} \sum_{i=1}^{n} \theta(x - T^n_B)
\]

constructed from the samples \( T^n_G \) and \( T^n_B \). By the Glivenko–Cantelli theorem the empirical DF converges to the original DF with probability one:

\[
P \left\{ \sup_x |P_{GB}(x) - P^n_{GB}(x)| \to 0 \right\} = 1, \quad n \to \infty
\]

\[
P \left\{ \sup_x |P_{BG}(x) - P^n_{BG}(x)| \to 0 \right\} = 1.
\]

This means that for some fixed \( n \) the right-hand sides and kernel functions of (7.46) and (7.47) are not precisely known. Thus, one can only obtain approximate solutions \( \hat{g}(t) \) and \( \hat{b}(t) \) rather than the exact functions \( g(t) \) and \( b(t) \). Since \( g(t) \) and \( b(t) \) are hazard rates, we can treat the solution of (7.46) and (7.47) as a solution of integral equation (7.20) and apply the regularization method to estimate the hazard rate (see p. 189). Then one can estimate the PDF of the transition time to the ‘bad’ state from the formula

\[
\hat{f}_{GB}(x) = \hat{g}(x) (1 - P^n_{GB}(x))
\]

and the PDF of the transition time to the ‘good’ state from

\[
\hat{f}_{BG}(x) = \hat{b}(x) (1 - P^n_{BG}(x)).
\]

The estimation of PDFs \( f_{GB}(x) \) and \( f_{BG}(x) \) may also be treated as the solution of Fredholm’s equation (2.8), using the data \( T^n_G \) and \( T^n_B \).
The mean time spent in the ‘good’ state – for example, error-free operation – in the time interval $[0, d]$ is given by

$$T_G = \int_0^d P_G(x) \, dx.$$  

The mean time spent in the ‘bad’ state – characterized by packet loss – is determined by

$$T_B = \int_0^d P_B(x) \, dx.$$  

The r.v.s $T_G$ and $T_B$ may be heavy-tailed distributed. On/off-periods are an example of such r.v.s. To estimate the transition rates $g(t)$ and $b(t)$ in this case one can transform the data to a compact interval (Section 7.4).

### 7.8.2 Estimation of a retrial process

Retrial queues are characterized by the following feature: if an arriving call finds the server free, it immediately occupies the server and leaves the system after service completion. But if the server is busy then this blocked call becomes a potential repeated call. The customer may repeat his request after some random delay. This situation plays a special role in several computer and communication networks. Many papers have been devoted to retrial queues (an extensive survey can be found in Falin, 1990), but relatively little is known about the statistics of this problem. Falin (1995) gives an estimate of the retrial rate and its asymptotic variance when the observation period is long. The difficulty of estimating the retrial rate is due to the lack of available empirical information:

- In most cases we cannot distinguish primary and repeated calls.
- Retrial queues cannot be fully observed (in particular, it is difficult in practice to observe customers in orbit, that is, abandoned customers who are making another attempt to get service).

We always observe

- a joint arrival flow of primary and repeated calls to the servers, i.e. the number of incoming calls (attempts) in a unit period of time;
- holding times of calls (a mean per unit time);
- the number of rejected calls (a mean per unit time).

Additionally, more detailed information might be measured:
• time spent in orbit,
• times between two consecutive attempts to get service or occupy the same line after the first abandonment (inter-attempts).

The latter information could be available from interviews. The time spent in orbit is the time interval between the time of the first case abandonment and the time of the last abandonment (the last unsuccessful attempt to get service) or of the first successful call of the customer.

If inter-attempts of several customers are available one can calculate the retrial rate \( \frac{r(t)}{\text{orbit}} \) directly by one of the methods for the estimation of the hazard rate; see, for example, Section 7.5.1 or Kooperberg et al. (1994). Normally, such information is not available.

Due to the lack of information, we elaborate on a probabilistic approach based on the description of a call process at the individual level which progresses in a random jump-like manner and possesses the Markov property. A class of compartmental semi-Markov models is considered for the modeling of stochastic changes of customer states. The main feature of this class is that models of different complexity can be generated from the simpler models by the addition of new states. The semi-Markov property of the models is important since it reflects the dependence of the transition rates between different states of the sojourn times in the states. For example, the retrial rate depends on the length of time spent in orbit: the longer one is waiting for service, the more impatient one becomes.

The identification of semi-Markov models uses estimates of transition rates between different states. The problem of model identification is considered in terms of the solution of an integral equation system on the basis of empirical data. The analytical expressions for the probabilities of being in different states are presented in the form of integral relationships.

The structure of the model depends on the available empirical data. To construct semi-Markov models it is important to be clear as to what a retrial call is. One may define a retrial call as the attempt of some customer to occupy a fixed channel (or to get service) within a limited time period without attempting to occupy any other channel (definition A). But within the waiting time, the customer may occupy another channel and then return to orbit, that is, restart the attempt to occupy the previous channel (definition B). These different understandings of the retrial call lead to different estimates of the retrial rate.

**First Semi-Markov model**

First, let us accept definition A and formulate the simplest compartment model of the retrial process (see Figure 7.12). This model has three states: ‘server’, ‘orbit’, and ‘out’. Calls in the server state are receiving service. Calls in the out state have either been serviced or are waiting to make their first attempt to get service. The orbit state contains abandoned customers. The rate of the transition from the server
state to the out state is the service rate $\mu(t)$. Abandoned customers are characterized by the rejection rate $\mu_0(t)$, that is, the transition rate from the server state to the orbit state. The arrival rate of primary calls is given by $\lambda_1(t)$ and corresponds to the transition rate from the out state to the server state. The transition rate $r(t)$ from the orbit state to the server state is the intensity of repeated calls. The total arrival rate $\lambda(t)$ is given by

$$\lambda(t) = \lambda_1(t) + r(t).$$

(7.48)

The customer may leave orbit, that is, pull out of the next attempt to get service and return to the out state at rate $h(t)$ – this function thus characterizes the impatience of customers and the abandonment of retrial attempts.

Let the functions $P_1(t)$, $P_2(t)$ and $P_3(t)$ stand for the probabilities of observing a customer at time instant $t$ in the server, out, and orbit states, respectively. Note that

$$P_1(t) + P_2(t) + P_3(t) = 1.$$  

(7.49)

Let us consider the structures of these probabilities in more detail. We start with $P_3(t)$. Let $u(t, z)$ denote the PDF relating to a customer in the orbit state at time $t$ with waiting duration from $z$ to $(z + dz)$.

Considering the transitions from server to orbit, from orbit to server, and from orbit to out, we can write

$$u(t, z) = P_1(t - z)\mu_0(t - z)S_r(t, z)S_h(t, z),$$

where

$$S_r(t, z) = \exp \left( - \int_{t-z}^{t} r(u)du \right), \quad S_h(t, z) = \exp \left( - \int_{t-z}^{t} h(u)du \right)$$

are the probabilities that a customer does not pass to the server and out states in the time interval $(t - z, t)$.
The probability \( P_3(t) \) of finding a customer in orbit at time \( t \) is equal to the integral of the function \( u(t, t - y) \) with respect to \( y \). Here \( y \) is the time instant when the transition occurs from the server state to the orbit state:

\[
P_3(t) = \int_0^t \mu_0(y)P_1(y)S_r(t, t - y)S_s(t, t - y)dy.
\] (7.50)

The probabilities \( P_1(t) \) and \( P_2(t) \) satisfy the differential equations

\[
P_1'(t) = -\left( \mu(t) + \mu_0(t) \right)P_1(t) + P_2(t)\lambda_1(t) + P_3(t)r(t),
\] (7.51)

\[
P_2'(t) = -\lambda_1(t)P_2(t) + P_1(t)\mu(t) + P_3(t)h(t).
\] (7.52)

Equations (7.49)–(7.52) form a system that describes the dynamics of the changes in the probabilities of sojourn in different states.

The transition rates \( \mu(t) \), \( \lambda(t) \) might be estimated using empirical data, namely from holding times and the number of incoming calls, respectively. The rates \( h(t) \) and \( \mu_0(t) \) may be estimated qualitatively as a proportion of leaving or rejected customers (e.g., 20% and 3%) or from the lengths of time spent in orbit and the number of rejected calls, respectively (if the latter data are available). Then by (7.48)–(7.52) one can estimate the unknown intensities \( \lambda_1(t) \) and \( r(t) \), and thus separate the primary and repeated call rates.

In the simplest case, one may assume that the holding times are exponentially distributed, and the total arrival stream is a nonhomogeneous Poisson stream. Then \( \mu(t) = 1/\overline{T_h} \) for any \( t \), where \( \overline{T_h} \) is the mean holding time, and one can estimate the intensity of the nonhomogeneous Poisson process \( \lambda(t) \) by formula (7.45).

One can solve equations (7.51) and (7.52) using the following initial conditions:

\[
P_1(0) = 0, \quad P_2(0) = 1.
\]

Hence

\[
P_1(t) = \int_0^t \left( \lambda_1(y)P_2(y) + r(y)P_3(y) \right) \exp \left( -\int_y^t (\mu(u) + \mu_0(u)) \, du \right) \, dy,
\]

\[
P_2(t) = \int_0^t \left( \mu(y)P_1(y) + h(y)P_3(y) \right) \exp \left( -\int_y^t \lambda_1(u) \, du \right) \, dy
\]

\[+ \exp \left( -\int_0^t \lambda_1(u) \, du \right) .\]

Equation (7.50) describes the probability that a customer is in the orbit state in time \( t \). The probability \( P_3(t) \) may be estimated as the proportion of customers in orbit.

In practice, the number of rejected calls per unit time is available. The probability of the transition from the server state to the orbit state in the interval \([\tau, \tau + d\tau]\) for small values of \( d\tau \) is equal to \( \mu_0(\tau)P_1(\tau)d\tau \). Thus, the probability of the transition to the orbit state or the abandonment of the call in the time interval \([t_0, t]\) is

\[
P^*_{\tau}(t_0, t) = \int_{t_0}^t \mu_0(\tau)P_1(\tau) \, d\tau.
\]
If $N_t$ is the total number of customers at time $t$, then the number of rejected calls in time interval $[t_0, t]$ is determined by

$$m(t_0, t) = N_t P^*(t_0, t).$$

To estimate $P_1(t)$ we can now write the equation

$$\frac{m(t_0, t)}{N_t} = \int_{t_0}^{t} \mu_0(\tau) P_1(\tau) d\tau$$

and use it instead of (7.50). A preliminary estimate of $\mu_0(\tau)$ is required. Note that $\mu_0(\tau)$ characterizes the probability of the customer being abandoned in time interval $[t, t + \Delta t]$ under the assumption that he was not abandoned before $t$. For small $\Delta t$,

$$\mu_0(t) \Delta t \approx \frac{m(t, t + \Delta t)}{N^{*}_t},$$

where $m(t, t + \Delta t)$ is the number of rejections in $[t, t + \Delta t]$, and $N^{*}_t$ is the number of calls arriving up to time $t$.

**Remark 12** The length of time spent in orbit may have a heavy-tailed distribution. Hence, to estimate the rate $h(t)$ one can apply the transformation approach (Section 7.4). The total arrival stream may be a renewal process. In particular, its inter-arrival time distribution may be heavy-tailed. To estimate $\lambda(t)$ it may then be useful to apply the methodology described in Chapter 8.

**Second semi-Markov model**

Let us now accept definition B. Figure 7.13 shows the semi-Markov model for the retrial call process in this case. This model has the same states as the first one, but some of the transition rates are in principle different. As earlier, $\mu(t), \lambda_1(t), \mu_0(t), h(t)$ correspond to the service rate, the primary call arrival rate, the rejection rate, and customer impatience, respectively. Now, however, one may occupy another line or use another service while waiting in orbit. This means that the transition rate from the orbit state to the server state combines the primary call and repeated call rates. Therefore, one may transit from the orbit state to the server state at rate $\lambda(t)$, satisfying (7.48). After the successful or not successful call one may return, that is, resume the attempt to occupy the former line or access service. This means that one may move from the out state to the orbit state at rate $\mu_1(t)$.

Let us now write the system of the equations describing the dynamics of this setup. Let $P_1(t), P_2(t), P_3(t)$ again be the probabilities of being in the server, out, and orbit states. Then the following equations hold:

\[
\dot{P}_1(t) = -(\mu(t) + \mu_0(t)) P_1(t) + P_2(t) \lambda_1(t) + P_3(t) \lambda(t),
\]

\[
\dot{P}_2(t) = -(\lambda_1(t) + \mu_1(t)) P_2(t) + P_1(t) \mu(t) + P_3(t) h(t),
\]

\[
\dot{P}_3(t) = -(\lambda(t) + h(t)) P_3(t) + P_1(t) \mu_0(t) + P_2(t) \mu_1(t).
\]
To identify this semi-Markov model one needs some more detailed empirical information. The transition rate $\mu_1(t)$ might be estimated quantitatively as the proportion of customers restarting their attempts to get previous service in the time intervals between the end of the successful or unsuccessful call and the attempt to occupy the previous service, if they are available. The rates $\mu(t)$, $\lambda(t)$, $\mu_0(t)$, $h(t)$ are calculated as described earlier. The description becomes more complicated than the first model, since $\mu_1(t)$ depends on the holding times and is restricted by the potential waiting time of the customer. Equations (7.48), (7.49), (7.53) provide the estimation of the retrial rate $r(t)$ as well as the rate of primary calls $\lambda_1(t)$.

**Remark 13** The solutions of Volterra’s integral equations, using the first and second Semi-Markov retrial call models, are extremely sensitive to noise in the empirical data. The solutions must be stabilized as discussed in Section 7.3.

### 7.9 Exercises

1. Rough hazard rate estimation.

   Generate $X^n$ according to Pareto distribution (4.8). Estimate the hazard rate function
   
   $$h(x) = \frac{f(x)}{(1 - F(x))},$$
   
   where $f(x)$ is the PDF, and $F(x)$ is the DF of the underlying r.v., by following procedures:

   - (a) replacing $f(x)$ by a histogram and $F(x)$ by an empirical DF $F_n(x)$ of the sample $X^n$ for $x < X_{(n)}$ (for $x \geq X_{(n)}$, $F_n(x) = 1$ and the estimate $\hat{h}(x)$ of the hazard rate is infinite);

   - (b) by the transformation of $X^n$ to $Y^n$, $Y_i = T(X_i)$, $i = 1, \ldots, n$. The transformation is given by $T = 1 - (1 + \hat{\gamma}x)^{-1/(2\hat{\gamma})}$. Calculate the estimate of the EVI $\hat{\gamma}$ by Hill’s estimator (1.5).

   Calculate the hazard rate by formula (7.18), where the hazard rate $h^*(x)$ of a new r.v. $Y_1 = T(X_1)$ is estimated by the method indicated in (a).
2. Hazard rate estimation from (7.21) by a regularization method. Generate \( X^n \) according to Pareto distribution (4.8). Transform \( X^n \) to \( Y^n \), \( Y_i = T(X_i), i = 1, \ldots, n \) by transformation \( T = 1 - (1 + \hat{\gamma} x)^{-1/(2\gamma)} \). Calculate the estimate of the EVI \( \hat{\gamma} \) by Hill’s estimator (1.5). Estimate the hazard rate \( h^\delta(x) \) by \( Y^n \) from the equation

\[
\int_0^t h^\delta(x) \, dx = -\ln(1 - G(t)), \quad t \in [0, 1],
\]

by a regularization method. Determine the estimate by \( \hat{h}^\delta(x) = \sum_{k=1}^{N} \alpha_k^\gamma \psi_k(x) \), where \( \{\psi_k(x) = \sqrt{2} \cos(\pi k x), k = 1, 2, \ldots\} \), \( \alpha_k^\gamma = (\gamma I + A_n^T A_n)^{-1} A_n^T z_n \), the regularization parameter \( \gamma \in \{0.01, 0.1, 0.5\} \), \( z_n \) is the vector \( \{-\ln(1 - G_n(Y_1)), \ldots, -\ln(1 - G_n(Y_n))\} \). \( G_n(x) \) is an empirical DF with respect to \( Y^n \), the elements of matrix \( A_n \) are defined as \( a_{i,j} = \int_{0}^{Y_i} \psi_j(x) \, dx, i = 1, \ldots, n, j = 1, \ldots, N \). Estimate \( h(x) \) by formula (7.18) and plot it for different values of \( \gamma \).

3. Hazard rate estimation from (7.20) by a regularization method. Repeat Exercise 2 but estimate the hazard rate \( h^\delta(x) \) by \( Y^n \) from the equation

\[
\int_0^t h^\delta(x) (1 - G(x)) \, dx = G(t), \quad t \in [0, 1],
\]

by a regularization method. The difference is that \( z_n = \{G_n(Y_1), \ldots, G_n(Y_n)\} \), and the elements of the matrix \( A_n \) are defined as \( a_{i,j} = \int_{0}^{Y_i} \psi_j(x) (1 - G_n(x)) \, dx \).

4. Estimating the intensity function of a nonhomogeneous Poisson arrival process. Consider \( m \) disjoint subintervals \( \Delta_i = [i - 1, i), i = 1, \ldots, m, m = 24 \) (hours), of equal length \( \Delta = 1 \) covering the observation interval \([0, s], s = 24 \). For each \( \Delta_i \) generate a Poisson distributed random sample of arrivals \( X_i^l = \{X_{i,1}, \ldots, X_{i,l}\}, l = 100 \), with intensity \( \lambda_i \):

\[
\lambda_i = \frac{\exp\left(-\frac{(i-a_1)^2}{2\sigma_1^2}\right)}{2\sqrt{2\pi\sigma_1^2}} + \frac{\exp\left(-\frac{(i-a_2)^2}{2\sigma_2^2}\right)}{2\sqrt{2\pi\sigma_2^2}}, \quad i = 1, \ldots, m,
\]

with parameters \( a_1 = 10, \sigma_1 = 3, a_2 = 21, \sigma_2 = 1 \). that is, the intensity function \( \lambda(t) = \sum_{i=1}^{m} \lambda_i I(t \in \Delta_i) \) is simulated. Estimate \( \lambda(t) \) by formula (7.45) and plot it.
We now consider the estimation of the renewal function (RF) within a finite time interval and for infinite time. A nonparametric histogram-type estimator, its asymptotical properties and smoothing methods are presented.

The chapter is organized as follows. Section 8.1 provides motivation for RF estimation, and in Section 8.2 the function is defined and its approximations for large time intervals are presented. In Section 8.3 the histogram-type estimate of the RF is described. Section 8.4 provides some results (Theorems 23–26) on the almost sure uniform convergence of this estimate to the RF. The selection of the parameter $k$ of this estimate by the bootstrap method (Section 8.5) and by a plot (Section 8.6) is described for different time intervals. Section 8.7 contains a simulation study on the accuracy of the proposed estimate for different inter-arrival-time distributions and different values of $k$ selected by the bootstrap and plot methods. A comparison with Frees’ estimate is given. An application to TCP flow inter-arrival time data is presented in Section 8.8. Further discussion is provided in Section 8.9. The proofs of the theorems are presented in Appendix F.
8.1 Traffic modeling by recurrent marked point processes

Let us consider measurements at the burst and/or packet level. Then the generated Internet traffic can be characterized by a marked point process \( N \equiv \{(\tau_i, Z_i) \mid i = 1, 2, \ldots \} \) of inter-arrival times \( \tau_i \) of bursts (or, in a more general setting, of IP packets) of sizes \( Z_i \). Then \( t_n = \sum_{i=1}^{n} \tau_i, t_0 = 0 \), is the \( n \)th arrival time based on the sampled inter-arrival (or inter-renewal) times \( \{\tau_1, \ldots, \tau_i\} \) with an assumed absolutely continuous DF \( F(x) \). Then the overall volume of IP packets or Web traffic in an observation period \([0, t]\) is determined by

\[
V(t) = \sum_{i:t_i\leq t} Z_i = \sum_{i=1}^{N_t} Z_i,
\]

where \( N_t = \max\{n : t_n < t\} \) is the number of burst (or IP packet) arrivals in \([0, t]\).

Assuming that \( Z_1, Z_2, \ldots \) are i.i.d. r.v.s with \( E|Z_i| < \infty \) and \( E(N_t) < \infty \), we obtain from Wald’s equation that, for all \( t > 0 \),

\[
E(V(t)) = E \left( \sum_{i=1}^{N_t} Z_i \right) = E(Z_i)E(N_t) = E(Z)H(t)
\]

(Trivedi, 1997). Here \( H(t) = E(N_t) \) denotes the RF of the corresponding nonmarked arrival process \( \{\tau_i, i = 1, 2, \ldots\} \) of transferred files and pages, respectively. The definition of the RF is presented in Section 8.2.

For all fixed \( t > 0 \) the variance of the overall volume is determined by:

\[
\text{var}(V(t)) = \text{var}(Z_i)H(t) + (E(Z_i))^2 \text{var}(N_t)
\]

if \( EZ_i^2 < \infty \) holds (Trivedi, 1997). It may be evaluated computationally at any \( t > 0 \) by the estimation of \( H(t) \) and the expectation and variance of \( Z_i \). Hence, a key issue concerns the estimation of the RF for moderate time intervals \([0, t]\). The approximation of the variance \( \text{var}(N_t) \) for large \( t \),

\[
\text{var}(N_t) \approx \frac{\sigma^2}{\mu^3} t + \left\{ \frac{2\sigma^2}{\mu^2} + \frac{3}{4} + \frac{5\sigma^4}{4\mu^4} - \frac{2\mu_3}{3\mu^3} \right\},
\]

where \( \mu, \sigma^2, \) and \( \mu_3 \) are the first three moments of the inter-arrival time distribution, is given by Heyman and Sobel (1982).

We also consider an estimate of the RF using a limited number of independent observations of the inter-arrival times \( \{\tau_1, \tau_2, \ldots, \tau_i\} \) for an unknown inter-arrival-time distribution (ITD). The nonparametric estimate is derived from the representation of the RF as series of DFs of consecutive arrival times using a finite summation and approximations of the latter by empirical DFs. Due to the limited number of observed inter-arrival times the estimate is accurate only for closed time intervals \([0, t]\). An important aspect is determined by the selection of an optimal number of terms \( k \) of the finite sum. Two methods are proposed:
(i) an a priori choice of \( k \) as a function of the sample size \( l \) which provides almost surely the uniform convergence of the estimate to the RF for light- and heavy-tailed ITDs if the time interval is not too large, and

(ii) a data-dependent selection of \( k \) by a bootstrap method and by a plot.

To evaluate both the efficiency of the estimate and the selection methods of \( k \), a Monte Carlo study is carried out.

### 8.2 Introduction to renewal function estimation

Renewal processes have a wide range of applications in warranty control, in the reliability analysis of technical systems and particularly of telecommunication networks such as high-speed packet-switched networks like the Internet. Normally, measurement facilities count the events of interest – the number of requested and transferred Web pages, incoming or outgoing calls, frames, packets or cells in consecutive time intervals of fixed length. It is important for planning and control purposes (e.g., for intrusion detection) to estimate the traffic load in terms of the mean numbers of events counted and their variances in these intervals. In such applications the RF constitutes the basic characteristic of an underlying renewal process since by means of this function the expectation and variance of the number of arrivals of the relevant events before a fixed time instant can be calculated (Gnedenko, 1943; Feller, 1971). To estimate the RF, several realizations of the counting process may be required – for example, the observations of the number of calls over several days. Here we consider the estimation of the RF using inter-arrival times between events for only one realization of the process.

Let \( F(t) = P\{\tau_n < t\} \) with \( F(0+) = 0 \), denote the common DF of the i.i.d. inter-arrival times \( \{\tau_n, n = 1, 2, \ldots\} \) of these events. The renewal counting process \( \{N_t, t \geq 0\} \) denotes the number of events before time \( t \), \( N_t = \max\{n : \tau_n < t\} \) for \( t \geq 0 \), where \( \tau_n = \sum_{i=1}^{n} \tau_i \), \( \tau_0 = 0 \), are the arrival times. The RF \( H(t) \) is expressed by

\[
H(t) = E(N_t) = \sum_{n=1}^{\infty} P\{t_n < t\} = \sum_{n=1}^{\infty} F^{*n}(t),
\]

for \( t \geq 0 \), where \( F^{*n} \) denotes the \( n \)-fold recursive Stieltjes convolution of \( F \).

Several RF estimation methods have been developed for a known ITD. Unfortunately, explicit forms of the RF are obtained only in rare cases, for example, if the inter-arrival times have a uniform distribution, or for the wide class of matrix-exponential distributions (exponential and Erlang distributions belong to this class); see Asmussen (1996). Therefore, several attempts have been made to evaluate the RF computationally; see Chaudhry (1995), Deligönül (1985), McConalogue (1981), and Xie (1989).

In many problems, Smith’s (1954) key renewal theorem may be useful.
Theorem 22 Let the distribution \( F(t) \) be continuous \((F(0) = 0, F(\infty) = 1)\) and \( Q(t) \geq 0 \) be a monotone nonincreasing integrable function on \((0, \infty)\). Then

\[
\lim_{t \to \infty} \int_0^t Q(t - \tau) dH(\tau) = \frac{1}{\mu} \int_0^\infty Q(x) dx.
\]

Example 12 One can get \( H(t) = \lambda t \) for the exponential distribution using \( Q(t) = \theta(t - \tau) \).

We denote the mean inter-arrival time by \( \mu = E(\tau_n) \). If the variance \( \sigma^2 = \text{var}(\tau_n) \) of \( F \) is finite, then, applying Smith’s theorem, the RF \( H(t) \) may be approximated for large \( t \) by the expression

\[
H(t) = \frac{t}{\mu} + \frac{\sigma^2}{2\mu^2} - \frac{1}{2} + o(1),
\]

widely used in the literature. If \( \mu \) is finite, but \( \sigma^2 \) is not finite, then

\[
H(t) \sim \frac{t}{\mu} + G_F(t), \quad t \to \infty,
\]

where

\[
G_F(t) = \frac{1}{\mu^2} \int_0^t \left( \int_0^\infty \left(1 - F(x)\right) dx \right) dy.
\]

Note that \( G_F(t) \to \infty \) as \( t \to \infty \) if and only if \( \sigma^2 = \infty \) (Sgibnev, 1981). For regularly varying distributions \( 1 - F(x) = x^{-\alpha} \ell(x) \) and some \( 1 < \alpha < 2 \), where \( \ell(x) \) is a slowly varying function, it was shown that

\[
H(t) \sim \frac{t}{\mu} + \frac{t^2 (1 - F(t))}{\mu^2(\alpha - 1)(2 - \alpha)}, \quad t \to \infty
\]

(Teugels, 1968). This result has been extended to the case \( 1 < \alpha \leq 2 \),

\[
H(t) - \frac{t}{\mu} \sim \frac{1}{\mu^2} \int_0^t \left( \mu - \int_0^y \left(1 - F(x)\right) dx \right) dy, \quad t \to \infty,
\]

in Mohan (1976).

In Chaudhry (1995) a closed-form expression for the RF is stated if the Laplace–Stieltjes transform \( \overline{f}(s) = \int_0^\infty e^{-st} dF(t), \text{ Res} > 0 \), of \( F(t) \) is a rational function. It is assumed that \( \overline{f}(s) \) may be represented by the ratio

\[
\overline{f}(s) = \frac{P(s)}{Q(s)}
\]

1 The notation \( \sim \) means that the ratio between the two functions of variable \( t \) converges to 1 as \( t \to \infty \).
of two polynomials $P(s)$ and $Q(s)$ of degree $k$ and less than $k$, respectively. Denoting by $s_i, i = 1, 2, \ldots, k$, those roots of the polynomial $Q(s) - P(s)$ with $\text{Re}(s_i) \leq 0$, we obtain

$$H(t) = A_k t + \sum_{i=1}^{k-1} \frac{A_i}{s_i} \exp(s_i t) - \sum_{i=1}^{k-1} \frac{A_i}{s_i},$$

where

$$A_i = \frac{P(s_i)}{Q'(s_i) - P'(s_i)}, \quad i = 1, 2, \ldots, k.$$ 

It can be shown that $A_k = 1/\mu$. For large $t$, one can use the asymptotic expression in terms of roots

$$H(t) \simeq A_k t - \sum_{i=1}^{k-1} \frac{A_i}{s_i}.$$ 

Asymptotic estimates do not perform well for small $t$ relative to $\mu$, which is especially important for the load control of telecommunication systems and the warranty control of devices (Frees, 1986a).

In practice, it is more realistic for the distribution to be unknown or only general information describing it to be available. The estimation of the DF or the PDF, if the latter exists, may become complicated if the distribution of the r.v. is heavy-tailed (Section 3.1). Here we focus on the estimation of the RF with no information on the form of the underlying distribution and we use only a sample $T' = \{\tau_n, n = 1, 2, \ldots, l\}$ of the nonnegative i.i.d. inter-arrival times between events of size $l$. The nonparametric estimate (8.4) is related to a histogram-type estimate where the unknown probabilities $P\{t_n < t\}$ in (8.1) are replaced by the corresponding empirical DFs and a limited number $k$ of terms are used in the summation. A similar nonparametric estimate,

$$H(t, k) = \sum_{n=1}^{k} F^{(n)}_l(t),$$

was proposed by Frees (1986a,b) and further investigated in Schneider et al. (1990). These authors used

$$F^{(n)}_l(t) = \binom{l}{n}^{-1} \sum_{c} \theta(t - (\tau_{i_1} + \ldots + \tau_{i_n}))$$

as an estimate of the arrival-time distribution. Here $\sum_{c}$ denotes the sum over all $\binom{l}{n}$ distinct index combinations $\{i_1, i_2, \ldots, i_n\}$ of length $n$. The $U$-statistic $F^{(n)}_l(t)$

---

2 Matrix-exponential distributions have rational Laplace transforms, whereas the Weibull distribution does not have a closed-form Laplace–Stieltjes transform.

3 The main problem is to estimate the roots accurately.
is a minimum-variance unbiased estimator of $F^{*n}(t)$. In contrast to estimate (8.4), which uses just one combination of adjacent inter-arrival times, the computation of the $H_l(t, k)$ is awkward.

The accuracy of such types of estimates depends on $k$. Frees (1986b) obtained the almost sure uniform consistency of $H_l(t, k)$ on compact intervals $[0, t], 0 \leq t < \infty$, under the assumptions that $k = l$ and $F(t)$ has a positive mean and finite variance, and the asymptotic normality of $H_l(t, k)$ for each fixed point $t > 0$. Under some moment conditions on the r.v. $\min(0, \tau_i)$, the almost sure consistency and the asymptotic normality are proved for real-valued inter-arrivals $\tau_i$ (Frees, 1986b). However, the data-dependent selection of $k$ (which is important for moderate samples) was not considered in Frees (1986a, b, or Schneider et al. (1990). Grübel and Pitts (1993) proved the convergence of a noncomputational empirical RF

$$H_l(t) = \sum_{n=1}^{\infty} \hat{F}^{*n}_l(t)$$

on $R$ as $l \to \infty$. Here, $\hat{F}^{*n}_l(t)$ is the $n$-fold convolution of the empirical DF $F_l(t)$ based on the sample $T^l$.

In our discussion of the estimate (8.4) below, an unbiased estimate of $F^{*n}(t)$ is used, but its variance is not minimal. This inaccuracy is compensated by the data-dependent selection of $k$ and the use of larger samples. An a priori choice of $k$ as a function of the sample size $l$ is considered to obtain almost surely the uniform convergence of the estimate to the RF as $l \to \infty$ for light- and heavy-tailed PDFs. The bootstrap and plot methods are applied for a data-dependent choice of $k$.

### 8.3 Histogram-type estimator of the renewal function

We consider the estimate of the RF $H(t)$ which was first introduced in Markovitch and Krieger (2002b) and investigated in Markovich (2004) and Markovich and Krieger (2006a, b). Let $[r]$ denote the integer part of a real number $r$. Substituting the empirical mean for $E(N_i)$, we replace the DF $P\{t_n < t\}$ by the empirical DF $F_{l_n}(t) = \frac{1}{l_n} \sum_{i=1}^{l_n} \theta(t - t^n_i)$ (an unbiased estimate), where $t^n_i = \sum_{q=1}^{n_i-1} \tau_q$, $i = 1, \ldots, l_n, t^n_n = \left[ \frac{t}{l_n} \right], n = 1, \ldots, k$, are the observations of the r.v. $t_n$. Then we can estimate the renewal function $H(t)$ based on the samples of independent renewal-time observations $t_1 = \{t^1_1, \ldots, t^1_{l_1}\}, \ldots, t_k = \{t^k_1, \ldots, t^k_{l_k}\}$ by

$$\tilde{H}(t, k, l) = \sum_{n=1}^{k} \frac{1}{l_n} \sum_{i=1}^{l_n} \theta(t - t^n_i).$$
Note that \( \tilde{H}(t, k, l) = k \) holds for \( t \in [t_{\text{max}}(k), \infty) \), where \( t_{\text{max}}(k) = \max_{1 \leq n \leq k} \max_{1 \leq i \leq l_n} t_{n,i}^{(i)} \) and \( k \) is some fixed number.

Errors arise in the estimation from both the approximation of \( H(t) \) in (8.4) by a finite sum and the approximation of \( P \{ t_n < t \} \) by the empirical DF \( F_{t_n}(t) \):

\[
\| H(t) - \tilde{H}(t, k, l) \| = \left\| \sum_{n=k+1}^{\infty} P \{ t_n < t \} + \sum_{n=1}^{k} (P \{ t_n < t \} - F_{t_n}(t)) \right\|. \tag{8.5}
\]

From this formula one can see that \( \tilde{H}(t, k, l) \) as well as the estimator (8.2) are biased since \( k \) is limited. A rough upper bound for the bias is given by

\[
\text{bias}(t, k, l) = H(t) - E\tilde{H}(t, k, l) = \sum_{n=k+1}^{\infty} P \{ t_n < t \} \leq \sum_{n=k+1}^{\infty} (F(t))^n = \frac{(F(t))^{k+1}}{1 - F(t)}. \tag{8.6}
\]

For small \( t \), \( F(t) \) is generally small and \( F(t) < 1 \), thus, this error is small.

To provide a good approximation of \( P \{ t_n < t \} \) by the empirical DF, according to the Glivenko–Cantelli theorem sufficiently large values \( l_n \) should be used \((k < l)\). Note that \( l_n = 1 \) for \( 1/2 < k \leq l \), that is, the sample \( t_k \) contains only one point. Therefore, it is reasonable to take \( k \leq l/2 \). In the following, we provide an optimized estimate of \( k \). On the other hand, to provide a good approximation of \( H(t) \) by means of \( \tilde{H}(t, k, l) \) in general, the value of \( k \) should be large enough. Therefore, the estimate \( \tilde{H}(t, k, l) \) is sensitive to the choice of \( k \) and the length of the estimation interval \([0, t]\). Obviously, the estimate \( \tilde{H}(t, k, l) \) may only be accurate within the interval \([0, t_{\text{max}}(k)]\), since the sample size \( l \) is limited.

## 8.4 Convergence of the histogram-type estimator

The convergence of the estimator (8.4) to the RF in the metric of the space \( C \) of continuous functions (Theorems 23–25) was proved in Markovich and Krieger (2006a). To estimate the risk (8.5), one is interested in the \( t \)-regions \([0, t] \subseteq [0, t_{\text{max}}(k)]\). The main problem is to estimate the systematic error \( \sum_{n=k+1}^{\infty} P \{ t_n < t \} \). To do so, one needs some information about the DF \( F(t) \) of the r.v. \( \tau \), perhaps the existence of the moment generating function (Cramér’s condition); see Petrov (1975). Then one can use precise large-deviation results for \( P \{ t_n < t \} \). The r.v. \( \tau \) satisfies Cramér’s condition if there exists \( \theta > 0 \) such that \( E(e^{\theta \tau}) < \infty \). Cramér’s condition is equivalent to an exponential decay rate of \( 1 - F(t) \) and is satisfied for light-tailed distributions; it provides the existence of all moments of the r.v. \( \tau \).

---

4 With the exception of Theorem 26 and Corollary 3, this section is taken from *Stochastic Models*, 22(2), pp. 175–199, Nonparametric estimation of the renewal function by empirical data, Markovich NM and Krieger UR, Section 2.1, © 2006 Taylor and Francis Group, LLC. With permission from Taylor and Francis Group.
Theorem 23 Let \( \{\tau_1, \ldots, \tau_l\} \) be a sequence of i.i.d. nonnegative r.v.s and \( t \in [0, t_{\text{max}}(k)] \). We suppose that \( E|\tau_i|^m < \infty \) for some integer \( m \geq 1 \), \( E\tau_i = \mu \), \( \text{var}(\tau_i) = \sigma^2 \), and that the parameter \( k \) obeys
\[
k = k(l) \sim l^p \quad \text{as} \quad l \to \infty, \quad 0 < p < 1/3.
\] (8.7)
Then
\[
P \left\{ \omega : \lim_{l \to \infty} \sup_t |H(t) - \tilde{H}(t, k, l)| = 0 \right\} = 1.
\]
The rate of this uniform convergence may be proved for the class \( \tilde{S} \) of ITDs such that
\[
1 - F(t) \geq \exp(-\nu t)
\] for any \( t \in [0, T] \) and some \( \nu > 0 \). We assume, without loss of generality, that \( [0, T] = [0, 1] \). The class \( \tilde{S} \) includes, for example, the exponential distribution and the Weibull distribution with shape parameter greater than 1. Hence, it follows for the estimate of the right-hand side of (8.6) that
\[
\frac{(F(t))^{k+1}}{1 - F(t)} \leq \frac{(1 - \exp(-\nu t))^{k+1}}{\exp(-\nu t)}.
\]
Then, for \( F(t) \in \tilde{S} \) the error of an approximation by (8.4) in the metric of \( C \) is estimated by
\[
\sup_t |H(t) - \tilde{H}(t, k, l)| \leq \sup_t \left( \frac{(1 - \exp(-\nu t))^{k+1}}{\exp(-\nu t)} + \left| \sum_{n=1}^{k} (P \{ t_n < t \} - F_{t_n}(t)) \right| \right)
\] (8.8)

Theorem 24 If \( \{\tau_1, \ldots, \tau_l\} \) is an i.i.d. nonnegative sample with DF \( F \in \tilde{S} \), \( t \in [0, 1] \) and the parameter \( k = c \cdot l^p \) \( (c = c(\nu) > 0) \), \( 0 < \rho < 1/3 - (2/3)\alpha \), \( 0 < \alpha < 1/2 \), then the asymptotic rate of convergence of the estimate \( \tilde{H}(t, k, l) \) to \( H(t) \) is given by the expression
\[
P \left\{ \omega : \lim_{l \to \infty} \sup_t l^\alpha |H(t) - \tilde{H}(t, k, l)| \leq c_1 \right\} = 1,
\]
where \( c_1 \) is a constant that is independent of \( l \).
Then the following confidence interval is derived for the RF.

Corollary 2 If the assumptions of Theorem 24 hold, then, with probability at least \( 1 - \chi \), \( 0 < \chi < 1 \),
\[
\tilde{H}(t, k, l) - D \leq H(t) \leq \tilde{H}(t, k, l) + D
\] (8.9)
where

\[ D = l^{-\alpha} + k \sqrt{-\frac{k \ln (\chi/2)}{2l}}. \]

In practice, inter-arrival times are often described by distributions with heavy tails (Chistyakov, 1964; Goldie and Klüppelberg, 1998). Two classes of heavy-tailed distributions are well known: the distributions with regularly varying tails where \( 1 - F(t) = t^{-\alpha} \ell(t), \ t > 0, \alpha > 0, \) and \( \ell(x) \) is a slowly varying function; and the subexponential distributions, with the property that for any \( \epsilon > 0 \) there exists \( T = T(\epsilon, F) \) such that for any \( t > T, \ 1 - F(t) > \exp(-\epsilon t) \). It is the specific feature of heavy-tailed distributions that they do not satisfy Cramér’s condition.

If \( t \) is not too large, an approximation of \( (8.12) \)

\[ \bar{S}_C \left\{ \tau_k \mid \tau_k \right\} \left( 1 - F(t) \right) \sim -\epsilon n \]

may be used for heavy-tailed distributions, namely,

\[ P\{t_n > t\} \sim \Phi \left( \frac{t}{\sqrt{n}} \right) \]

(8.10)

(this means that \( \lim_{n \to \infty} \sup_{0 < \epsilon < c_n/h_n} |P\{t_n > t\}/\Phi(\epsilon) - 1| = 0 \) if \( t \in (0, c_n/h_n) \) for any choice of the sequence \( h_n \to \infty \) as \( n \to \infty \) (Mikosch and Nagaev, 1998).

Several threshold sequences \( c_n \) are proposed by different authors. For example, for a Weibull distribution with shape parameter \( 0 < \alpha \leq 0.5, \ c_n/h_n \sim n^{1/(2-\alpha)}, \) and for \( 0.5 < \alpha < 1, \ c_n/h_n \sim n^{2/3}, \) for distributions with regularly varying tails and \( \alpha > 2, \ c_n/h_n \) may be \( \sim n^{0.5} \ln^{0.5} n, \) (Mikosch, 1999).

**Theorem 25** If \( \{\tau_1, \ldots, \tau_l\} \) is a sequence of i.i.d. nonnegative r.v.s with heavy-tailed DF \( F(t), \ t \in (0, \min(t_{\max}(k), c_k/h_k)), \) and the parameter \( k \) obeys (8.7), then

\[ P \left\{ \omega \colon \lim_{l \to \infty} \sup_{t} |H(t) - \tilde{H}(t, k, l)| = 0 \right\} = 1. \]

The next theorem, proved in Markovich (2004), gives the rate of uniform convergence for distributions with regularly varying tails. By the Karamata representation theorem (Embrechts et al., 1997; Mikosch, 1999; Resnick 2006) the slowly varying function \( \ell(x) \) can be rewritten in the form

\[ \ell(x) = c(x) \exp \left\{ \int_{x_0}^{x} \frac{\varepsilon(y)}{y} dy \right\}, \quad x \geq x_0, \]

(8.11)

for some \( x_0 > 0, \) where \( c(\cdot) \) is a measurable nonnegative function such that \( \lim_{x \to \infty} c(x) = c_0 \in (0, \infty) \) and \( \varepsilon(x) \) is continuous function, \( \lim_{x \to \infty} \varepsilon(x) = 0. \) It is assumed that \( c(x) \) is a monotone decreasing or increasing function and \( \varepsilon(x) \) is a nonpositive function.

**Theorem 26** Let \( \{\tau_1, \ldots, \tau_l\} \) be i.i.d. nonnegative regularly varying r.v.s with tail \( F(x) = \ell(x)x^{-\alpha}, \ x > 0, \alpha > 0, \) parameter \( k = d \cdot \rho^p, \ d \geq -A, \)

\[ 0 \leq \rho < (1 - 2\beta)/3, \]

(8.12)
\[ A = A(\beta) = \frac{\beta + (\alpha - \varepsilon(\theta)) (1 + \eta) \ln l - \ln e^* + \varepsilon(\theta) \ln x_0}{\ln \left( 1 - c^* l^{(1 + \eta)(-\alpha + \varepsilon(\theta))} x_0^{-\varepsilon(\theta)} \right)} + 1, \]

where \( c^* = \min(c_0, c(a)), \) \( \eta > 1/(\alpha - \varepsilon^*), 0 < \varepsilon^* < \alpha, \) \( 0 < \beta < 1/2, \) \( \theta \in [x_0, t_{\max}(k)], \) \( x_0 > 0, \) \( t \in [a, t_{\max}(k)], \) \( a > 0. \) Then

\[
P \left\{ \omega : \lim_{l \to \infty} \sup_t l^\beta |H(t) - \tilde{H}(t, k, l)| \leq c_1 \right\} = 1,
\]

where \( c_1 \) is a constant that is independent of \( l. \)

**Corollary 3** If assumptions of Theorem 26 hold, \( \eta > (1 - \ln v/\ln l)/(\alpha - \varepsilon^*), \) \( 0 < \varepsilon^* < \alpha, \) then with probability at least \( 1 - v, \) \( 0 < v < 1, \) inequality (8.9) holds, where

\[
D = l^{-\beta} + k \sqrt{-\frac{k}{2l} \ln \left( \frac{\nu - l^{1-\eta(\alpha-\varepsilon^*)}}{2} \right)}.
\]

The theorems determine the values of \( k \) as functions of the sample size \( l. \) These values \( k \) are given only up to a rough asymptotic equivalence. For instance, \( k \) can be multiplied by any positive constant and the theorems remain valid. In practice, one needs exact optimal values of \( k \) which are adapted to the empirical data. Therefore, we subsequently consider a data-dependent selection of \( k. \)

**8.5 Selection of \( k \) by a bootstrap method**

Using empirical data, \( k \) can be chosen automatically by minimizing the bootstrap estimate of the mean squared error of \( H(t) \) for fixed \( t \) (Markovich and Krieger, 2006a), i.e.

\[
\text{MSE}(t, k, l) = E(\tilde{H}(t, k, l) - H(t))^2 \to \min_k.
\]

The bootstrap estimate is obtained by drawing resamples with replacement from the original data set \( T^l. \) Some observations from \( T^l \) may appear more than once, while others do not appear at all.

The bias of the estimate (8.4) is given by (8.6) and the variance by

\[
\text{var} \tilde{H}(t, k, l) = E(\tilde{H}(t, k, l))^2 - \left( E \tilde{H}(t, k, l) \right)^2 = \sum_{n=1}^{k} \sum_{m=1}^{k} P\{\max(t_n, t_m) < t\} - \left( \sum_{n=1}^{k} P\{t_n < t\} \right)^2.
\]

---

5 This section is based on Markovich and Krieger (2006a, Section 2.2).
For the solution of several problems of statistics such as the choice of smoothing parameters of the kernel estimator for a PDF, a nonparametric regression or Hill’s estimate of a tail index, it is recommended to use smaller resamples of size \( l_1 < l \). The goal is to avoid the situation where the bootstrap estimate of the bias is equal to zero regardless of the nonzero true bias of the estimator (Hall, 1990).

The bootstrap estimate of the RF that is constructed from \( T_l^i \) by some of the resamples \( T_{l_1}^i = \{ \tau_1^*, \ldots, \tau_l^* \} \) of size \( l_1 \) is given in a way similar to (8.4) by

\[
\tilde{H}^*(t, k_1, l_1) = \sum_{n=1}^{k_1} \frac{1}{l_n^1} \sum_{i=1}^{l_n^1} \theta(t - t_n^{i}), \quad l_n^1 = \lfloor l_1/n \rfloor, \quad t_n^{i} = \sum_{q=1+n(i-1)}^{n} \tau_q^*.
\]

The values \( l_1 \) and \( l \) may be related by

\[
l_1 = l^\beta, \quad 0 < \beta < 1. \tag{8.13}\]

The values \( k_1 \) and \( k \) are related by

\[
k = k_1(l/l_1)^\alpha, \quad 0 < \alpha < 1. \tag{8.14}\]

What values of \( \alpha \) and \( \beta \) should be taken? Considering the related problem of choosing the smoothing parameter in the case of a Kernel PDF estimator or linear regression, Hall (1990) has shown by means of asymptotic theory that \( \beta = 1/2 \) leads to the most accurate results. For the bootstrap estimation of the parameter \( k \) of Hill’s estimate, \( \alpha = 2/3 \) has been recommended.

The bias and variance of \( \tilde{H}^* (t, k_1, l_1) \) are given by

\[
b^*(t, k_1, l_1) = E\{\tilde{H}^* (t, k_1, l_1) | T_l^i \} - \tilde{H} (t, k, l) \tag{8.15}\]

and

\[
var^*(t, k_1, l_1) = E\{\tilde{H}^* (t, k_1, l_1)^2 | T_l^i \} - \left( E\{\tilde{H}^* (t, k_1, l_1) | T_l^i \} \right)^2, \tag{8.16}\]

respectively. Here, \( T_l^i \) is fixed and the expectation is calculated over all theoretically possible resamples \( T_{l_1}^i \) of \( T_l^i \) with size \( l_1 \). The bootstrap estimate of MSE\((t, k, l)\) is determined by

\[
\text{MSE}^*(t, k_1, l_1) = E\{(\tilde{H}^* (t, k_1, l_1) - \tilde{H} (t, k, l))^2 | T_l^i \}
\]

\[
= E\left\{ \left( \tilde{H}^* (t, k_1, l_1) \right)^2 | T_l^i \right\} - 2\tilde{H} (t, k, l) E \left\{ \tilde{H}^* (t, k_1, l_1) | T_l^i \right\}
\]

\[
+ \tilde{H} (t, k, l)^2.
\]

Since \( \tilde{H} (t, k, l)^2 \) does not depend on \( k_1 \), the problem reduces to the minimization of

\[
E\left\{ \left( \tilde{H}^* (t, k_1, l_1) \right)^2 | T_l^i \right\} - 2\tilde{H} (t, k, l) E \left\{ \tilde{H}^* (t, k_1, l_1) | T_l^i \right\}
\]

with respect to \( k_1 \).
In the following, we will show that minimizing $\text{MSE}^\star(t, k_1, l_1)$ with respect to $k_1$ is as awkward as the calculation of the estimator (8.2). The problem arises from the calculation of the subsequently considered statistic $F_{\tilde{p}}(t)$. It coincides with the statistic $\hat{F}_{\tilde{p}}(t)$ – see (8.3) – and it is close to the $U$-statistic $F_{\tilde{p}}^{(n)}(t)$. We note the difference that $F_{\tilde{p}}(t)$ is calculated over all combinations of $n$ observations with possible repetitions. Since $l^\star$ is the number of possible $t_n^{(i)}$, we get

$$E\{\tilde{H}^\star(t, k_1, l_1) \mid T^l\} = E \left\{ \sum_{i=1}^{k_1} \frac{1}{l_1} \sum_{i=1}^{l_1} \theta(t - t_n^{(i)}) \mid T^l \right\} = \sum_{i=1}^{k_1} \frac{1}{l_1} \sum_{i=1}^{l_1} E\{\theta(t - t_n^{(i)}) \mid T^l\} = \sum_{i=1}^{k_1} \frac{1}{l_1} \sum_{i=1}^{l_1} \theta(t - t_n^{(i)}) = \sum_{n=1}^{k_1} F_{\tilde{p}}(t). \quad (8.17)$$

Hence, by (8.15) and (8.17) one can see that the bias of the bootstrap approach does not depend on $l_1$, that is,

$$b^\star(t, k_1, l) = \sum_{n=1}^{k_1} F_{\tilde{p}}(t) - \tilde{H}(t, k, l). \quad (8.18)$$

When resamples of size $l$ are used, then the bias of the bootstrap

$$b^\star(t, k, l) = \sum_{n=1}^{k} F_{\tilde{p}}(t) - \tilde{H}(t, k, l)$$

may be close to zero for sufficiently large $l$ (since $F_{\tilde{p}}(t)$ and $F_{\tilde{p}}(t)$ may not differ so much), but for $k = 1$ it is equal to zero since $l^\star = l_n$ regardless of the true bias of $\tilde{H}(t, 1, l)$ (see (8.6)).

Independent of the values $k_1$ and $l_1$, the bootstrap variance is equal to zero. This property follows from (8.16), (8.17) and the expression

$$E\{\tilde{H}^\star(t, k_1, l_1) \mid T^l\}^2 = \sum_{n=1}^{k_1} \sum_{m=1}^{k_1} \frac{1}{l_1} \sum_{i=1}^{l_1} \sum_{j=1}^{l_1} E\{\theta(t - \max(t_n^{(i)}, t_m^{(j)}) \mid T^l\} = \sum_{n=1}^{k_1} \sum_{m=1}^{k_1} F_{\tilde{p}}(t) F_{\tilde{p}}(t).$$

One should not mix the variance and the bootstrap variance. The latter is a r.v.

The bootstrap estimate of $\text{MSE}(t, k, l)$ is given by

$$\text{MSE}^\star(t, k_1, l) = \sum_{n=1}^{k_1} \sum_{m=1}^{k_1} F_{\tilde{p}}(t) F_{\tilde{p}}(t) - 2\tilde{H}(t, k, l) \sum_{n=1}^{k_1} F_{\tilde{p}}(t).$$

---

6 See also Example 2 in Section 1.2.2.
However, the fact that the statistic $F_p(t)$ cannot be computed easily is a problem. Using the empirical estimate of $b^*(t, k_1, l)$,

$$
\hat{b}^*(t, k_1, l_1) = \frac{1}{B} \sum_{b=1}^{B} H^b(t, k_1, l_1) - \tilde{H}(t, k, l)
$$

(where $B$ denotes the number of $l_1$-sized resamples), instead of the actual bootstrap bias may give rough results. Here, we denote by $H^b(t, k_1, l_1)$ the estimate (8.4) constructed from some resample. In practice one can minimize the empirical estimate of $\text{MSE}^*(t, k_1, l)$,

$$
\text{MSE}^*(t, k_1, l_1) = \hat{b}^*(t, k_1, l_1)^2 + \hat{\text{var}}^*(t, k_1, l_1),
$$

(8.19)

where

$$
\hat{\text{var}}^*(t, l_1, k_1) = \frac{1}{B-1} \sum_{b=1}^{B} \left( H^b(t, k_1, l_1) - \frac{1}{B} \sum_{b=1}^{B} H^b(t, k_1, l_1) \right)^2
$$

is an empirical estimate of the bootstrap variance. All possible values of $k_1$ should be examined, where $k_1$ is an integer in the interval $[1, [l_1/2]]$.

The estimate $H_l(t, k)$ requires $A_1 = \sum_{n=1}^{k} \binom{l}{n}(n+1) = 2^l(1 + l/2) - \sum_{n=k+1}^{l} \binom{l}{n} = \sum_{n=k+1}^{l} n \binom{l}{n} - 1$ operations, whereas $\tilde{H}(t, k, l)$ requires $A_2 = \sum_{n=1}^{k} \left[ \frac{l}{n} \right] (n+1)$ operations. The selection of $k$ in $\tilde{H}(t, k, l)$ by means of an empirical bootstrap method (i.e., the minimization of (8.19)) requires $A_2 + A_3$ operations, where $A_3 = \sum_{k_1=1}^{[l_1/2]} \left( B \sum_{n=1}^{k_1} \left[ \frac{l_1}{n_1} \right] (n_1 + 1) + 6 \right)$. Note that

$$
S_k = \sum_{n=1}^{k} \frac{1}{n} = c + \psi(k) + \frac{1}{k},
$$

where $c \approx 0.5772$ is Euler’s constant, $\psi(z) = \Gamma'(z)/\Gamma(z)$, and $\Gamma(\cdot)$ is the gamma function (Prudnikov et al., 1981). We suppose for simplicity, that $[l/n] = l/n$, $[l_1/n_1] = l_1/n_1$, $[l/2] = l_1/2$ and $k = [l/2] = l/2$. Since

$$
\sum_{n=0}^{l/2} \binom{l}{n} = 2^{l-1} + \frac{1 + (-1)^{l/2}}{4} \binom{l}{l/2},
$$

(Prudnikov et al., 1981), we have

$$
A_1 = 2^{l-1}(1 + l) + \frac{1 + (-1)^{l/2}}{4} \binom{l}{l/2} - 1 - \sum_{n=l/2+1}^{l} n \binom{l}{n}, \quad A_2 = l(l/2 + S_{l/2}),
$$

$$
A_3 = Bl_1 \left( \frac{l_1(2+l_1)}{8} + \sum_{k_1=1}^{l_1/2} S_{k_1} \right) + 3l_1.
$$
Hence, for example when \( k = 10, \ l = 20 \) we have \( A_1 = 5.86 \cdot 10^6, A_2 = 258.579, \ A_1/A_2 = 2.266 \cdot 10^4 \). Let \( \ B = 50, \ l_1 = 6 \approx l^{0.6} \) then \( A_3 = 3.118 \cdot 10^3 \) and \( A_1/(A_2 + A_3) = 1.735 \cdot 10^3 \) (see Table 8.1).

## 8.6 Selection of \( k \) by a plot

As a practical tool for the visual selection of \( k \), the plot of the histogram-type estimate for a fixed \( t \) against \( k \) may be used. An example of a similar approach is given by the Hill’s plot for selecting the tail index. The idea of the plot is based on the uniform convergence of estimate (8.4) to the true RF as \( k \) increases and \( l \to \infty \). Then one can select for a fixed \( t \) the smallest \( k \) corresponding to the interval of stability of the plot, namely,

\[
   k^* = \arg \min \{ k : \tilde{H}(t, k, l) = \tilde{H}(t, k + 1, l), k = 1, \ldots, l - 1 \} \tag{8.20}
\]

(Markovich, 2004).

Figure 8.1 shows plots of the histogram-type estimate (8.4) against \( k \) for different fixed time intervals \([0, t], \ t \in \{1, 3, 5, 10\} \). The Weibull distribution with shape parameter \( s = 3 \) and the sample size \( l = 50 \) is considered. The table to the right of the figure shows the values \( k \) selected by the bootstrap with parameters \( \alpha = 0.5 \) and \( \beta = 0.7 \) and by the plot. In this example, the bootstrap recommends larger \( k \) than the plot method. The choice of \( k \) is determined by the trade-off between two terms in the sum (8.5). The smaller \( k \) corresponds to a larger bias and a better estimate of the DF by the empirical DF. A larger \( k \) leads to a reduction in the bias.

In Figure 8.2 the histogram-type estimates for a Weibull \((s = 3)\) and a gamma \((s = 0.55, \lambda = 1)\) distribution are shown. The parameter \( k \) is selected by the

![Figure 8.1](image-url)  
**Figure 8.1** Histogram-type estimate of the RF against \( k \) for a Weibull distribution: \( t = 1 \) (solid horizontal line), \( t = 3 \) (dotted line), \( t = 5 \) (dot-dashed line), \( t = 10 \) (solid line).
bootstrap and by the plot method. The corresponding curves coincide for Weibull distribution.

Using larger sample sizes $l$ and an adaptive selection of $k$ from the data $T_l$, the histogram-type estimate may provide a smaller mean squared error compared to Frees’ estimate (Markovich, 2004). In contrast to (8.4), Frees’ estimate requires a lot of calculations and hence cannot be evaluated for large sample sizes.

The numbers of operations required to calculate $H_l(t, k)$ and $\tilde{H}(t, k, l)$ with fixed, bootstrap-selected and plot-selected $k$ are shown in Table 8.1. The plot method can also be applied to Frees’ estimate.

### Table 8.1 Number of operations for Frees’ and the histogram-type estimator with fixed, bootstrap-estimated, and plot-estimated $k$.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>No. of operations</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frees’ estimator with fixed $k$</td>
<td>$A_1 = \sum_{n=1}^{k} \binom{l}{n} (n+1)$</td>
<td>$k = 10$, $l = 20$, $B = 50$, $l_1 = 6$</td>
</tr>
<tr>
<td></td>
<td>$2^l (1 + l/2) - \sum_{n=k+1}^{l} \binom{l}{n} - \sum_{n=k+1}^{l} n \binom{l}{n} - 1$</td>
<td>$5.86 \cdot 10^6$</td>
</tr>
<tr>
<td>Histogram-type estimator with fixed $k$</td>
<td>$A_2 = \sum_{n=1}^{k} \left[ \frac{l}{n} \right] (n+1)$</td>
<td>258</td>
</tr>
<tr>
<td>Histogram-type estimator with bootstrap-selected $k$</td>
<td>$A_2 + A_3$, $A_3 = \sum_{k=1}^{[l/2]} \left[ \frac{l_1}{n_1} \right] (n_1 + 1) + 6$</td>
<td>3376</td>
</tr>
<tr>
<td>Histogram-type estimator with plot-selected $k$</td>
<td>$\sum_{k=1}^{[l/2]} A_2$, $\sum_{k=1}^{[l/2]} \sum_{n=1}^{k} \left[ \frac{l}{n} \right] (n+1)$</td>
<td>1544</td>
</tr>
</tbody>
</table>
8.7 Simulation study

To evaluate the performance of the bootstrap approach, we have to investigate the influence of the values $\alpha$ and $\beta$ in (8.13) and (8.14) by a Monte Carlo simulation. We consider the values $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$ and $\beta \in \{0.3, 0.5, 0.7\}$. Let $[0, T]$ be the interval of the estimation, where $T \in \{0.5, 2.5, 4.5, 6.5, 8.5, 10\}$.

We generate samples with known PDFs

\[
    f_1(t) = \begin{cases} 
    \lambda \exp(-\lambda t), & t \geq 0, \\
    0, & t < 0,
    \end{cases}
\]

of an exponential distribution with parameter $\lambda = 1$,

\[
    f_2(t) = \begin{cases} 
    t^{s-1} \lambda^{-s} \exp(-t/\lambda) / \Gamma(s), & t > 0, \\
    0, & t \leq 0,
    \end{cases}
\]

of a gamma distribution with parameters $s = 2$ and $\lambda = 1$, and

\[
    f_3(t) = \begin{cases} 
    st^{s-1} \exp(-t^s), & t > 0 \\
    0, & t \leq 0
    \end{cases}
\]

of a Weibull distribution with $s = 0.5$. The latter distribution is heavy-tailed and subexponential. It is one of the most interesting distributions in reliability engineering where the PDF is singular.

For $f_1(t)$ and $f_2(t)$ the RFs are determined by

\[
    H_1(t) = \lambda t, \quad H_2(t) = 0.5 \left( t - 0.5 + 0.5 \exp(-2t) \right),
\]

respectively. For $f_3(t)$ the explicit form of the RF is unknown. Therefore, we use the results of a numerical approximation by Xie’s RS method for $H_3(t)$ since this method provides rather accurate results for a known PDF and a correctly selected step size $h = t/N$. Here $N$ is the number of points inside the interval $[0, t]$ (Xie, 1989). Strictly speaking, $H(i)$ is recursively calculated by

\[
    H(i) \approx \frac{F_1(i) + \sum_{j=1}^{i-1} F(i-j) (H(j) - H(j-1)) - F_0 H(i-1)}{1 - F_0},
\]

where $0 = z_0 < z_1 < \ldots < z_N = t$ and $H(i) = H(i/N), F(i) = F \left( \frac{(i+0.5)t}{N} \right), F_0 = F \left( \frac{0.5t}{N} \right), F_1(i) = F \left( \frac{i}{N} \right)$ are used.

Tables 8.2 and 8.3 show the bias and mean squared error of the estimate (8.4) calculated by 200 repeated samples for the given $f_1(t)$ and $f_2(t)$. The parameter $k$ in

---

7 This section is taken from Stochastic Models, 22(2), pp. 175–199, Nonparametric estimation of the renewal function by empirical data, Markovich NM and Krieger UR, Section 3. © 2006 Taylor and Francis Group, LLC. With permission from Taylor and Francis Group.
<table>
<thead>
<tr>
<th>$T$</th>
<th>$\alpha$</th>
<th>$\beta = 0.3$</th>
<th>$\beta = 0.5$</th>
<th>$\beta = 0.7$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BIAS $\cdot 10^3$</td>
<td>MSE $\cdot 10^4$</td>
<td>BIAS $\cdot 10^3$</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
<td>-26.99</td>
<td>17.5</td>
<td>82.3</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>26.3</td>
<td>20.06</td>
<td>-16.99</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>23.3</td>
<td>15.67</td>
<td>18.3</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>59.3</td>
<td>23.45</td>
<td>-26.7</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>-31.57</td>
<td>19.62</td>
<td>39.3</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>14.3</td>
<td>18.14</td>
<td>50.43</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>-68.7</td>
<td>19.59</td>
<td>32.3</td>
</tr>
<tr>
<td>2.5</td>
<td>0.1</td>
<td>-280</td>
<td>210</td>
<td>-270</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>-530</td>
<td>170</td>
<td>-270</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>-110</td>
<td>290</td>
<td>-390</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>-140</td>
<td>250</td>
<td>-120</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>76.86</td>
<td>250</td>
<td>36.35</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>64.45</td>
<td>230</td>
<td>83.03</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>-68.84</td>
<td>260</td>
<td>-76.26</td>
</tr>
<tr>
<td>4.5</td>
<td>0.1</td>
<td>-1840</td>
<td>910</td>
<td>-1660</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>-1400</td>
<td>1130</td>
<td>-1310</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>-240</td>
<td>800</td>
<td>-1010</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>260</td>
<td>670</td>
<td>-200</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>120</td>
<td>690</td>
<td>99.1</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>-26.1</td>
<td>600</td>
<td>230</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>110</td>
<td>630</td>
<td>-1040</td>
</tr>
<tr>
<td>6.5</td>
<td>0.1</td>
<td>-4700</td>
<td>4360</td>
<td>-3090</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>-3970</td>
<td>4770</td>
<td>-2680</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>-330</td>
<td>1540</td>
<td>2330</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>-280</td>
<td>1600</td>
<td>-320</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>-870</td>
<td>1260</td>
<td>-390</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>410</td>
<td>1340</td>
<td>-540</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>-7.029</td>
<td>1270</td>
<td>260</td>
</tr>
<tr>
<td>8.5</td>
<td>0.1</td>
<td>-7070</td>
<td>11100</td>
<td>-6160</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>-8920</td>
<td>18220</td>
<td>-6060</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>-1720</td>
<td>3530</td>
<td>-5280</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>-1140</td>
<td>3570</td>
<td>-710</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>290</td>
<td>2220</td>
<td>-170</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>-130</td>
<td>2150</td>
<td>-170</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>150</td>
<td>2430</td>
<td>320</td>
</tr>
<tr>
<td>10</td>
<td>0.1</td>
<td>-11340</td>
<td>18530</td>
<td>-8180</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>-17940</td>
<td>46720</td>
<td>-7290</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>-4560</td>
<td>8850</td>
<td>-10660</td>
</tr>
</tbody>
</table>
Table 8.2  (Continued)

<table>
<thead>
<tr>
<th>T</th>
<th>α</th>
<th>β = 0.3</th>
<th></th>
<th>β = 0.5</th>
<th></th>
<th>β = 0.7</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BIAS</td>
<td>MSE</td>
<td></td>
<td>BIAS</td>
<td>MSE</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>.10⁴</td>
<td>.10⁴</td>
<td></td>
<td>.10⁴</td>
<td>.10⁴</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>-4370</td>
<td>8770</td>
<td>-1120</td>
<td>3290</td>
<td>-4750</td>
<td>13830</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>-340</td>
<td>3360</td>
<td>-1620</td>
<td>4510</td>
<td>-6370</td>
<td>18730</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>87.66</td>
<td>3240</td>
<td>300</td>
<td>2750</td>
<td>-3920</td>
<td>11810</td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>-240</td>
<td>2680</td>
<td>-570</td>
<td>3200</td>
<td>450</td>
<td>3180</td>
<td></td>
</tr>
</tbody>
</table>

Reprinted from *Stochastic Models*, 22(2), pp. 175–199, Nonparametric estimation of the renewal function by empirical data, Markovich NM and Krieger UR, Table 1. © 2006 Taylor and Francis Group, LLC. With permission from Taylor and Francis Group.

Table 8.3  Quality of estimate (8.4): Exp (λ = 1, Eτ = 1), sample size l = 50.

<table>
<thead>
<tr>
<th>T</th>
<th>α</th>
<th>β = 0.3</th>
<th></th>
<th>β = 0.5</th>
<th></th>
<th>β = 0.7</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BIAS</td>
<td>MSE</td>
<td></td>
<td>BIAS</td>
<td>MSE</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>.10⁴</td>
<td>.10⁴</td>
<td></td>
<td>.10⁴</td>
<td>.10⁴</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
<td>-170</td>
<td>120</td>
<td>-280</td>
<td>130</td>
<td>-220</td>
<td>130</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>-250</td>
<td>140</td>
<td>-120</td>
<td>130</td>
<td>-100</td>
<td>140</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>63.17</td>
<td>180</td>
<td>-110</td>
<td>160</td>
<td>-220</td>
<td>140</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>-170</td>
<td>130</td>
<td>-180</td>
<td>180</td>
<td>-87</td>
<td>170</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>38.92</td>
<td>140</td>
<td>160</td>
<td>160</td>
<td>41.12</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>-94.96</td>
<td>200</td>
<td>96.71</td>
<td>160</td>
<td>-21.37</td>
<td>130</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>-14.29</td>
<td>160</td>
<td>39.46</td>
<td>160</td>
<td>83.21</td>
<td>170</td>
</tr>
<tr>
<td>2.5</td>
<td>0.1</td>
<td>-5580</td>
<td>4620</td>
<td>-4160</td>
<td>2920</td>
<td>-2760</td>
<td>2190</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>-4290</td>
<td>4430</td>
<td>-4750</td>
<td>3780</td>
<td>-3100</td>
<td>2250</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>-810</td>
<td>2240</td>
<td>-4600</td>
<td>4100</td>
<td>-3300</td>
<td>2809</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>-1760</td>
<td>1990</td>
<td>-520</td>
<td>1705</td>
<td>-2060</td>
<td>2540</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>-300</td>
<td>1500</td>
<td>-360</td>
<td>1800</td>
<td>-1620</td>
<td>2520</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>-240</td>
<td>1780</td>
<td>-91.67</td>
<td>1780</td>
<td>-1480</td>
<td>2340</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>200</td>
<td>1900</td>
<td>-310</td>
<td>1790</td>
<td>10.83</td>
<td>1980</td>
</tr>
<tr>
<td>4.5</td>
<td>0.1</td>
<td>-13190</td>
<td>24050</td>
<td>-11850</td>
<td>21340</td>
<td>-8880</td>
<td>12660</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>-14970</td>
<td>35240</td>
<td>-12300</td>
<td>27260</td>
<td>-8190</td>
<td>14160</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>-5920</td>
<td>11690</td>
<td>-11990</td>
<td>34250</td>
<td>-11140</td>
<td>23220</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>-6370</td>
<td>12122</td>
<td>8570</td>
<td>7344</td>
<td>15500</td>
<td>24025</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>-1530</td>
<td>4914</td>
<td>8980</td>
<td>8064</td>
<td>-8940</td>
<td>23720</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>-340</td>
<td>5970</td>
<td>-690</td>
<td>5530</td>
<td>-10160</td>
<td>28880</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>-340</td>
<td>6050</td>
<td>-580</td>
<td>4760</td>
<td>-840</td>
<td>4750</td>
</tr>
<tr>
<td>6.5</td>
<td>0.1</td>
<td>-29770</td>
<td>93881</td>
<td>-17910</td>
<td>53084</td>
<td>-16540</td>
<td>45496</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>-43990</td>
<td>196426</td>
<td>-25250</td>
<td>96100</td>
<td>-17600</td>
<td>56929</td>
</tr>
</tbody>
</table>
Reprinted from Stochastic Models, 22(2), pp. 175–199, Nonparametric estimation of the renewal function by empirical data, Markovich NM and Krieger UR, Table 2. © 2006 Taylor and Francis Group, LLC. With permission from Taylor and Francis Group.

(8.4) is determined by the bootstrap method. The sample size is \( l = 50 \), and \( B = 50 \) bootstrap resamples were taken. To understand better the results of Tables 8.2 and 8.3, it may be helpful to examine Figures 8.3 and 8.4 and Table 8.4. The values \( \overline{\text{MSE}} \) and \( \overline{\text{BIAS}} \) are the averages of MSE and BIAS values over all different \( T \) for each fixed couple \((\alpha, \beta)\). In Figures 8.3 and 8.4 the left-hand figures correspond to Table 8.2 and the right-hand figures to Table 8.3. Table 8.4 shows the corresponding smallest values of \( \overline{\text{MSE}} \) and \( \overline{\text{BIAS}} \) for a gamma and an exponential distribution. From Tables 8.2–8.4 and Figures 8.3, 8.4 it is evident that

- \( \alpha = 0.7, \beta = 0.3 \) give the smallest \( \overline{\text{MSE}} \);
- the best trade-off between the averages \( \overline{\text{MSE}} \) and \( \overline{\text{BIAS}} \) is provided by \( \alpha \in \{0.6, 0.7\}, \beta \in \{0.3, 0.5\} \);
- the mean squared error increases if the time interval \([0, T]\) of the estimation is extended.

Figures 8.5–8.7 show Xie’s estimate and the histogram-type estimate for the PDF \( f_3(t) \) and the PDF

\[
 f_4(t) = \begin{cases} 
 \frac{c}{:\rho} \left( \frac{\rho}{\rho+\tau} \right)^{c+1}, & t > 0, \\
 0, & t \leq 0,
\end{cases}
\]
Figure 8.3 Averages of the MSEs from Tables 8.2 and 8.3 over different \( T \) for fixed \( \alpha = 0.1(0.1)0.7, \beta \in \{0.3, 0.5, 0.7\} \) and for a gamma distribution (left) and an exponential distribution (right): \( \beta = 0.3 \) (solid line), \( \beta = 0.5 \) (dotted line), \( \beta = 0.7 \) (dot-dashed line). Reprinted from *Stochastic Models*, 22(2), pp. 175–199, Nonparametric estimation of the renewal function by empirical data, Markovich NM and Krieger UR, Figure 1. © 2006 Taylor and Francis Group, LLC. With permission from Taylor and Francis Group.

Figure 8.4 Averages of the BIAses in Tables 8.2 and 8.3 over different \( T \) for fixed \( \alpha = 0.1(0.1)0.7, \beta \in \{0.3, 0.5, 0.7\} \) and for a gamma distribution (left) and an exponential distribution (right): \( \beta = 0.3 \) (solid line), \( \beta = 0.5 \) (dotted line), \( \beta = 0.7 \) (dot-dashed line). Reprinted from *Stochastic Models*, 22(2), pp. 175–199, Nonparametric estimation of the renewal function by empirical data, Markovich NM and Krieger UR, Figure 2. © 2006 Taylor and Francis Group, LLC. With permission from Taylor and Francis Group.
Table 8.4  Minimal values minMSE and minBIAS of averages MSE and BIAS calculated by Tables 8.2 and 8.3 and corresponding $\alpha$ and $\beta$.

<table>
<thead>
<tr>
<th></th>
<th>Gamma distribution</th>
<th>Exponential distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>minMSE</td>
<td>0.121</td>
<td></td>
</tr>
<tr>
<td>minBIAS</td>
<td>$7.757 \cdot 10^{-4}$</td>
<td></td>
</tr>
<tr>
<td>($\alpha, \beta$)</td>
<td>$(0.7, 0.3)$</td>
<td>$(\alpha, \beta) = (0.7, 0.5)$</td>
</tr>
</tbody>
</table>


![H(t) vs. t](image)

**Figure 8.5** Estimation of the RF of a Weibull distribution by $\tilde{H}(t, k, l)$ against $t$, with $\alpha = 0.7$, $\beta = 0.3$ (step line); $\alpha = 0.7$, $\beta = 0.5$ (dot-dashed line); $\alpha = 0.7$, $\beta = 0.7$ (solid line with circles); $\alpha = 0.1$, $\beta = 0.5$ (solid line with crosses); $\alpha = 0.4$, $\beta = 0.5$ (dotted line). Xie’s estimate is shown by the solid line. Reprinted from *Stochastic Models*, 22(2), pp. 175–199, Nonparametric estimation of the renewal function by empirical data, Markovich NM and Krieger UR, Figure 3. © 2006 Taylor and Francis Group, LLC. With permission from Taylor and Francis Group.

of a Pareto distribution with parameters $c = 0.5$, $\rho = 0.5$ on the time interval $[0, 5]$, as well as for the exponential PDF with $\lambda = 1$. The sample size is $l = 100$. The parameter $k$ was selected by the bootstrap method with parameters $(\alpha, \beta) \in \{(0.7, 0.3); (0.7, 0.5); (0.7, 0.7); (0.1, 0.5); (0.4, 0.5)\}$. For the bootstrap $B = 50$ resamples was taken. In Figures 8.6 and 8.7 the lines corresponding to $(\alpha, \beta) \in \{(0.7, 0.3); (0.7, 0.5); (0.7, 0.7)\}$ coincide with each other. In Figure 8.5
Figure 8.6  Estimation of the RF of a Pareto distribution by $\tilde{H}(t, k, l)$ against $t$, with $\alpha = 0.7$, $\beta = 0.3$ (step line); $\alpha = 0.7$, $\beta = 0.5$ (dot-dashed line); $\alpha = 0.7$, $\beta = 0.7$ (solid line with circles); $\alpha = 0.1$, $\beta = 0.5$ (solid line with crosses); $\alpha = 0.4$, $\beta = 0.5$ (dotted line). Xie’s estimate is shown by the solid line. Reprinted from *Stochastic Models*, 22(2), pp. 175–199, Nonparametric estimation of the renewal function by empirical data, Markovich NM and Krieger UR, Figure 4. © 2006 Taylor and Francis Group, LLC. With permission from Taylor and Francis Group.

Figure 8.7  Estimation of the RF of an exponential distribution by $\tilde{H}(t, k, l)$ against $t$, with $\alpha = 0.7$, $\beta = 0.3$ (step line); $\alpha = 0.7$, $\beta = 0.5$ (dot-dashed line); $\alpha = 0.7$, $\beta = 0.7$ (solid line with circles); $\alpha = 0.1$, $\beta = 0.5$ (solid line with crosses); $\alpha = 0.4$, $\beta = 0.5$ (dotted line). Xie’s estimate is shown by the solid line. Reprinted from *Stochastic Models*, 22(2), pp. 175–199, Nonparametric estimation of the renewal function by empirical data, Markovich NM and Krieger UR, Figure 5. © 2006 Taylor and Francis Group, LLC. With permission from Taylor and Francis Group.
the line corresponding to \((\alpha, \beta) = (0.7, 0.7)\) differs from those lines with \((\alpha, \beta) \in \{(0.7, 0.3); (0.7, 0.5)\}\) (the latter two lines coincide) approximately on the interval [2,3,3]. One can see that the curves corresponding to \(k\) selected by bootstrap with \((\alpha, \beta) \in \{(0.7, 0.3); (0.7, 0.5)\}\) for \(f_\beta(t)\) and with \((\alpha, \beta) \in \{(0.7, 0.3); (0.7, 0.5); (0.7, 0.7)\}\) for the Pareto and exponential PDFs are closer to the true RF than all other curves. The line with \((\alpha, \beta) = (0.4, 0.5)\) is better than that with \((\alpha, \beta) = (0.1, 0.5)\), especially for the Pareto PDF. The figures support our previous conclusion regarding the prevalence of \(\alpha = 0.7\) and \(\beta \in \{0.3, 0.5\}\).

The figures also illustrate the following phenomenon. Referring to formula (8.4), one can see that for some fixed \(t\) the value of \(\tilde{H}(t, k, l)\) may not change anymore as \(k\) increases. For example, we have for \(f_\alpha(t)\) (Figure 8.5) and \(t = 3\) that \(k \in \{3; 3; 26; 29; 36\}\) for \((\alpha, \beta) \in \{(0.1, 0.3); (0.4, 0.5); (0.7, 0.7); (0.7, 0.3); (0.7, 0.5)\}\), respectively. This reflects the situation where the corresponding \(\{t'_k\}\) are larger than \(t\) and the corresponding terms in the sum (8.4) are equal to 0.

The second part of the simulation study relates to the comparison with the tables presented in Frees (1986a). For this purpose samples of the lognormal distribution with PDF \(f(x) = \left(x \sigma \sqrt{2\pi}\right)^{-1} \exp\left(-\frac{(\log x - \mu)^2}{2\sigma^2}\right)\), where \(\mu = 0\) and \(\sigma^2 = 1\), and of the Weibull distribution with \(s = 3\) (not heavy-tailed Weibull) were generated. The gamma distribution \((s = 0.55, \lambda = 1)\) presented in Frees (1986a) was not considered. Since the generator of gamma r.v.s (Ahrens and Dieter, 1974) used in Frees (1986a) is not reliable for small samples, it has an adverse influence on the accuracy of the results of a simulation study.

\(T \in \{0.25, 0.5, 0.75, 1.0, 1.25\}\) are the times provided and \(l \in \{10, 15, 20, 25, 30, 100\}\) are the sample sizes. As in Frees (1986a), two characteristics, the bias and the mean squared error of the estimates, were calculated over 500 Monte Carlo repetitions and \(H(t)\) is the true RF. For the fixed number of points \(T\) and for the distributions mentioned, \(H(t)\) was taken from tables in Baxter et al. (1982). The results of the calculation are presented in the Tables 8.5–8.8. Frees’ results are included here; \(H_{3n}(t)\) denotes the estimate (8.2). Since (8.2) requires much computational effort, only \(k \in \{5, 10\}\) and \(l \leq 30\) were considered. Considering (8.4), the parameter \(k\) is calculated by the bootstrap method, that is, by minimizing (8.19) with respect to \(k_1\), where \(l_1\) and \(k_1\) are related to \(l\) and \(k\) by the formulas (8.13) and (8.14) with parameters \(\alpha = 0.7\) and \(\beta = 0.3\). \(B = 50\) bootstrap resamples was taken. The parameter \(k\) is also calculated by the plot method. Tables 8.5–8.8 show that, for all estimates:

- the mean squared error increases as \(T\) increases;
- for any fixed \(T\) the mean squared error decreases as \(l\) increases;
- the bias does not exhibit stable behavior.
Table 8.5 Monte Carlo study of the bias and the mean squared error of Frees’ estimate $H_{3n}(t)$ and histogram-type estimate $\tilde{H}(t, k, l)$ of the renewal function for the lognormal distribution with parameters $(\mu = 0, \sigma = 1)$ and $E \tau = \exp(1/2)$ for sample sizes $l \in \{10, 15\}$, different time intervals $[0, T]$ and values of the parameter $k$.

<table>
<thead>
<tr>
<th>size</th>
<th>$T$</th>
<th>$H_{3n}(t)$</th>
<th>$\tilde{H}(t, k, l)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$k = 5$</td>
<td>$k = 10$</td>
</tr>
<tr>
<td></td>
<td>BIAS $\cdot 10^4$</td>
<td>MSE $\cdot 10^4$</td>
<td>BIAS $\cdot 10^4$</td>
</tr>
<tr>
<td>10</td>
<td>0.25</td>
<td>-1 77</td>
<td>-1 77</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>149 255</td>
<td>149 255</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>128 422</td>
<td>128 422</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>163 659</td>
<td>-163 659</td>
</tr>
<tr>
<td></td>
<td>1.25</td>
<td>117 918</td>
<td>117 918</td>
</tr>
<tr>
<td>15</td>
<td>0.25</td>
<td>1 54</td>
<td>1 54</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>123 179</td>
<td>123 179</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>99 296</td>
<td>99 296</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>133 443</td>
<td>123 443</td>
</tr>
<tr>
<td></td>
<td>1.25</td>
<td>93 619</td>
<td>93 619</td>
</tr>
</tbody>
</table>


Table 8.6 Monte Carlo study of the bias and the mean squared error of Frees’ estimate $H_{3n}(t)$ and histogram-type estimate $\tilde{H}(t, k, l)$ of the renewal function for the lognormal distribution with parameters $(\mu = 0, \sigma = 1)$ and $E \tau = \exp(1/2)$ for sample sizes $l \in \{20, 25, 30, 100\}$, different time intervals $[0, T]$ and values of the parameter $k$.

<table>
<thead>
<tr>
<th>size</th>
<th>$T$</th>
<th>$H_{3n}(t)$</th>
<th>$\tilde{H}(t, k, l)$</th>
<th>$k_{\text{boot}}$</th>
<th>$k_{\text{plot}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$k = 5$</td>
<td></td>
<td>BIAS $\cdot 10^4$</td>
<td>MSE $\cdot 10^4$</td>
</tr>
<tr>
<td>20</td>
<td>0.25</td>
<td>-10 38</td>
<td>-52.57 $\cdot 10^4$</td>
<td>55.214 $\cdot 10^4$</td>
<td>44.624 $\cdot 10^4$</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>94 124</td>
<td>-22.15 $\cdot 10^4$</td>
<td>-115.966 $\cdot 10^4$</td>
<td>127.713 $\cdot 10^4$</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>95 218</td>
<td>59.99 $\cdot 10^4$</td>
<td>-65.762 $\cdot 10^4$</td>
<td>236.264 $\cdot 10^4$</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>141 329</td>
<td>-61.07 $\cdot 10^4$</td>
<td>-73.716 $\cdot 10^4$</td>
<td>387.564 $\cdot 10^4$</td>
</tr>
<tr>
<td></td>
<td>1.25</td>
<td>126 452</td>
<td>55.8 $\cdot 10^4$</td>
<td>-72.275 $\cdot 10^4$</td>
<td>590.623 $\cdot 10^4$</td>
</tr>
<tr>
<td>25</td>
<td>0.25</td>
<td>16 29</td>
<td>16.62 $\cdot 10^4$</td>
<td>-40.129 $\cdot 10^4$</td>
<td>31.635 $\cdot 10^4$</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>111 101</td>
<td>-37.62 $\cdot 10^4$</td>
<td>4.626 $\cdot 10^4$</td>
<td>99.883 $\cdot 10^4$</td>
</tr>
</tbody>
</table>
Table 8.7 Monte Carlo study of the bias and the mean squared error of Frees’ estimate \( H_{3n}(t) \) and histogram-type estimate \( \tilde{H}(t, k, l) \) of the renewal function for the Weibull distribution with parameter \( s = 3 \) and \( E\tau = 0.89 \) for sample sizes \( l \in \{10, 15\} \), different time intervals \([0, T]\) and values of the parameter \( k \).

<table>
<thead>
<tr>
<th>size</th>
<th>( T )</th>
<th>( H_{3n}(t) )</th>
<th>( \tilde{H}(t, k, l) )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>( k = 5 )</td>
<td>( k = 10 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BIAS ( \cdot 10^4 )</td>
<td>MSE ( \cdot 10^4 )</td>
</tr>
<tr>
<td>10</td>
<td>0.25</td>
<td>0.075</td>
<td>123</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>10</td>
<td>179</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>1.25</td>
<td>185</td>
</tr>
<tr>
<td>30</td>
<td>0.25</td>
<td>0.25</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.75</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1.25</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>1.25</td>
<td>n.a.</td>
<td>96</td>
</tr>
<tr>
<td>100</td>
<td>0.25</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td></td>
<td>1.25</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
</tbody>
</table>

Table 8.8 Monte Carlo study of the bias and the mean squared error of Frees’ estimate $H_{3n}(t)$ and histogram-type estimate $\tilde{H}(t, k, l)$ of the renewal function for the Weibull distribution with parameter $s = 3$ and $E\tau = 0.89$ for sample sizes $l \in \{20, 25, 30, 100\}$, different time intervals $[0, T]$ and values of the parameter $k$.

<table>
<thead>
<tr>
<th>size</th>
<th>$T$</th>
<th>$H_{3n}(t)$</th>
<th>$\tilde{H}(t, k, l)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$k = 5$</td>
<td>$k_{\text{boot}}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BIAS $\cdot 10^4$</td>
<td>MSE $\cdot 10^4$</td>
</tr>
<tr>
<td>20</td>
<td>0.25</td>
<td>$-12$</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>1</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>43</td>
<td>130</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>34</td>
<td>174</td>
</tr>
<tr>
<td></td>
<td>1.25</td>
<td>35</td>
<td>171</td>
</tr>
<tr>
<td>25</td>
<td>0.25</td>
<td>$-13$</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>$-12$</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>1</td>
<td>104</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>$-12$</td>
<td>130</td>
</tr>
<tr>
<td></td>
<td>1.25</td>
<td>13</td>
<td>125</td>
</tr>
<tr>
<td>30</td>
<td>0.25</td>
<td>$-14$</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>$-39$</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>$-35$</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>$-46$</td>
<td>114</td>
</tr>
<tr>
<td></td>
<td>1.25</td>
<td>$-47$</td>
<td>111</td>
</tr>
<tr>
<td>100</td>
<td>0.25</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td></td>
<td>1.25</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
</tbody>
</table>

Comparing $\tilde{H}(t, k, l)$ with $H_{3n}(t)$, one may conclude that:

- the biases and mean squared errors of $\tilde{H}(t, k, l)$ and $H_{3n}(t)$ are comparable for the same sample sizes;
- the increasing the sample size provides better accuracy with regard to $\tilde{H}(t, k, l)$ as shown in Table 8.9, where $\overline{\text{MSE}}_{H_{3n}}, \overline{\text{MSE}}_{\tilde{H}}$ and $|\text{BIAS}|_{H_{3n}}, |\text{BIAS}|_{\tilde{H}}$ are averages of the MSE and $|\text{BIAS}|$ over different $T$. The averaging was provided using the results of Table 8.6 and 8.8 when the sample size is equal to 30 in the case of $H_{3n}$ and to 100 for $\tilde{H}$. 

Table 8.9  Comparison of $H_{3n}(t)$ for the sample size 30 and $\tilde{H}(t)$ for the sample size 100 regarding averages of the MSE and the $|BIAS|$ for different distributions and two selection methods of $k$ (the bootstrap and the plot method).

| Distribution | $\frac{MSE_{H_{3n}}}{MSE_{\tilde{H}}}$ | $\frac{|BIAS|_{H_{3n}}}{|BIAS|_{\tilde{H}}}$ |
|--------------|---------------------------------|---------------------------------|
|              | bootstrap | plot | bootstrap | plot |
| Lognormal    | 2.62      | 2.62 | 2.70      | 1.99 |
| Weibull      | 2.26      | 2.32 | 1.70      | 2.04 |

Comparing the bootstrap and the plot methods from Table 8.9, one may conclude that both methods demonstrate similar MSE and $|BIAS|$.

8.8 Application to the inter-arrival times of TCP connections

The estimator (8.4) was applied to 1000 TCP flow inter-arrival times measured in a mobile network (Markovich, 2008). Table 8.10 gives the descriptive statistics of this sample.

To apply (8.4) we need to check the independence of the inter-arrivals. Tests (see Section 1.3) show that the inter-arrivals of TCP connections are heavy-tailed distributed (Figure 8.8). The distribution of inter-arrivals is close to an exponential one. They may be independent (Figure 8.9).

The parameter $k$ in (8.4) was calculated by formula (8.20). The estimate looks close to a straight line on time interval $[0.5, 5]$ but on smaller time intervals the curve is not quite linear (Figure 8.10). This implies that the inter-arrivals of TCP connections cannot be considered as a pure Poisson process. Formula (8.4) provides a more exact estimate of the mean number of TCP connections.

Table 8.10  Description of the TCP flow inter-arrival times data.

<table>
<thead>
<tr>
<th>Unit</th>
<th>Min.</th>
<th>Max.</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>sec</td>
<td>10^{-5}</td>
<td>2.237</td>
<td>0.235</td>
<td>0.085</td>
</tr>
</tbody>
</table>

Figure 8.8 Left: Sample mean excess function $e_n(u)$ (1.41) against the threshold $u$ for the inter-arrivals of TCP connections. The plot is close to a constant, which indicates closeness to the exponential distribution. Right: EVI $\gamma$ estimation by Hill’s (dotted line) and group estimator (solid line) for the inter-arrivals of TCP connections. Horizontal lines show the plot-selected values: $\hat{\gamma}^H(n, k) = 0.388$ and $\hat{\gamma}^G(n, k) = 0.356$. The latter values imply infinite first two moments of the inter-arrival distribution and, hence, heaviness of the tail.

Figure 8.9 The sample ACF (1.43) of the inter-arrivals of TCP connections. The horizontal dotted lines indicate 95% asymptotic confidence bounds $\pm 1.96/\sqrt{n}$ corresponding to the ACF of i.i.d. Gaussian r.v.s. The inter-arrivals are located inside the confidence interval and, hence, may be independent.
8.9 Conclusions and discussion

In this chapter, we have discussed nonparametric estimation of the RF which does not require any knowledge of the form of the ITD.

Due to the limited number of empirical data the histogram-type estimate (as well as Frees’ estimate (8.2)) can be applied for closed time intervals $[0, t]$ with relatively small $t$. Compared to Frees’ estimate $F_l^{(n)}(t)$ of the arrival-time distribution $F^*(t)$ in (8.2), the estimate (8.4) uses a simpler and rougher estimate of $F^*(t)$. The RF $H(t)$ is approximated by a finite sum of estimates of the arrival-time distributions with $k$ terms. The parameter $k$ is selected to compensate the error of the risk function. The estimate (8.4) may be computed for sufficiently large $l$ and $k$, which is not realistic for (8.2).

Theorems 23 and 25 state, both for heavy- and light-tailed ITDs, those values of the parameter $k$ as functions of the sample size which provide almost surely the uniform convergence of the histogram-type estimate (8.4) to the true RF for sufficiently small $t$. It is proved that a smaller value of $k$ ($k < l$) than in Frees (1986b) is sufficient to get a reliable estimate of the RF.

In Theorem 24 the rate of uniform convergence and a confidence interval of the RF for the specific class of ITDs with an exponential decay rate of the tails are presented. But these theorems determine $k$ only up to a rough asymptotic equivalence. Such a value $k$ does not depend on the empirical data. This feature may influence the accuracy of the estimation.

To estimate $k$ from samples of moderate size, the bootstrap method is used. Following Hall (1990), a smaller resample size $l_1$ (and $k_1$) is used to avoid the situation where the bootstrap estimate of the bias is equal or close to zero regardless of the true bias of the estimate. Then the bootstrap estimate $E\left\{\left(\widetilde{H}^*(t, l_1, k_1) - \widetilde{H}(t, l, k)\right)^2 \mid T'\right\}$ of the mean squared error $\text{MSE}(t, k, l) = E\left\{\left(\widetilde{H}(t, l, k) - H(t)\right)^2 \right\}$ is minimized with respect to $k_1$, where $\widetilde{H}^*(t, l_1, k_1)$ is the
RF estimate derived from one of the resamples. The relevant relationships between $l$ and $l_1$ as well as between $k$ and $k_1$ are found by a Monte Carlo study.

As an alternative data-dependent method to estimate $k$ from samples of a moderate size, the plot method is proposed.

The histogram-type estimate (8.4) tends to decrease the MSE in comparison with Frees’ estimate (8.2) by using larger samples.

The number of operations required for (8.4) with the bootstrap selection of $k$ is much less than that for Frees’ estimate, and with the plot selection is less than that for a bootstrap selection.

As usual, the main disadvantage of the bootstrap method is that it requires the choice of additional parameters $\alpha$ and $\beta$ that determine the resample size to estimate $k$. In contrast, the plot method does not require any additional parameters.

### 8.10 Exercises

1. **RF estimation at infinity.**

   Generate 100 inter-arrivals which are regularly varying with DF
   \[
   F(x) = 1 - x^{-\alpha}, \quad x \geq 1, \quad \text{for } \alpha = 3 \text{ and } 1 < \alpha < 2.
   \]

   Approximate the behavior of the RF $H(t)$ for large $t$ (e.g., $100 < t < 10000$) applying the formulas
   \[
   H(t) \approx t/\mu + \sigma^2/(2\mu^2) - 1/2, \quad \text{for } \alpha = 3, \quad (8.21)
   \]
   \[
   H(t) \approx t/\mu + t^2/\left(\mu^2(\alpha - 1)(2 - \alpha)\right) \left(1 - F(t)\right), \quad \text{for } 1 < \alpha < 2.
   \]

   Determine the true values of the mean and variance of $F(x)$. Calculate the empirical mean and variance as estimates of $\mu$ and $\sigma^2$. Estimate $F(x)$ in (8.22) by the empirical DF and by $\hat{F}(x) = 1 - x^{-\hat{\alpha}}$. Estimate the tail index $\alpha = 1/\gamma$ by Hill’s estimator, $\hat{\alpha} = 1/\hat{\gamma}^H(n,k)$. Compare the estimates and the approximations (the latter are calculated for the known $\mu, \sigma^2$ and $F(x)$).

2. **RF estimation at infinity.**

   Generate $l = 1000$ Weibull distributed random numbers with $s = 0.3$ (p. 234) as $T_l$. Estimate the tail index by some method and determine the number of finite moments. Approximate the behavior of the RF $H(t)$ at infinity by formulas (8.21) and (8.22) if possible.

3. **RF estimation on a finite interval $[0, t]$.**

   Generate $l = 100$ exponentially distributed random numbers with $\lambda = 0.5$ and $l = 100$ gamma distributed random numbers with $s = 2$ and $\lambda = 1$.

   Calculate a histogram-type estimate of the RF on $[0, t], t \in \{0.5, 0.75, 1, 1.25\}$ by formula (8.4). Take different $k \leq l/2$. Compare the estimates with the true RFs, $H(t) = \lambda t$ and $H(t) = 0.5(t - 0.5 + 0.5 \exp(-2t))$, of the exponential and gamma distributions.
4. RF estimation on a finite interval \([0, t]\).

Generate random numbers as indicated in Exercise 3. Construct the plot of the histogram-type estimate \(\tilde{H}(t, k, l)\) against \(k\) for different \([0, t]\), \(t \in \{1, 2, 3, 5, 10\}\). Determine the value of \(k\) by the plot method for each fixed \(t\), that is, by formula (8.20). This implies that for each curve related to a fixed \(t\) the value \(k = k(t)\) is selected that corresponds to the beginning of a stable interval of the curve.

Calculate \(\tilde{H}(t, k, l)\) with time-dependent \(k\). Compare the estimates with the true RFs of the exponential and gamma distributions.
Appendix A

Proofs of Chapter 2

Proof of Lemma 2. By (2.48), the left-hand side of (2.49) clearly does not exceed

\[
\sum_{i=1}^{n} \int \int |K_h(x-u)K_h(x-v)||f_i(u,v) - f(u)f(v)|\,du\,dv \\
\leq \beta_n \int \int |K_h(x-u)K_h(x-v)|\,du\,dv \leq (h+\eta)^2 \beta_n.
\]
Appendix B

Proofs of Chapter 4

Proof of Theorem 3. Suppose that \( h_* \not\to 0 \) as \( n \to \infty \). This implies that for any integer \( N > 0, \exists n > N \) such that \( h_*(n) > H_*, \) where \( H_* \) is some positive constant. We shall prove that, for such \( h_* \), \( \sup_{x \in \Omega} |F_n(x) - F_{h_*}(x)| \not\to 0 \) as \( n \to \infty \).

For any solution \( h_* \) and \( x \in \Omega_* \) one may represent the divergence in (4.28) using the substitution \( u = (t - X_i)/h_* \):

\[
F_n(x) - F_{h_*}(x) = \frac{1}{n} \sum_{i=1}^{n} \left( \theta(x - X_i) - \frac{1}{h_*} \int_{0}^{x} K \left( \frac{t - X_i}{h_*} \right) 1\{|t - X_i| \leq h_*\} dt \right)
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \left( \theta(x - X_i) - \int_{t_i^1}^{t_i^2} K(u) 1\{|u| \leq 1\} du \right),
\]

where we denote \( t_i^1 = t_i^1(h_*) = -X_i/h_* \), \( t_i^2 = t_i^2(h_*) = (x - X_i)/h_* \), and

\[
\theta(x) = \begin{cases} 
1, & x \geq 0, \\
0, & x < 0.
\end{cases}
\]

Furthermore, we omit \( 1\{|u| \leq 1\} \), bearing in mind that \( K(u) \) is compactly supported on \([-1, 1]\). Without loss of generality, we can consider the sequence

\[
h_1 \leq h_2 \leq \ldots \leq h_j \leq \ldots ,
\]
where \( h_j = h_\ast (n_j) = H_\ast + j\Delta, \Delta \) is some positive constant, and \( N < n_1 \leq n_2 \leq \ldots \leq n_j \leq \ldots \). Since, for any fixed \( i \), the sequences \(-X_i/h_\ast (n_j) \to 0 \) and \((x - X_i)/h_\ast (n_j) \to 0 \) as \( n_j \to \infty \), we have
\[
\int_{t_i}^{\bar{t}} K(u)du \to 0, \quad \text{as } n_j \to \infty.
\]
Therefore,
\[
\theta (x - X_i) - \int_{t_i}^{\bar{t}} K(u)du \to \theta (x - X_i), \quad n_j \to \infty.
\]
Hence,
\[
\sup_{x \in \Omega^*} |F_n(x) - F_h(x)| \to 1, \quad \text{as } n_j \to \infty.
\]
Therefore, \( h_\ast \to 0 \) as \( n \to \infty \), since \( \sup_{x \in \Omega^*} |F_n(x) - F_h(x)| \to 0 \) as \( n \to \infty \) according to (4.28).

**Proof of Theorem 4.** Obviously,
\[
\sup_{x \in \Omega^*} |F_n(x) - F_h(x)| \leq \sup_{x \in \Omega^*} |F_n(x) - F(x)| + \sup_{x \in \Omega^*} |F(x) - F_h(x)|. \tag{B.1}
\]
We shall estimate \( \sup_{x \in \Omega^*} |F(x) - F_h(x)| \). Note that
\[
\hat{f}_h(x) = \frac{1}{h} \int_{\Omega^*} K \left( \frac{x - y}{h} \right) 1 \{|x - t| \leq h\} dF_n(t),
\]
\[
E\hat{f}_h(x) = \frac{1}{h} \int_{\Omega^*} K \left( \frac{x - y}{h} \right) 1 \{|x - t| \leq h\} dF(t).
\]
Furthermore,
\[
\sup_{x \in \Omega^*} |F(x) - F_h(x)| \leq \sup_{x \in \Omega^*} \left| \int_0^x \left( f(t) - E\hat{f}_h(t) \right) dt \right| + \sup_{x \in \Omega^*} \left| \int_0^x \left( E\hat{f}_h(t) - \hat{f}_h(t) \right) dt \right|. \tag{B.2}
\]
The first term on the right-hand side of the latter inequality can be estimated from (4.25). For the second term we get
\[
\sup_{x \in \Omega^*} \left| \int_0^x \left( E\hat{f}_h(t) - \hat{f}_h(t) \right) dt \right| = \frac{1}{h} \sup_{x \in \Omega^*} \left| \int_0^x \left( \int_{\Omega^*} K \left( \frac{t - y}{h} \right) 1 \{|t - y| \leq h\} d(F(y) - F_n(y)) dt \right) \right|
\]
\[
\leq \sup_{x \in \Omega^*} \int_0^x \left| \int_{|u| \leq 1} K(u) d(F(t - hu) - F_n(t - hu)) \right| dt
\]
\[
\leq C \sup_{x \in \Omega^*} \int_0^x \left| F(t - h) - F_n(t - h) - (F(t + h) - F_n(t + h)) \right| dt
\]
\[
\leq 2C \sup_{x \in \Omega^*} |F(x) - F_n(x)|.
\]
From (4.25), (4.29), (B.1) and (B.2) we get

\[
\sup_{x \in \Omega^*} |F_n(x) - F_h(x)| \leq (2C + 1) \sup_{x \in \Omega^*} |F(x) - F_n(x)| + \eta_2 h^2 K_1/2.
\] (B.3)

since \( |K_1^f| \leq \eta_2 K_1 \). Assume that

\[
\sup_{x \in \Omega^*} |F_n(x) - F(x)| \leq (2 (2C + 1))^{-1} n^{-\alpha}.
\] (B.4)

Hence, for any solution \( h_* \) of (4.28) we get, from (B.1) and (B.3),

\[
n^{-\alpha}/2 \leq \eta_2 h_*^2 K_1/2.
\]

Hence, it follows that

\[
h_* \geq \rho_1 n^{-\alpha/2},
\] (B.5)

where \( \rho_1 = (\eta_2 K_1)^{-1/2} \). We shall now prove that \( h_* \leq \rho_2 n^{-\alpha/2} \). For this purpose, we consider the auxiliary function

\[
I(x, h) = \int_{-1}^{1} [(F(x - hy) - F(x)) - (F_n(x - hy) - F_n(x))] K(y) dy.
\]

Applying Taylor’s expansion to \( F(x - hy) \) up to the term of order \( h^2 \), we get for any \( x \) that

\[
\int_{-1}^{1} |F(x - hy) - F(x)| K(y) dy = \frac{h^2}{2} \int_{-1}^{1} y^2 |f'(\theta hy)| K(y) dy \geq h^2 G,
\]

where \( G = (\eta_1/2) \int_{-1}^{1} y^2 K(y) dy = K_1 \eta_1/2 \) is a positive constant.

Assume (B.4). It follows that

\[
|I(x, h)| \leq \int_{-1}^{1} |F(x - hy) - F_n(x - hy)| K(y) dy + |F(x) - F_n(x)|
\]

\[
\leq (2C + 1)^{-1} n^{-\alpha}.
\] (B.6)

Since \( h \) is selected from (4.28), we have, from (B.4) and (B.6),

\[
h_*^2 G \leq \sup_x \int_{-1}^{1} |F(x - hy) - F(x)| K(y) dy
\]

\[
\leq \sup_x |I(x, h)| + \sup_x \int_{-1}^{1} |F_n(x - hy) - F_n(x)| K(y) dy
\]

\[
\leq (2/(2C + 1)) n^{-\alpha}
\]

as \( n \to \infty \). Hence, it follows that

\[
h_* \leq \rho_2 n^{-\alpha/2},
\] (B.7)
where $\rho_2 = 2 \left( (2C + 1)K_1 \eta_1 \right)^{-1/2}$. One can see that assumption (B.4) leads to (B.5) and (B.7). From (7.5) it follows that

$$1 - P\{\rho_1 n^{-\alpha/2} \leq h_s \leq \rho_2 n^{-\alpha/2} \} < P\{\sup_x |F(x) - F_n(x)| > n^{-\alpha}/(2(2C + 1))\}$$

$$\leq 2 \exp \left( -n^{1-2\alpha}/(2(2C + 1)^2) \right).$$

**Proof of Theorem 5.** Suppose that any solution $h_s$ of (4.28) obeys the conditions $h_s > \rho_2 n^{-\alpha/2}$ and $h_s < \rho_1 n^{-\alpha/2}$. Then from the assertion of Theorem 4 and (4.27) we get

$$\text{MSE}(\hat{f}_{h_s}) \leq \left( \rho_1^4 \left( K_1^\prime(x) \right)^2 / 4 \right) n^{-2\alpha} + n^{-1+\alpha/2} f(x) K^* / \rho_2 + O(n^{-1}).$$

This implies that

$$P\left\{ \text{MSE}(\hat{f}_{h_s}) > \left( \rho_1^4 \left( K_1^\prime(x) \right)^2 / 4 \right) n^{-2\alpha} + n^{-1+\alpha/2} f(x) K^* / \rho_2 + cn^{-1} \right\} < 1 - P\{\rho_1 n^{-\alpha/2} \leq h_s \leq \rho_2 n^{-\alpha/2} \}$$

for some positive constant $c$. From (4.30) it follows that

$$P\{\text{MSE}(\hat{f}_{h_s}) > c^* n^{-4/5} \} \leq 2 \exp \left( -n^{4/5}/(2(2C + 1)^2) \right) = \psi(n)$$

when $\alpha = 2/5$. The series $\sum_{n=1}^\infty \psi(n)$ converges and, by the Borel–Cantelli lemma, the assertion of the theorem holds.

**Proof of Theorem 6.** The proof is similar to the proof of Theorem 3. Suppose that $h_s \not\rightarrow 0$ as $n \rightarrow \infty$. This implies that, for any integer $N > 0, \exists n > N$ such that $h_s = h_s(n) > H_s$, where $H_s$ is some positive constant. We shall prove that, for such $h_s$, $\sup_{-\infty < x < \infty} |F_n(x) - F^A_{h_s,h_1}(x)| \not\rightarrow 0$ as $n \rightarrow \infty$.

For any solution $h_s$ one may represent the divergence in (4.31) using the substitution $u = (t - X_i \hat{f}_{h_1}(X_i))^{1/2}/h_s$:

$$F_{u}(x) - F^A_{h_s,h_1}(x) = \frac{1}{n} \sum_{i=1}^{n} \left( \theta (x - X_i) - \frac{\hat{f}_{h_1}(X_i)^{1/2}}{h_s} \int_{-\infty}^{x} K \left( \frac{t - X_i \hat{f}_{h_1}(X_i)}{h_s} \right) dt \right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left( \theta (x - X_i) - \int_{-\infty}^{t_i(h_s)} K(u) du \right), \quad (B.8)$$

where $t_i = t_i(h_s) = (x - X_i \hat{f}_{h_1}(X_i))^{1/2}/h_s$ and

$$\theta(x) = \begin{cases} 1, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Without loss of generality, one can consider the sequence

$$h_1 \leq h_2 \leq \ldots \leq h_j \leq \ldots,$$
where \( h_j = h_*(n_j) = H_* + j \Delta, \Delta \) is some positive constant, and \( N < n_1 \leq n_2 \leq \ldots \leq n_J \leq \ldots \).

Since for any fixed \( i, t_i(h_j) \to 0 \) as \( n_j \to \infty \), we have
\[
\int_{-\infty}^{t_i} K(u)du \to \int_{-\infty}^{0} K(u)du.
\]

Since \( \sup_x |K(x)| < \infty \), we get
\[
-\infty < -c \leq \int_{-\infty}^{0} K(u)du < 1
\]
(for a symmetric kernel \( K(x) = K(-x) \), we have \( \int_{-\infty}^{0} K(u)du = 1/2 \),
\[
-1 < \theta (x - X) - \int_{-\infty}^{0} K(u)du < 1 + c \quad \forall x.
\]

Hence
\[
\sup_{-\infty < x < \infty} |F_n(x) - F_{A,h_j,h_1}(x)| \to 1 + c, \quad \text{as } n_j \to \infty.
\]

This implies that the sequence \( \{F_n(x) - F_{A,h_j,h_1}(x), i = 1, 2, \ldots \} \), corresponding to \( h_1, h_2, \ldots, h_j, \ldots \), does not go to 0 as \( h_i \) increases for any \( x \). Hence,
\[
\sup_{-\infty < x < \infty} |F_n(x) - F_{h_*},h_1(x)| \not\to 0, \quad \text{as } n \to \infty.
\]

Therefore, \( h_* \to 0 \) as \( n \to \infty \).

**Proof of Theorem 7.** We denote
\[
I(x, h) = \int_{-\infty}^{\infty} [(F(x - hy) - F(x)) - (F_n(x - hy) - F_n(x))]K(y)dy.
\]

Using the fact that the kernel \( K(x) \) has order \( m + 1 \) and applying Taylor’s expansion to \( F(x - hy) \) up to the term of order \( h^{m+1} \), we get
\[
\sup_x \left| \int_{-\infty}^{\infty} (F(x - hy) - F(x))K(y)dy \right|
= h^{m+1} \sup_x \left| \int_{-\infty}^{\infty} \frac{y^{m+1}}{(m+1)!} F^{(m+1)}(x - \theta hy)K(y)dy \right|
\geq h^{m+1}G,
\]
where \( G = (1/(m + 1)! \sup_x | \int_{-\infty}^{\infty} f^{(m)}(x - \theta hy) y^{m+1}K(y)dy | \) is a positive constant, \( 0 < \theta < 1 \), since \( f^{(m)}(x) \) is bounded.

Suppose that, for \( \alpha > 2 \),
\[
\sup_x |F(x) - F_n(x)| \leq n^{-1/\alpha}. \quad (B.9)
\]
Then,
\[
|I(x, h)| \leq \int_{-\infty}^{\infty} |F(x - hy) - F_n(x - hy)|K(y)dy + |F(x) - F_n(x)| \leq 2n^{-1/\alpha}. \quad (B.10)
\]
Since $h$ is selected from (4.31), we have, from (B.9) and (B.10),

$$h^{m+1} G \leq \sup_x \left| \int_{-\infty}^{\infty} (F(x - hy) - F(x)) K(y) dy \right|$$

$$\leq \sup_x |I(x, h)| + \sup_x \int_{-\infty}^{\infty} |(F_n(x - hy) - F_n(x)) K(y)| dy$$

$$\leq 2n^{-1/\alpha} + 2A\delta n^{-1/2},$$

as $n \to \infty$. Hence, from (B.9) it follows that $h \leq \rho n^{-1/(\alpha(m+1))}$, where $ho = (2(1 + A\delta)/G)^{1/(m+1)}$, since $\alpha > 2$. We now use the well-known inequality (7.5), due to Prakasa Rao (1983),

$$P\{\sup_x |F_n(x) - F(x)| > \eta \} \leq 2 \exp \left(-2n\eta^2\right),$$

to conclude that

$$P\{h > \rho n^{-1/(\alpha(m+1))}\} < P\{\sup_x |F(x) - F_n(x)| > n^{-1/\alpha}\} \leq 2 \exp \left(-2n^{1-2/\alpha}\right).$$

**Proof of Theorem 8.** Denote $\varphi(x) = (d/dx)^4 1/f(x)$, $K_3 = \int_{-\infty}^{+\infty} x^4 K(x) dx$. According to Hall and Marron (1988), it follows for the solution $h_*$ of (4.31) and for the assumed $K(x)$ that

$$\tilde{f}^A(x|h_1, h_*) = \hat{f}^A(x|h_*) + cZ(nh_*)^{-1/2} + o((nh_*)^{-1/2});$$

see also (3.18). Then the bias of $\tilde{f}^A(x|h_1, h_*)$ is the same as for $\hat{f}^A(x|h_*)$, that is,

$$E\tilde{f}^A(x|h_1, h_*) - f(x) = \frac{K_3}{24} h_*^4 \varphi(x) + o(h_*^4)$$

(Hall and Marron, 1988). Suppose that $h_* \leq \rho n^{-1/(\alpha(m+1))}$, where $\rho$ is defined in Theorem 7. Then it follows that

$$E\tilde{f}^A(x|h_1, h_*) - f(x) \leq \frac{K_3}{24} \varphi(x) \rho^4 n^{-4/(\alpha(m+1))} + o(n^{-4/(\alpha(m+1))}).$$

For $\alpha = 9/(m+1)$ the bias of $\tilde{f}^A(x|h_1, h_*)$ has order $n^{-4/9}$ for any positive integer $m < 3.5$, since $\alpha > 2$. Then, we have

$$P \left\{ \frac{E\tilde{f}^A(x|h_1, h_*) - f(x)}{\frac{K_3}{24} \varphi(x) \rho^4 n^{-4/9}} > \frac{K_3}{24} \varphi(x) \rho^4 n^{-4/9} \right\} < P\{h_* > \rho n^{-1/(\alpha(m+1))}\}$$

$$\leq 2 \exp \left(-2n^{1-2(m+1)/9}\right) = \psi(n).$$

Since the series $\sum_{n=1}^{\infty} \psi(n)$ converges, the assertion of the theorem holds by the Borel–Cantelli lemma.

**Proof of Corollary 1.** Denote $K_2^* = \int K^2(t) dt$. From (B.11) and since $E(Z \cdot \hat{f}^A(x|h_*) = 0)$, the variance of $\tilde{f}^A(x|h_1, h_*)$ is

$$\text{var} (\tilde{f}^A(x|h_1, h_*)) = \text{var} (\hat{f}^A(x|h_*)) + c^2 (nh_*)^{-1} + o((nh_*)^{-1})$$

$$= (nh_*)^{-1} \left( c^2 + f(x)^{3/2}K_2^* \right) + o((nh_*)^{-1}).$$

(B.13)
From Theorem 7 it follows that \( h_\ast = O\left(n^{-1/9}\right) \) if \( \alpha = 9/(m+1) \) and \( m < 3.5 \). Hence, from (B.12) and (B.13) we have that

\[
\text{MSE}(\hat{f}^A(x|h_1, h_\ast)) = (K_3/24) h_\ast^8 (\varphi(x))^2 + (nh_\ast)^{-1} \left( c^2 + f(x)^{3/2} K_2^* \right) + o(h_\ast^8) \sim n^{-8/9},
\]

as \( n \to \infty \), if a maximal solution \( h_\ast \) of (4.31) has order \( n^{-1/9} \).

**Proof of Theorem 9.** From (2.12) and (2.14) we obtain

\[
\int_0^1 \left( \hat{f}^*_{\gamma_n}(x) - f(x) \right)^2 dx = \frac{1}{2} \sum_{j=1}^{\infty} (\lambda_j a_j - \theta_j)^2.
\]

Let \( N \) be an arbitrary integer. Then

\[
\frac{1}{2} \sum_{j=1}^{N} (\lambda_j a_j - \theta_j)^2 = \frac{1}{2} \sum_{j=1}^{N} (\lambda_j a_j - \theta_j)^2 + \frac{1}{2} \sum_{j=N+1}^{\infty} (\lambda_j a_j - \theta_j)^2
\]

\[
\leq \sum_{j=1}^{N} \left( \frac{a_j - \theta_j}{1 + (\pi j \gamma)^{2k+2}} \right)^2 + \sum_{j=1}^{N} \left( \frac{(\pi j \gamma)^{2k+2}}{1 + (\pi j \gamma)^{2k+2}} \right)^2 \theta_j^2
\]

\[
+ \sum_{j=N+1}^{\infty} \left( \frac{a_j}{1 + (\pi j \gamma)^{2k+2}} \right)^2 + \sum_{j=N+1}^{\infty} \theta_j^2.
\]

We estimate each term on the right-hand side of this inequality. For the first term, we have

\[
\sum_{j=1}^{N} \left( \frac{a_j - \theta_j}{1 + (\pi j \gamma)^{2k+2}} \right)^2 \leq \sup_{1 \leq j \leq N} (a_j - \theta_j)^2 \sum_{j=1}^{N} \left( \frac{1}{1 + (\pi j \gamma)^{2k+2}} \right)^2 \leq 2 \frac{\epsilon_n^2}{\pi \gamma},
\]

where

\[
\epsilon_n = \sup_{1 \leq j \leq N} (a_j - \theta_j).
\]

Since \( f(x) \in \varphi \) holds, according to Fikhtengol’ts (1965), for its Fourier coefficients we have the inequality

\[
|\theta_j| \leq 2 V_k / j^{k+1}, \quad j = 1, 2, \ldots,
\]

where \( V_k \) is the variation of the function \( f^{(k)}(x) \). Therefore, for the second term on the right-hand side of (B.14), we have

\[
\sum_{j=1}^{N} \left( \frac{(\pi j \gamma)^{2k+2}}{1 + (\pi j \gamma)^{2k+2}} \right)^2 \leq 4 V_k^2 \sum_{j=1}^{N} \frac{(\pi j \gamma)^{2k+2}}{1 + (\pi j \gamma)^{2k+2}} \frac{1}{j^{2k+2}}
\]

\[
= 4 V_k^2 (\pi \gamma)^{2k+2} \sum_{j=1}^{N} \left( \frac{(\pi j \gamma)^{k+1}}{1 + (\pi j \gamma)^{2k+2}} \right)^2
\]
$< 4V_k^2 (\pi \gamma)^{2k+1} (1 + \pi \gamma/2)$

$< 8V_k^2 (\pi \gamma)^{2k+1}$.

To estimate the third term, we take into account that $|a_j| \leq 2$:

$$
\sum_{j=N+1}^{\infty} \left( \frac{a_j}{1 + (\pi j \gamma)^{2k+2}} \right)^2 \leq \sum_{j=N+1}^{\infty} \left( \frac{a_j}{(\pi j \gamma)^{2k+2}} \right)^2 \leq \frac{4}{(\pi \gamma)^{4k+4} (4k+3) N^{4k+3}}.
$$

From (B.16), we have for the fourth term

$$
\sum_{j=N+1}^{\infty} \theta_j^2 \leq \frac{4V_k^2}{(2k+1) N^{2k+1}}.
$$

Drawing these results together, we obtain

$$
\| \hat{f}_\gamma^n(x) - f(x) \|^2 \leq 2 \frac{\epsilon_n^2}{\pi \gamma} + 8V_k^2 (\pi \gamma)^{2k+1} + \frac{4}{(\pi \gamma)^{4k+4} (4k+3) N^{4k+3}} + \frac{4V_k^2}{(2k+1) N^{2k+1}}.
$$

Since $N$ is an arbitrary number, we take $N = n^{1/(k+1)}$. By the assumption of the theorem, it follows $\gamma = n^{-1/(2k+2)}$. Therefore

$$
\| \hat{f}_\gamma^n(x) - f(x) \|^2 \leq 2 \frac{\epsilon_n^2}{\pi} n^{1/(2k+2)} + c_1 n^{-(2k+1)/(2k+2)} + c_2 n^{-(2k+1)/(k+1)},
$$

where $c_1 = 8V_k^2 \pi^{2k+1}$, $c_2 = 4 \left( 1/(\pi^{4k+4}(4k+3)) + V_k^2/(2k+1) \right)$. Thus

$$
\frac{n^{(2k+1)/(2k+2)}}{\ln n} \| \hat{f}_\gamma^n(x) - f(x) \|^2 \leq \frac{2n \epsilon_n^2}{\pi \ln n} + \frac{c_1}{\ln n} + \frac{c_2}{\ln n} n^{-(2k+1)/(2k+2)} = A_n + B_n,
$$

where

$$
A_n = \frac{2n \epsilon_n^2}{\pi \ln n}, \quad B_n = (\ln n)^{-1} \left( c_1 + c_2 n^{-(2k+1)/(2k+2)} \right).
$$

For sufficiently large $n$, $B_n \leq 1$ and $A_n$ is a random variable. If $A_n \leq 8$, then

$$
\frac{n^{(2k+1)/(2k+2)}}{\ln n} \| \hat{f}_\gamma^n(x) - f(x) \|^2 \leq 9.
$$

Consequently,

$$
P \left\{ \frac{n^{(2k+1)/(2k+2)}}{\ln n} \| \hat{f}_\gamma^n(x) - f(x) \|^2 > 9 \right\} < P \{A_n > 8\}.
$$

We estimate the right-hand side using Hoeffding’s inequality (Petrov, 1975). According to this inequality,

$$
P \{ \epsilon_n > k \} < 2N \exp \left( -nk^2/8 \right). \quad (B.17)
$$
Then
\[ P\{A_n > 8\} = P\left\{ \epsilon_n > \sqrt{\frac{8\pi \ln n}{2n}} \right\} < 2n^{1/(k+1)} \exp\left( -\frac{\pi \ln n}{2} \right) = 2n^{1/(k+1)-\pi/2}. \]

Since \( k \geq 1 \), we have
\[
\sum_{n=1}^\infty n^{1/(k+1)-\pi/2} < \infty,
\]
and, according to the Borel–Cantelli lemma the first assertion of Theorem 9 holds.

In order to prove the second assertion, instead of (B.15) one has to use the estimate
\[
\sum_{j=1}^N \frac{E(a_j - \theta_j)^2}{(1 + (\pi j \gamma)^{2k+2})} = \sum_{j=1}^N \frac{(4/n) \left( E\varphi_j^2(X_1) - \theta_j^2/4 \right)}{(1 + (\pi j \gamma)^{2k+2})^2} \leq \frac{c_3}{n \pi \gamma},
\]
where \( c_3 \) is some constant. Then Theorem 9 is proved.

For the proofs of Theorems 10 and 11 we need the following lemmas, in each of which we assume that \( X_1, \ldots, X_n \) is a sample of i.i.d. r.v.s with PDF \( f(x) \in \varphi \).

**Lemma 6** If in estimate (2.14) we have \( \gamma = (G/n)^{1/(2k+3)} \), where \( G \) is some constant, then we have the inequality
\[
P\{ \hat{f}_\gamma^{pr}(x, X^n) > B \} < 2n^{1/(2k+1)} \exp\left( -n^{(2k-1)/(2k+1)} \right),
\]
where \( B = 2 + 4V_k + 2\sqrt{2} \).

**Lemma 7** For any \( N > 0 \), we have the inequality
\[
\int_0^1 (F_n(x) - F^\gamma(x))^2 \, dx < 4V_k^2 \pi^{2k+1} \gamma^{2k+3} (1 + 2\pi \gamma) + \frac{\epsilon_n^2 \gamma (1 + 2\pi \gamma)}{\pi} + \frac{4}{\pi^2 N}.
\]

**Lemma 8** Assume that the regularization parameter \( \gamma > 0 \) in the estimate of the PDF \( \hat{f}_\gamma^{pr}(x, X^n) \) can be found from (4.34). Then we have the inequality
\[
\sum_{j=1}^N \left( \frac{a_j (\pi j \gamma)^{2k+2}}{j \left( 1 + (\pi j \gamma)^{2k+2} \right)} \right)^2 \leq \frac{2\pi^2 \beta}{\beta n},
\]
where \( \beta > 0 \) (see (4.35)).

**Proof of Lemma 6.** From expression (2.14) for the estimation of \( \hat{f}_\gamma^{pr}(x, X^n) \) we obtain
\[
\hat{f}_\gamma^{pr}(x, X^n) \leq 1 + \sum_{j=1}^\infty \lambda_j |a_j|.
\]
We divide the sum on the right-hand side of this inequality into a finite sum of \( N \) terms and the remainder starting from the \((N + 1)\)th index \((N \) is an arbitrary integer). From (B.16) and from the fact that \(|\lambda_j| \leq 1, |a_j| \leq 2, j = 1, 2, \ldots,\) we obtain

\[
\sum_{j=1}^{\infty} \lambda_j |a_j| < \sum_{j=1}^{N} |a_j| + 2 \sum_{j=N+1}^{\infty} \lambda_j < \sum_{j=1}^{N} |a_j - \theta_j| + \sum_{j=1}^{N} |\theta_j| + 2 \sum_{j=N+1}^{\infty} \frac{1}{1 + (\pi j \gamma)^{2k+2}} \leq \epsilon_n N + 2V_k \sum_{j=1}^{N} \frac{1}{j^{k+1}} + 2 \sum_{j=N+1}^{\infty} \frac{1}{(\pi j \gamma)^{2k+2}} \leq K_N + M_N,
\]

where

\[
K_N = \epsilon_n N, \quad M_N = 4V_k + \frac{2}{(2k+1)(\pi \gamma)^{2k+2}N^{2k+1}}.
\]

Since \( N \) is arbitrary here, we take \( N = n^{1/(2k+1)} \). Then for \( \gamma = (G/n)^{1/(2k+3)} \), where \( G \) is some constant, and for sufficiently large \( n \) we have \( M_N \leq 1 + 4V_k \). If \( K_n \leq 2\sqrt{2} \), then

\[
\hat{f}_{\gamma}^p (x, X^n) < 2 + 4V_k + 2\sqrt{2} = B.
\]

Consequently,

\[
P\{\hat{f}_{\gamma}^p (x, X^n) > B\} < P\{K_N > 2\sqrt{2}\}.
\]

We estimate the right-hand side by Hoeffding’s inequality. From (B.16) we obtain

\[
P\{K_N > 2\sqrt{2}\} = P\{\epsilon_n > 2\sqrt{2}n^{-1/(2k+1)}\} < 2n^{1/(2k+1)} \exp \left( -n^{(2k-1)/(2k+1)} \right),
\]

and the assertion of the lemma holds.

**Proof of Lemma 7.** \( F_{\gamma}(x) \) for estimate (2.14) is determined in the following manner:

\[
F_{\gamma}(x) = x + \frac{1}{\pi} \sum_{j=1}^{\infty} \frac{a_j}{j \left( 1 + (\pi j \gamma)^{2k+2} \right)} \sin(\pi j x).
\]

The expansion of the empirical DF

\[
F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \theta(x - X_i)
\]

into a sine Fourier series on \([0,1]\) has the form

\[
F_n(x) - x \sim \frac{1}{\pi} \sum_{i=1}^{\infty} \frac{a_j}{j} \sin(\pi j x),
\]

(B.19)
where $a_j$ are the same coefficients as in (2.14). The Fourier series of the function $F_n(x)$ does not converge to it on the entire segment $[0,1]$. As is known, at the points of discontinuity of the first kind $X_i$, $i = 1, \ldots, n$, the series converges to the value

$$\frac{F_n(X_i + 0) + F_n(X_i - 0)}{2}$$

(Fikhtengol’ts, 1965). Then

$$F_n(x) - F^n(x) \sim \frac{1}{\pi} \sum_{i=1}^{n} \frac{a_j (\pi j \gamma)^{2k+2}}{j \left(1 + (\pi j \gamma)^{2k+2}\right)} \sin(\pi j x),$$

$$\int_0^1 (F_n(x) - F^n(x))^2 \, dx = \frac{1}{2\pi^2} \sum_{i=1}^{\infty} \left(\frac{a_j (\pi j \gamma)^{2k+2}}{j \left(1 + (\pi j \gamma)^{2k+2}\right)}\right)^2. \quad \text{(B.20)}$$

Let $N$ be an arbitrary integer. We divide the sum on the right-hand side of (B.20) into a finite sum of $N$ terms and the remainder starting from the $(N+1)$th index. We estimate each of the obtained terms:

$$\frac{1}{2} \sum_{i=1}^{N} \left(\frac{a_j (\pi j \gamma)^{2k+2}}{j \left(1 + (\pi j \gamma)^{2k+2}\right)}\right)^2 \leq \sum_{i=1}^{N} \left(\frac{\theta_j (\pi j \gamma)^{2k+2}}{j \left(1 + (\pi j \gamma)^{2k+2}\right)}\right)^2 + \varepsilon_n^2 \sum_{i=1}^{\infty} \left(\frac{(\pi j \gamma)^{2k+2}}{j \left(1 + (\pi j \gamma)^{2k+2}\right)}\right)^2. \quad \text{(B.21)}$$

From (B.16) we obtain the inequality

$$\sum_{i=1}^{N} \left(\frac{\theta_j (\pi j \gamma)^{2k+2}}{j \left(1 + (\pi j \gamma)^{2k+2}\right)}\right)^2 \leq 4V_k^2 (\pi \gamma)^{2k+2} \sum_{i=1}^{N} \left(\frac{(\pi j \gamma)^{2k+2}}{j \left(1 + (\pi j \gamma)^{2k+2}\right)}\right)^2 \quad \text{(B.22)}$$

$$< 4V_k^2 (\pi \gamma)^{2k+3} (1 + 2\pi \gamma).$$

We estimate the second sum on the right-hand side of (B.21):

$$\varepsilon_n^2 \sum_{i=1}^{N} \left(\frac{(\pi j \gamma)^{2k+2}}{j \left(1 + (\pi j \gamma)^{2k+2}\right)}\right)^2 < \varepsilon_n^2 \pi \gamma (1 + 2\pi \gamma). \quad \text{(B.23)}$$

Since $|a_j| \leq 2$, $j = 1, 2, \ldots$ and $(\pi j \gamma)^{2k+2} / \left(1 + (\pi j \gamma)^{2k+2}\right) < 1$, we have

$$\sum_{j=N+1}^{\infty} \left(\frac{a_j (\pi j \gamma)^{2k+2}}{j \left(1 + (\pi j \gamma)^{2k+2}\right)}\right)^2 \leq 4 \sum_{j=N+1}^{\infty} \frac{1}{j^2} < \frac{4}{N}. \quad \text{(B.24)}$$
From (B.20)–(B.24) we have
\[
\int_0^1 \left(F_n(x) - F^\gamma(x)\right)^2 \, dx < 4V_k^2 \pi^{2k+1} \gamma^{2k+3} \left(1 + 2\pi\gamma\right) + \frac{\varepsilon_n^2 \gamma (1 + 2\pi\gamma)}{\pi} + \frac{4}{\pi^2 N}.
\]

Then the lemma is proved.

**Proof of Lemma 8.** From (4.33) it follows that
\[
\frac{\delta}{n} \geq \beta \int_0^1 \left(F_n(x) - F^\gamma(x)\right)^2 \, dx.
\]
From this and from (B.20) the assertion of the lemma follows.

**Proof of Theorem 10.** Let
\[
\hat{f}_\gamma^{pr}(x, X^n) \leq B. \tag{B.25}
\]
Then from (4.34) and Lemma 8 for \( \gamma = \left(G/n\right)^{1/(2k+3)} \), we obtain the inequality
\[
\hat{\omega}_n^2 < \max\{B, \beta\} n \int_0^1 \left(F_n(x) - F^\gamma(x)\right)^2 \, dx
\]
\[
< \max\{B, \beta\} n \left[ 4V_k^2 \pi^{2k+1} \frac{G}{n} \left(1 + 2\pi \left(\frac{G}{n}\right)^{1/(2k+3)}\right) \right.
\]
\[
+ \varepsilon_n^2 \left(\frac{G}{n}\right)^{1/(2k+3)} \pi \left(1 + 2\pi \left(\frac{G}{n}\right)^{1/(2k+3)}\right) + \frac{4}{\pi^2 N}\right] = \Omega(n).
\]

Since \( N \) is arbitrary, we select \( N = n^2 \). We assume that
\[
\varepsilon_n \leq 5\sqrt{\ln n/n}. \tag{B.26}
\]
Then for \( G = \delta / \left(8V_k^2 \pi^{2k+1} \max\{B, \beta\}\right) \) and sufficiently large \( n \), the quantity \( \Omega(n) < \delta \) and
\[
\hat{\omega}_n^2 < \delta. \tag{B.27}
\]
On the other hand, inequality (4.33) holds by assumption. This means that for \( \gamma \to \infty \), according to (4.33), (B.18) and (B.19) we have
\[
\hat{\omega}_n^2 \to n \int_0^1 (x - F_n(x))^2 \, dx = \frac{n}{2\pi^2} \sum_{j=1}^{\infty} \left(\frac{a_j}{j}\right)^2 \geq \delta. \tag{B.28}
\]
Thus, under the conditions (B.25) and (B.26), from (B.27), (B.28) and from the continuity of \( \hat{\omega}_n^2 \) with respect to \( \gamma \) it follows that there exists \( \gamma \geq \left(G/n\right)^{1/(2k+3)} \) such that \( \hat{\omega}_n^2 = \delta \). In other words, if \( \gamma \) is the largest value of the smoothing parameter such that \( \hat{\omega}_n^2 = \delta \), then
\[
P\left\{ \gamma < \left(\frac{G}{n}\right)^{1/(2k+3)} \right\} < P\left\{ \varepsilon_n > 5\sqrt{\ln n/n} \right\} + P\left\{ \hat{f}_\gamma^{pr}(x, X^n) > B \right\}.$$
From (B.16) we obtain
\[ P \left\{ \varepsilon_n > 5 \sqrt{\frac{\ln n}{n}} \right\} < 2n^{-9/8} + 2n^{1/(2k+1)} \exp \left( -n^{(2k-1)/(2k+1)} \right). \]

From this it is clear that, starting from some \( n \), the assertion of the theorem holds.

**Proof of Theorem 11.** We take a sample of sufficiently large size \( n \). We consider (B.14). From Lemma 8 we obtain
\[
\sum_{i=1}^{N} \left( \frac{(\pi j \gamma)^{2k+2}}{1 + (\pi j \gamma)^{2k+2}} \right)^2 \theta_j^2 \leq 2 \sum_{i=1}^{N} \left( \frac{a_j (\pi j \gamma)^{2k+2}}{j \left( 1 + (\pi j \gamma)^{2k+2} \right)} \right)^2 j^2 \\
+ 2\varepsilon_n^2 \sum_{i=1}^{N} \left( \frac{(\pi j \gamma)^{2k+2}}{1 + (\pi j \gamma)^{2k+2}} \right)^2 \leq \frac{4\pi^2 \delta}{\beta n} N^2 + 2\varepsilon_n^2 N.
\]

From Lemma 8 we also obtain
\[
\sum_{i=N+1}^{\infty} \left( \frac{a_j}{1 + (\pi j \gamma)^{2k+2}} \right)^2 = \sum_{i=N+1}^{\infty} \left( \frac{a_j (\pi j \gamma)^{2k+2}}{j \left( 1 + (\pi j \gamma)^{2k+2} \right)} \right)^2 j^2 \\
< \frac{2\delta}{\beta n \gamma^{4k+4} (\pi N)^{4k+2}}.
\]

We estimate the other terms on the right-hand side of (B.14) in the same way as in Theorem 9. Then
\[
\| \hat{f}^p_{\gamma}(x, X^n) - f(x) \|^2 \leq \frac{4\pi^2 \delta}{\beta n} N^2 + 2\varepsilon_n^2 N + \frac{2\delta}{\beta n \gamma^{4k+4} (\pi N)^{4k+2}} + 2 \frac{\varepsilon_n^2}{\pi \gamma} \\
+ \frac{4V_k^2}{(2k+1) N^{2k+1}}.
\]

Since \( N \) is an arbitrary number, we select \( N = n^{1/(2k+3)} \). By assumption, \( \gamma \) is obtained by the \( \omega^2 \) method. We assume that \( \gamma \geq (G/n)^{1/(2k+3)} \). Then we obtain
\[
\| \hat{f}^p_{\gamma}(x, X^n) - f(x) \|^2 \leq \tilde{c}_1 \varepsilon_n^2 n^{1/(2k+3)} + \tilde{c}_2 n^{-(2k+1)/(2k+3)},
\]
where
\[
\tilde{c}_1 = 2 \left( \pi^{-1} G^{-1/(2k+3)} + 1 \right),
\]
\[
\tilde{c}_2 = 2 \left( \frac{\delta}{\beta} \left( 2\pi^2 + \pi^{-4k+2} G^{-(4k+4)/(2k+3)} \right) + 2V_k^2 / (2k+1) \right).
\]

Then
\[
n^{(2k+1)/(2k+3)} \| \hat{f}^p_{\gamma}(x, X^n) - f(x) \|^2 \leq \tilde{c}_1 \varepsilon_n^2 n^{(2k+2)/(2k+3)} + \tilde{c}_2 = H_n + \tilde{c}_2.
\]
Thus, if \( H_n \leq 8\tilde{c}_1 \) and \( \gamma \geq (G/n)^{1/(2k+3)} \), then
\[
n^{(2k+1)/(2k+3)} \| \hat{f}_\gamma^\rho (x, X^n) - f(x) \|^2 \leq 8\tilde{c}_1 + \tilde{c}_2.
\]
Consequently,
\[
P\{ n^{(2k+1)/(2k+3)} \| \hat{f}_\gamma^\rho (x, X^n) - f(x) \|^2 > 8\tilde{c}_1 + \tilde{c}_2 \} < P\{ H_n > 8\tilde{c}_1 \}
+ P\{ \gamma < (G/n)^{1/(2k+3)} \}.
\]
In the same way as in Theorem 9, we estimate the first term on the right-hand side by Hoeffding’s inequality (Petrov, 1975):
\[
P\{ H_n > 8\tilde{c}_1 \} = P\{ \epsilon_n > 2\sqrt{2n^{-(k+1)/(2k+3)}} \} < 2n^{1/(2k+3)} \exp \left( -n^{1/(2k+3)} \right).
\]
We estimate the second term on the right-hand side by Theorem 10.

Now we consider the case when \( \gamma \) is obtained in accordance with part two of the \( \omega^2 \) method (4.33)–(4.35): \( \gamma = n^{-1/(2k+2)} \). Then by the proof of Theorem 9 we obtain that, for sufficiently large \( n \),
\[
P\{ n^{(2k+1)/(2k+3)} \| \hat{f}_\gamma^\rho (x, X^n) - f(x) \|^2 > 9 \} < 2n^{1/(k+1) - \pi/2}.
\]
Therefore, if \( \gamma \) is obtained by the \( \omega^2 \) method, then we have the inequality
\[
P\{ n^{(2k+1)/(2k+3)} \| \hat{f}_\gamma^\rho (x, X^n) - f(x) \|^2 > \max \{ 8\tilde{c}_1 + \tilde{c}_2, 9 \} \}
\leq \min \left\{ 2n^{1/(2k+3)} \exp \left( -n^{1/(2k+3)} \right) + 3n^{-9/8}, 2n^{1/(k+1) - \pi/2} \right\} = \psi(n).
\]
The series \( \sum_{n=1}^{\infty} \psi(n) \) converges and, by the Borel–Cantelli lemma, the assertion of Theorem 11 holds.
Appendix C

Proofs of Chapter 5

Proof of Theorem 12. Consider equation (5.4). In order to estimate the first term on the right-hand side, we note that

$$E(\hat{p}_i \hat{g}_i(x)) = p_i K_h * g_i(x),$$

where the asterisk denotes convolution. Hence

$$\int_{0}^{1} |q_i(T^{-1}(x))(p_i g_i(x) - E(\hat{p}_i \hat{g}_i(x)))| dx$$

$$= \int_{0}^{1} \left| q_i(T^{-1}(x)) \left( p_i g_i(x) \int_{-h}^{h} K_h(t) dt - p_i K_h * g_i(x) \right) \right| dx$$

$$\leq \int_{0}^{1} q_i(T^{-1}(x)) \int_{-1}^{1} |(p_i g_i(x - uh) - p_i g_i(x)) K(u)| du dx$$

$$\leq \sup_{x \in [0,1], |u| \leq 1} |g_i(x - uh) - g_i(x)| \int_{0}^{1} q_i(T^{-1}(x)) \int_{-1}^{1} K(u) du dx$$

$$= \sup_{x \in [0,1], |u| \leq 1} |g_i(x - uh) - g_i(x)| \int_{0}^{1} q_i(T^{-1}(x)) dx.$$

Since the PDF $g_i(x)$ is triangular, that is, $g_i(x) \simeq (1 - x) \mathbf{1}(x \in [0, 1])$, we have

$$g_i(x - uh) - g_i(x) \simeq uh.$$
By assumption (5.5),

\[
\int_0^1 |q_i(T^{-1}(x))(p_i g_i(x) - E(\hat{p}_i g_i(x)))| \, dx = O(h).
\]

We now estimate the second term on the right-hand side of (5.4). For the objects in the \(i\)th class, we have

\[
\hat{p}_i g_i(x) = \frac{1}{nh} \sum_{j=1}^n K \left( \frac{x - y_j}{h} \right) 1[\eta_j = i] \hat{p}_i, \quad 1 \leq i \leq M,
\]

where \(\eta_j\) are the labels of the objects. Therefore, \(\hat{p}_i g_i(x)\) can be expressed as

\[
\hat{p}_i g_i(x) = \frac{1}{h} \int_{0,|x-t|\leq h} K \left( \frac{x - t}{h} \right) \hat{p}_i dG_n^i(t),
\]

where \(G_n^i(t)\) is the empirical DF constructed from transformed observations \(y_j = T(x_j), j = 1, \ldots, n\), corresponding to the \(i\)th class. Obviously,

\[
E(\hat{p}_i g_i(x)) = \frac{1}{h} \int_{0,|x-t|\leq h} K \left( \frac{x - t}{h} \right) \hat{p}_i dG^i(t),
\]

where \(G^i(t)\) is the DF of \(y_j = T(x_j), j = 1, \ldots, n\), if \(O \in P_i\).

Hence we have

\[
\int_0^1 |q_i(T^{-1}(x))(\hat{p}_i g_i(x) - E(\hat{p}_i g_i(x)))| \, dx \\
\leq \frac{1}{h} \int_0^1 |q_i(T^{-1}(x))| \int_{0,|x-t|\leq h} K \left( \frac{x - t}{h} \right) d(G_n^i(t) - G^i(t)) \, dx \\
\leq \int_0^1 |q_i(T^{-1}(x))| \int_{-1}^1 K(u) d(G_n(x - uh) - G^i(x - uh)) \, dx \\
\leq C_1 \int_0^1 |G_n(x - h) - G^i(x - h) - (G_n(x + h) - G^i(x + h))| \, dx \\
\leq 2C_1 \sup_x |G_n^i(x) - G^i(x)|.
\]

Substituting into (5.4), we obtain

\[
\int_0^1 |q_i(x)(p_i f_i(x) - \hat{p}_i \hat{f}_i(x))| \, dx < 2C_1 \sup_x |G_n^i(x) - G^i(x)| + O(h).
\]

Since \(h = n^{-\beta}\), from (5.3) we obtain

\[
|\tilde{\mathcal{L}}(\eta_{EB}) - \mathcal{L}(\eta_B)| n^d < 2C_1 (M + 1) n^d \sup_x |G_n^i(x) - G^i(x)| + (M + 1) O(n^{d-\beta}).
\]

Let \(B_n\) denote the first term on the right-hand side of this inequality. For sufficiently large \(n\) and \(d < \beta\), the second term is less than one. Hence, if \(B_n \leq 1\), then \(n^d|\tilde{\mathcal{L}}(\eta_{EB}) - \mathcal{L}(\eta_B)| \leq 2\). Therefore,

\[
P\{n^d|\tilde{\mathcal{L}}(\eta_{EB}) - \mathcal{L}(\eta_B)| > 2\} < P\{B_n > 1\}.
\]
To estimate $P\{B_n > 1\}$ we use Prakasa Rao’s (1983) result,

$$P\{\sup_x |G'_n(x) - G'(x)| > \epsilon\} \leq 2\exp(-2\epsilon^2 n),$$

which is valid for any $i$. Hence

$$P\{B_n > 1\} \leq 2\exp\left(-\frac{n^{1-2d}}{2(M + 1)^2 C^2}\right).$$

Let $H(n, d)$ denote the right-hand side of this inequality. If $d < \min(0, 5; \beta)$, then the series $\sum_{n=1}^\infty H(n, d)$ converges and, by the Borel–Cantelli lemma, we arrive at the assertion of the theorem.

**Proof of Theorem 13.** Let us consider (5.4). By the Hölder inequality,

$$\left|\int_0^1 q_i(T^{-1}(x))(p_i g_i(x) - \hat{p}_i \hat{g}_i(x))\,dx\right|$$

\[ \leq \left(\int_0^1 |q_i(T^{-1}(x))|^2\,dx\right)^{\frac{1}{2}} \left(\int_0^1 |p_i g_i(x) - \hat{p}_i \hat{g}_i(x)|^2\,dx\right)^{\frac{1}{2}}. \quad (C.1) \]

Let us assume that

$$\left(\int_0^1 |p_i g_i(x) - \hat{p}_i \hat{g}_i(x)|^2\,dx\right)^{\frac{1}{2}} \leq n^{-d}.$$

Then, from (5.3)–(5.5) and (C.1), for $C_1 > 0$, we obtain

$$n^d |\mathcal{L}(\eta_{EB}) - \mathcal{L}(\eta_B)| \leq (M + 1)C_1.$$

Hence,

$$P\{n^d |\mathcal{L}(\eta_{EB}) - \mathcal{L}(\eta_B)| > (M + 1)C_1\} < P\left\{\left(\int_0^1 |p_i g_i(x) - \hat{p}_i \hat{g}_i(x)|^2\,dx\right)^{\frac{1}{2}} > n^{-d}\right\}$$

\[ \leq P\left\{\sup_{x \in [0, 1]} (g_i(x) - \hat{g}_i(x))^2 > n^{-2d}\right\} \]

\[ = P\{\hat{g}_i(x_*) > g_i(x_*) + n^{-d}\} + P\{\hat{g}_i(x_*) < g_i(x_*) - n^{-d}\}, \]

where $x_* \in [0, 1]$.

Let $z = \hat{g}_i(x)$ and $g_i = g_i(x)$. By the assumptions of the theorem, the polygram has an asymptotic distribution (as $n \to \infty$)

$$p(z) = \frac{(g_i/z)^{5/2}}{g_i \sqrt{2\pi/L}} \exp\left\{-\frac{1}{2} \left(\frac{1 - g_i/z}{g_i/(Lz)} + \frac{L (g_i/z)^2}{m(1 - g_i/(mz))}\right)\right\} (1 + O(1/\sqrt{n})).$$
(Tarasenko, 1976). Therefore,

\[ P\{\hat{g}_i(x) > g_i(x) + n^{-d}\} = \sqrt{\frac{L}{2\pi}} \int_{g_i + 1/n^d}^{\infty} \frac{(g_i/z)^{5/2}}{g_i} \cdot \exp \left\{ -\frac{1}{2} \left( \frac{(1 - g_i/z)^2}{g_i/(Lz)} + \frac{L (g_i/z)^2}{m(1 - g_i/(mz))} \right) \right\} \left(1 + O(1/\sqrt{n})\right) dz. \]

Applying the substitution \( u = g_i/z \), we obtain

\[ \ldots = \sqrt{\frac{L}{2\pi}} \int_{0}^{g_i/(g_i + 1/n^d)} \sqrt{u} \exp \left\{ -\frac{1}{2} \left( \frac{L(1-u)^2}{u} + \frac{Lu^2}{m(1-u/m)} \right) \right\} \left(1 + O(1/\sqrt{n})\right) du \]

\[ \leq \sqrt{L/(2\pi)} \exp \left\{ -\frac{L}{n^{2d}(g_i + n^{-d})g_i} \right\} \simeq \sqrt{L} \exp(-L/n^{2d}) \simeq \sqrt{n^\beta} \exp(-n^{\beta-2d}). \]

This can also be proved for \( P\{\hat{g}_i(x) < g_i(x) - n^{-d}\} \).

Let \( H(n, d, \beta) \) denote the right-hand side. Since \( 0 < d < \beta/2 \), we find that \( \sum_{n=1}^{\infty} H(n, d, \beta) \) converges and, by the Borel–Cantelli lemma, we arrive at the assertion of the theorem.
Appendix D

Proofs of Chapter 6

Proof of Theorem 14. Let us find first the distribution of \( \log (x_w/x_p) \). From (6.7) we have

\[
\hat{\gamma} \log c(\hat{\gamma}) = \hat{\gamma} \log \{1 + X_{(n-k)}^{-1/\hat{\gamma}} + X_{(n-k)}^{-2/\hat{\gamma}}\} \simeq \hat{\gamma} X_{(n-k)}^{-1/\hat{\gamma}} + o(X_{(n-k)}^{-1/\hat{\gamma}}). \tag{D.1}
\]

We use the relation \( X_{(i)} = d\, F^{-1}(\exp(-S_i)), 1 \leq i \leq n, \) where \( F(x) = 1 - F(x), S_i \) are order statistics, corresponding to the sample of size \( n \) of independent exponentially distributed r.v.s with unit expectation, and \( \exp(-S_i) \) are order statistics of the sample of independent uniform distributed r.v.s on \((0,1)\).

For distributions of Pareto type (6.3) we have

\[
x_p = d\, F^{-1}(p) = \left( \frac{c}{p} \right)^{\gamma} \{1 + \gamma d c^{-\gamma\beta} p^{\gamma\beta} + o(p^{\gamma\beta})\}, \tag{D.2}
\]

\[
X_{(n-k)} = d\, \left( \frac{c}{\exp(-S_{n-k})} \right)^{\gamma} \{1 + \gamma d c^{-\gamma\beta} \exp(-\gamma\beta S_{n-k}) + o(\exp(-\gamma\beta S_{n-k}))\}, \tag{D.3}
\]

\[
\log X_{(n-k)} = d\, \gamma (\log c + S_{n-k}) + \gamma d c^{-\gamma\beta} \exp(-\gamma\beta S_{n-k}) + o(\exp(-\gamma\beta S_{n-k})) \tag{D.4}
\]

(all equations are satisfied in probability). From (6.5) and (D.1)–(D.4) it follows that
\[
\log \left( \frac{x_{w}^{u}}{x_{p}} \right) = \log X_{(n-k)} - \log x_{p} + \gamma \log a_{n}
\]
\[= d \gamma \left( S_{n-k} - \log \left( \frac{n}{k} \right) \right) + (\hat{\gamma} - \gamma) \log \left( \frac{k}{np} \right)
\]
\[+ \gamma d c^{-\gamma \beta} \exp(\gamma \beta S_{n-k}) - p^{\gamma \beta} + o(\exp(\gamma \beta S_{n-k})) + o(p^{\gamma \beta}).
\]
From the Rényi representation
\[
S_{n-k} = d \sum_{j=1}^{n-k} \frac{z_{j}}{n-j+1} = \xi_{n-k} + T_{n-k},
\]
where \(z_{j}\) are independent exponentially distributed r.v.s with unit expectation, and
\[
T_{n-k} = d \sum_{j=1}^{n-k} \frac{z_{j} - 1}{n - j + 1},
\]
\[
\xi_{n-k} = \sum_{j=1}^{n-k} \frac{1}{n-j+1} = \sum_{j=k+1}^{n} \frac{1}{j} = \log \left( \frac{n}{k+1} \right) + O \left( \frac{1}{n} \right),
\]
for \(1 \leq k < n\), we get the expectation and the variance of \(S_{n-k}\):
\[
ES_{n-k} = \log \left( \frac{n}{k+1} \right) + O \left( \frac{1}{n} \right),
\]
\[
\text{var} S_{n-k} = \frac{1}{(k+1)^{2}} + \ldots + \frac{1}{n^{2}} \lesssim \frac{1}{k+1} - \frac{1}{n} + O \left( \frac{1}{k} \right).
\]
Furthermore, we have
\[
\exp(\gamma \beta S_{n-k}) = \exp(\gamma \beta (S_{n-k} - E(S_{n-k}))) \exp(-\gamma \beta (S_{n-k} - E(S_{n-k})))
\]
\[\approx \left( \frac{n}{k+1} \right)^{-\gamma \beta} \exp \left( -\gamma \beta \left( S_{n-k} - \log \left( \frac{n}{k+1} \right) \right) \right)
\]
\[= \left( \frac{k+1}{n} \right)^{\gamma \beta} \left( 1 + \gamma \beta \left( \log \frac{n}{k+1} - S_{n-k} \right) + o \left( \log \frac{n}{k+1} - S_{n-k} \right) \right).
\]
Taking into account (D.8), we consider equation (D.5). We obtain that
\[
\log \left( \frac{x_{w}^{u}}{x_{p}} \right) = d \left( S_{n-k} - \log \left( \frac{n}{k} \right) \right) \gamma \left( 1 - \gamma \beta d c^{-\gamma \beta} \left( \frac{k+1}{n} \right)^{\gamma \beta} \right)
\]
\[+ (\hat{\gamma} - \gamma) \log \left( \frac{k}{np} \right) + \gamma d c^{-\gamma \beta} \left( \left( \frac{k+1}{n} \right)^{\gamma \beta} - p^{\gamma \beta} \right).
\]
It is known that
\[ \sqrt{k} (\hat{\gamma} - \gamma) \rightarrow^d N(0, \gamma^2), \]
\[ \sqrt{k} \left( S_{n-k} - \log \left( \frac{n}{k} \right) \right) \rightarrow^d N(0, 1); \]
see Embrechts et al. (1997, p. 341) or Beirlant et al. (2004, p. 109). Thus
\[ \log \left( \frac{k}{np} \right) (\hat{\gamma} - \gamma) \rightarrow^d N \left( 0, \left( \log \left( \frac{k}{np} \right) \right)^2 \frac{\gamma^2}{k} \right), \]
\[ \sigma \left( S_{n-k} - \log \left( \frac{n}{k} \right) \right) + a \rightarrow^d N \left( a, \frac{\sigma^2}{k} \right), \quad \forall \ a, d. \quad (D.10) \]
Furthermore, we have
\[ \log \left( \frac{x_{np}^w}{x_p} \right) \rightarrow^d N \left( a, \frac{\sigma^2}{k} \right), \]
where
\[ a = \gamma d c^{-\gamma \beta} \left( \frac{k+1}{n} \right)^{\gamma \beta} - \left( \frac{1}{p} \right)^{\gamma \beta}, \quad \sigma = \gamma \left( 1 - \gamma \beta d c^{-\gamma \beta} \left( \frac{k+1}{n} \right)^{\gamma \beta} \right). \]
Thus,
\[ \frac{\log \left( \frac{x_{np}^w}{x_p} \right) - a}{\left( \frac{\sigma^2}{k} + \left( \log(k/(np)) \right)^2 \left( \gamma^2/k \right) \right)^{1/2}} \rightarrow^d N(0, 1). \quad (D.11) \]
We now find the DF for \( \log(x_{np}^c/x_p) \). Note that
\[ \hat{\gamma} \log c(\hat{\gamma}) = \log \left( 1 + X_{(n-k)}^{-1/\hat{\gamma}} + X_{(n-k)}^{-2/\hat{\gamma}} \right) \hat{\gamma} \approx X_{(n-k)}^{-1}. \]
We take, for simplicity,
\[ X_{(n-k)} = \bar{F}_n^{-1}(\exp(-S_{n-k})), \]
where \( \bar{F}_n(x) = x^{-1/\gamma} \), that is, \( X_{(n-k)}^{-1} = \exp(-\gamma S_{n-k}) \). From (D.8) we have
\[ X_{(n-k)}^{-1} = \frac{d}{k+1} \left( \frac{k+1}{n} \right)^{\gamma} \left[ 1 + \gamma \left( \log \left( \frac{n}{k+1} \right) - S_{n-k} \right) + o \left( \log \left( \frac{n}{k+1} \right) - S_{n-k} \right) \right]. \quad (D.12) \]
From (D.10) it follows that
\[ X_{(n-k)}^{-1} \rightarrow^d N \left( \left( \frac{k+1}{n} \right)^{\gamma}, \frac{\gamma^2}{k} \left( \frac{k+1}{n} \right)^{2\gamma} \right). \]
Thus from (6.9), (D.9), and (D.12) we obtain

\[
\log \left( \frac{x_p^c}{x_p} \right) = d \left( S_{n-k} - \log \left( \frac{n}{k} \right) \right) \gamma \left( 1 - \gamma \beta d e^{-\gamma \beta} \left( \frac{k+1}{n} \right) ^\gamma - \left( \frac{k+1}{n} \right) ^\gamma \right) \\
+ \left( \frac{k+1}{n} \right) ^\gamma + \gamma d e^{-\gamma \beta} \left( \frac{k+1}{n} \right) ^\gamma - p^\gamma \beta - \left( \frac{k+1}{n} \right) ^\gamma \\
+ \log \left( \frac{k}{np} \right) (\hat{\gamma} - \gamma) \\
\rightarrow d N \left( a + \left( \frac{k+1}{n} \right) ^\gamma, \frac{\sigma^2}{k} + \left( \log \left( \frac{k}{np} \right) \right) ^2 \frac{\chi^2}{k} \right),
\]

where \( \sigma_* = \sigma - \gamma \left( \frac{k+1}{n} \right) ^\gamma \) and it follows that

\[
\frac{\log(x_p^c/x_p) - \left( a + ((k+1)/n) ^\gamma \right)}{\left( \frac{\sigma^2}{k} + \left( \log \left( \frac{k}{np} \right) \right) ^2 \frac{\chi^2}{k} \right)^{1/2}} \rightarrow d N \left( 0, 1 \right). 
\]  

\quad (D.13)

The theorem follows from (D.11) and (D.13).
Appendix E

Proofs of Chapter 7

Proof of Lemma 4. We denote by $\| \cdot \|_C$ the norm in the space $C[0, x_a]$. Then,

$$\| y - y_n \|_C = \sup_x | - \ln(1 - F(x)) + \ln(1 - F_n(x) + \eta^*(x)) |$$

$$= \sup_x \left| \ln \frac{1 - F_n(x) + \eta^*(x)}{1 - F(x)} \right|$$

$$= \sup_x \left| \ln \left( 1 + \frac{F(x) - F_n(x) + \eta^*(x)}{1 - F(x)} \right) \right|.$$ 

Let us introduce the sets

$A = \{ x : 0 \leq F(x) - F_n(x) + \eta^*(x) \}$ and $B = \{ x : 0 \leq F_n(x) - \eta^*(x) - F(x) \}$.

Then for $x \in A$,

$$\| y - y_n \|_C = \sup_{x \in A} \left| \ln \left( 1 + \frac{F(x) - F_n(x) + \eta^*(x)}{1 - F(x)} \right) \right|$$

$$= \sup_{x \in A} \ln \left( 1 + \frac{|F(x) - F_n(x) + \eta^*(x)|}{1 - F(x)} \right)$$

$$= \ln \left( 1 + \sup_{x \in A} \frac{|F_n(x) - \eta^*(x) - F(x)|}{1 - F(x)} \right)$$

$$= \ln \left( 1 + \sup_{F(x) \leq a} \frac{|F_n(x) - \eta^*(x) - F(x)|}{1 - F(x)} \right),$$
and for \( x \in B \),
\[
\| y - y_n \|_C = \sup_{x \in B} \left| \ln \left( 1 + \frac{F(x) - F_n(x) + \eta^*(x)}{1 - F(x)} \right) \right|
\]
\[
= \sup_{x \in B} \left| \ln \left( 1 - \frac{F_n(x) - \eta^*(x) - F(x)}{1 - F(x)} \right) \right|
\]
\[
= - \ln \left( 1 - \sup_{x \in B} \frac{|F_n(x) - \eta^*(x) - F(x)|}{1 - F(x)} \right)
\]
\[
= - \ln \left( 1 - \sup_{F(x) \leq a} \frac{|F_n(x) - \eta^*(x) - F(x)|}{1 - F(x)} \right).
\]

Hence,
\[
\| y - y_n \|_C \leq \max \left\{ \ln \left( 1 + \sup_{F(x) \leq a} \frac{|F_n(x) - \eta^*(x) - F(x)|}{1 - F(x)} \right), \right.
\]
\[
- \ln \left( 1 - \sup_{F(x) \leq a} \frac{|F_n(x) - \eta^*(x) - F(x)|}{1 - F(x)} \right) \right\}
\]
\[
= - \ln \left( 1 - \sup_{F(x) \leq a} \frac{|F_n(x) - \eta^*(x) - F(x)|}{1 - F(x)} \right) = \varepsilon (n).
\]

**Proof of Theorem 18.** Let us consider the operator equation (7.2). Let \( g^\gamma \) be the regularized estimate and \( g \) be the solution of (7.2). The proof is based on the Theorem 17 in Stefanyuk (1986). We use Prakasa Rao’s (1983) inequality, according to which
\[
P \{ \sup_{x \in B} \frac{|F_n(x) - \eta^*(x) - F(x)|}{1 - F(x)} > \varepsilon (n) \} \leq 2 \exp(-2n\varepsilon^2)
\]
for sufficiently large \( n \), and obtain with regard to the inequality of Lemma 4 that
\[
P \left\{ \frac{\| y_n - y \|_C}{\sqrt{\gamma}} > c_1 \right\} \leq P \left\{ - \ln \left( 1 - \frac{\sup_{x} |F_n(x) - \eta^*(x) - F(x)|}{1 - a} \right) > c_1 \sqrt{\gamma} \right\}
\]
\[
\leq 2 \exp(-2n ((1 - a)(1 - \exp(-c_1 \sqrt{\gamma}))-1)^2) = P(n, \gamma, a).
\]
Since \( \gamma \) is defined so that \( \gamma = \gamma(n) \rightarrow 0 \) as \( n \rightarrow \infty \), \( \sum_{n=1}^{\infty} P(n, \gamma, a) < \infty \) holds. In view of the fact that the operator \( A \) is defined precisely, we get from (7.7) for the solution of equation (7.21) that
\[
P \{ \omega : \| h^\gamma - h \|_C > \varepsilon \} \leq P \left\{ \frac{\| y_n - y \|_C}{\sqrt{\gamma}} > c_1 \right\}.
\]
Then Theorem 18 follows from the Borel–Cantelli lemma.
Proof of Theorem 19. The inequality
\[
\| h(x) - h^\gamma(x; A, y_n) \| \leq \| h(x) - h^\gamma(x; A, y) \| \\
+ \| h^\gamma(x; A, y) - h^\gamma(x; A, y_n) \|, \tag{E.2}
\]
where the function \( h^\gamma(x; A, y) \) is the solution of the equation
\[
\gamma h + A^* A h = A^* y, \quad \gamma > 0,
\]
is valid for the error of the regularized estimate of (7.21). To represent the solution explicitly, we make use of the method of E. Schmidt (Ivanov et al., 1978). We have
\[
h^\gamma(t; A, y) = \sum_{i=1}^{\infty} \frac{\lambda_i}{1 + \gamma \lambda_i^2} c_i \psi_i(t) = \sum_{i=1}^{\infty} \frac{a_i}{1 + \gamma \lambda_i^2} \psi_i(t).
\]
Hence,
\[
\| h(t) - h^\gamma(t; A, y) \| = \left\| \sum_{i=1}^{\infty} \frac{\gamma \lambda_i^2}{1 + \gamma \lambda_i^2} a_i \psi_i(t) \right\|
\]
by virtue of (7.27). We apply the Parseval equality and obtain
\[
\| h(x) - h^\gamma(x; A, y) \|^2 = \int_0^\infty \left( h(t) - h^\gamma(t; A, y) \right)^2 dt = \sum_{i=1}^{\infty} \left( \frac{\gamma \lambda_i^2}{1 + \gamma \lambda_i^2} a_i \right)^2.
\]
We denote \( R_\gamma = (\gamma I + A^* A)^{-1} A^* \) and obtain the estimate of the second term on the left-hand side of (E.2):
\[
\| h^\gamma(x; A, y) - h^\gamma(x; A, y_n) \| = \left\| (\gamma I + A^* A)^{-1} A^* (y - y_n) \right\| \leq \| R_\gamma \| e(n). \tag{E.3}
\]
It follows from (7.25) and (7.28) that
\[
\sum_{i=1}^{\infty} \left( \frac{\gamma \lambda_i^2}{1 + \gamma \lambda_i^2} \right)^2 a_i^2 \leq \left( \frac{V_k}{x_a} \right)^2 \sum_{i=1}^{\infty} \left( \frac{\gamma \left( \frac{m_i}{x_a} \right)^2}{1 + \gamma \left( \frac{m_i}{x_a} \right)^2} \right)^2 \frac{1}{i^{2(k+1)}}.
\]
For an arbitrary integer \( N \),
\[
\sum_{i=1}^{\infty} \left( \frac{\gamma \left( \frac{m_i}{x_a} \right)^2}{1 + \gamma \left( \frac{m_i}{x_a} \right)^2} \right)^2 \frac{1}{i^{2(k+1)}} \leq \left( \frac{V_k}{x_a} \right)^2 \sum_{i=1}^{N} \left( \frac{\gamma \left( \frac{m_i}{x_a} \right)^2}{1 + \gamma \left( \frac{m_i}{x_a} \right)^2} \right)^2 \frac{1}{i^{2(k+1)}} + \sum_{i=N+1}^{\infty} \left( \frac{\gamma \left( \frac{m_i}{x_a} \right)^2}{1 + \gamma \left( \frac{m_i}{x_a} \right)^2} \right)^2 \frac{1}{i^{2(k+1)}}. \tag{E.4}
\]
We estimate the first term on the right-hand side of (E.4). For $k \in \{0, 1\}$ and $N \leq \frac{1}{\sqrt{\gamma}}$, we get

$$\sum_{i=1}^{N} \left( \frac{\gamma \left( \frac{m_i}{\lambda_a} \right)^2}{1 + \gamma \left( \frac{m_i}{\lambda_a} \right)^2} \right)^2 \frac{1}{t^{2(k+1)}} = \left( \frac{\pi \sqrt{\gamma}}{\lambda_a} \right)^{2k+2} \sum_{i=1}^{N} \left[ \left( \frac{m_i \sqrt{\gamma}}{\lambda_a} \right)^2 \left( \frac{m_i}{\lambda_a \sqrt{\gamma}} \right)^k \right]^{2k+1} = O \left( \frac{1}{\gamma^{2k+1}} \right).$$

For $k \geq 2$, we get

$$\sum_{i=1}^{N} \left( \frac{\gamma \left( \frac{m_i}{\lambda_a} \right)^2}{1 + \gamma \left( \frac{m_i}{\lambda_a} \right)^2} \right)^2 \frac{1}{t^{2(k+1)}} = O \left( \gamma^2 \right).$$

Turning to the second term on the right-hand side of (E.4), we have that

$$\sum_{i=N+1}^{\infty} \left( \frac{\gamma \left( \frac{m_i}{\lambda_a} \right)^2}{1 + \gamma \left( \frac{m_i}{\lambda_a} \right)^2} \right)^2 \frac{1}{t^{2(k+1)}} \leq \sum_{i=N+1}^{\infty} \frac{1}{t^{2(k+1)}} \leq \frac{1}{(2k+1)N^{2k+1}}.$$

Since $N$ is an arbitrary number, we take

$$N = \min \left\{ \left[ \frac{1}{\sqrt{\gamma}} \right], \left[ \left( \frac{1}{\gamma} \right)^{\frac{2}{\pi \pi}} \right] \right\},$$

where $[\cdot]$ is the integer part of a real number, and obtain that

$$\| h(x) - h^\gamma(x; A, y) \|^2 = \begin{cases} O \left( \gamma^{(2k+1)/2} \right), & k = \{0, 1\}, \\ O \left( \gamma^2 \right), & k \geq 2. \end{cases} \quad (E.5)$$

We now estimate $\| R_\gamma \|$ (see (E.3)). Let $h(x) = \sum_{i=1}^{\infty} h_i \phi_i(x) \in L_2[0, x_a]$. Then,

$$R_\gamma h = \sum_{i=1}^{\infty} \frac{\lambda_i}{1 + \gamma \lambda_i^2} h_i \phi_i(t).$$

We obtain from the definition of norm and the Parseval equality that

$$\| R_\gamma \|^2 = \sup \left\{ \sum_{i=1}^{\infty} \left( \frac{\lambda_i}{1 + \gamma \lambda_i^2} \right)^2 h_i^2 : \sum_{i=1}^{\infty} h_i^2 \leq 1 \right\},$$

where the supremum is taken over sequences $\{ \lambda_i \}$ and $\{ h_i \}$. The function $g(\lambda) = \frac{\lambda}{1 + \gamma \lambda^2}$ at $\lambda = \frac{1}{\sqrt{\gamma}}$ reaches the maximum $1/(2\sqrt{\gamma})$. Hence,

$$\| R_\gamma \| \leq \frac{1}{2\sqrt{\gamma}}.$$
Then,
\[
\|h^\gamma(x; A, y) - h^\gamma(x; A, y_n)\| \leq \frac{1}{2\sqrt{\gamma}} \varepsilon(n).
\]

From this and from (E.2) and (E.5), we obtain
\[
\|h(x) - h^\gamma(x; A, y_n)\| \leq \frac{\varepsilon(n)}{2\sqrt{\gamma}} + \begin{cases} 
  c\gamma^{(2k+1)/4}, & k \in \{0, 1\}, \\
  c\gamma, & k \geq 2,
\end{cases}
\]
where \(c\) is a constant independent of \(n\).

Let \(\gamma = n^{-\alpha}, \alpha > 0\), be a constant. Then, for some \(\beta > 0\) and \(k \in \{0, 1\}\),
\[
n^\beta \|h(x) - h^\gamma(x; A, y_n)\| \leq A_n + B_n,
\]
where \(A_n = cn^{\beta - \alpha(2k+1)/4}\) and \(B_n = \varepsilon(n)n^{\frac{\alpha}{2} + \beta}/2\), while for \(k \geq 2\),
\[
n^\beta \|h(x) - h^\gamma(x; A, y_n)\| \leq C_n + B_n,
\]
where \(C_n = cn^{\beta - \alpha}\).

Let us consider the case \(k \geq 2\). Since \(\alpha \geq \beta\), for sufficiently large \(n\) we have that \(C_n \leq 1\) and \(B_n\) is a r.v. If \(B_n \leq 1\), then
\[
n^\beta \|\delta(x) - \delta^\gamma(x; A, y_n)\| \leq 2.
\]

Consequently,
\[
P\{n^\beta \|\delta(x) - \delta^\gamma(x; A, y_n)\| > 2\} < P\{B_n > 1\}.
\]

The right-hand side is estimated using inequality (E.1). By Lemma 4,
\[
P\{B_n > 1\} = P \left\{ - \ln \left( 1 - \sup_{F(x) \leq a} \frac{|F_n(x) - \eta^*(x) - F(x)|}{1 - F(x)} \right) > 2n^{-\frac{\alpha}{2} - \beta} \right\}
\]
\[
= P \left\{ \sup_x |F_n(x) - \eta^*(x) - F(x)| > \left( 1 - \exp\left(-2n^{-\frac{\alpha}{2} - \beta}\right) \right)(1-a) \right\}
\]
\[
\leq 2 \exp\left(-2n \left( (1-a) \left( 1 - \exp\left(-2n^{-\frac{\alpha}{2} - \beta}\right) \right) - 1 \right)^2 \right) = \psi(n).
\]

Since \(\alpha < 1 - 2\beta\), the series \(\sum_{n=1}^{\infty} \psi(n) < \infty\), and the assertion of the theorem is valid by the Borel–Cantelli lemma.

We now turn to the case \(k \in \{0, 1\}\). Since \(\alpha \geq 4\beta/(2k+1)\) or
\[
\beta - \alpha - \frac{2k+1}{4} \leq 0,
\]
\(A_n \leq C\) for sufficiently large \(n\) and some constant \(C\). If \(B_n \leq 1\), then
\[
n^\beta \|h(x) - h^\gamma(x; A, y_n)\| \leq 1 + C.
\]
Therefore,

\[ P\left\{ n^{\beta} \| h(x) - h^\gamma(x; A, y_n) \| > 1 + C \right\} < P\{ B_n > 1 \}. \]

The assertion of the theorem follows the same arguments as given before for estimating \( P\{ B_n > 1 \} \).

**Proof of Lemma 5.** Let us estimate the inaccuracy of defining the operator \( \| A_n g - Ag \|_C = \sup_{x} |A_n g - Ag| \leq \zeta(n, g), x \in [0, x_a] \), for equation (7.31). We have

\[ \| A g - A_n g \|_C \leq \sup_{x} \left| \int_{0}^{x} (I(y) - I_n (y)) g(x - y) dy \right| \]

\[ + \sup_{x} \left| \int_{0}^{x} (I_n (y) H_n (y) - I(y) H(y)) g(x - y) dy \right|. \tag{E.7} \]

Let us estimate the first term on the right-hand side of (E.7):

\[ \sup_{x} \left| \int_{0}^{x} (I(y) - I_n (y)) g(x - y) dy \right| \leq \sup_{x} \left| \int_{0}^{x} (I(y) - I_n (y)) g(x - y) dy \right| \]

\[ \leq \sup_{x} |I(x) - I_n (x)| \int_{0}^{x} g(x - y) dy \]

\[ = \sup_{x} |I(x) - I_n (x)|, \tag{E.8} \]

since \( g(x) \) is the PDF.

Turning to the second term of the right-hand side of (E.7), since

\[ |I_n (y) H_n (y) - I(y) H(y)| = |H_n (y) (I_n (y) - I(y)) + I(y) (H_n (y) - H(y))| \]

\[ \leq |I_n (y) - I(y)| + |I(y) (H_n (y) - H(y))|, \]

we obtain that

\[ \sup_{x} \left| \int_{0}^{x} (I_n (y) H_n (y) - I(y) H(y)) g(x - y) dy \right| \]

\[ \leq \sup_{x} \left| \int_{0}^{x} (I_n (y) H_n (y) - I(y) H(y)) g(x - y) dy \right| \]

\[ \leq \sup_{x} |I_n (x) - I(x)| + \sup_{x} |I(x) H_n (x) - H(x)|. \tag{E.9} \]
Then from (E.7)–(E.9) it follows that
\[ \|A_n g - Ag\|_C \leq 2\sup_x |I_{n_2}(x) - I(x)| + \sup_I(x) \sup_x |H_{n_1}(x) - H(x)|. \] 
\hfill (E.10)

In the notation of (7.34) we have
\[ |I_{n_2}(x) - I(x)| = \left| \frac{f_{n_2}(x)}{1 - F_{n_2}(x) + \eta^+(x)} - \frac{f(x)}{1 - F(x)} \right| \]
\[ \leq \frac{|f_{n_2}(x) - f(x)|}{(1 - a)C^*} \left( 1 + \sup_x \left| F_{n_2}(x) - F(x) \right| \right) + f(x) \sup_x \left| F(x) - F_{n_2}(x) + \eta^+(x) \right|. \] 
\hfill (E.11)

Furthermore, it follows that
\[ \sup_I(x) = \sup_x \frac{f(x)}{1 - F(x)} \leq \sup_x \frac{f(x)}{1 - a}. \]

Hence, from (E.10) and (E.11) we get the assertion of the lemma.

**Proof of Theorem 20.** We denote \( \sup_{x} = \sup_{x \in [0, x_1]} \). Note that \( \sup_{x} \eta^+(x) = \max(1/n, 1 - a) = \eta_{\max} \). Hence, by virtue of the condition, it follows from Lemma 5 that
\[ \|A_n g - Ag\|_C \leq \frac{2}{(1 - a)C^*} \left[ \sup_x \left| f_{n_2}^\gamma(x) - f(x) \right| \left( 1 + \sup_x \left| F_{n_2}(x) - F(x) \right| \right) \right. \]
\[ + \varepsilon \left( \eta_{\max} + \sup_x \left| F_{n_2}(x) - F(x) \right| \right) \] 
\[ \left. + \sup_x \left| H_{n_1}(x) - H(x) \right| \right]. \]

We fix the constants \( c_1, c_2 > 0 \) and \( c_3 > 2\varepsilon \eta/(C^*(1 - a)\sqrt{\gamma\Omega_{\min}}) \) and denote
\[ A = \left\{ \omega : \frac{4}{(1 - a)C^*} \sup_x \left| f_{n_2}^\gamma(x) - f(x) \right| \leq c_1 \sqrt{\gamma\Omega_{\min}} \right\}, \]
\[ B = \left\{ \omega : \frac{\varepsilon}{1 - a} \sup_x \left| H_{n_1}(x) - H(x) \right| \leq c_2 \sqrt{\gamma\Omega_{\min}} \right\}, \]
\[ C = \left\{ \omega : \frac{2\varepsilon}{(1 - a)C^*} \left( \sup_x \left| F_{n_2}(x) - F(x) \right| + \eta_{\max} \right) \leq c_3 \sqrt{\gamma\Omega_{\min}} \right\}. \]

If the events \( A, B \) and \( C \) occur simultaneously, then for any function \( g(x) \) we have
\[ \|A_n g - Ag\|_C \leq (c_1 + c_2 + c_3) \sqrt{\gamma\Omega_{\min}}. \]

Then,
\[ P \{ \|A_n g - Ag\|_C > (c_1 + c_2 + c_3) \sqrt{\gamma\Omega_{\min}} \} \leq P\{ \tilde{A} \} + P\{ \tilde{B} \} + P\{ \tilde{C} \}. \]

It follows from (7.8) that
We denote \( \overline{c_2} = c_2 (1 - a) \sqrt{\gamma \Omega_{\min}} / \varepsilon \) and \( \overline{c_3} = c_3 (1 - a) C^* \sqrt{\gamma \Omega_{\min}} / (2\varepsilon) - \eta_{\max} \). Then we obtain from (E.1) that

\[
P \left\{ \sup_x |H_{n_1}(x) - H(x)| > \overline{c_2} \right\} \leq 2 \exp \left( -2n_1 \overline{c_2}^2 \right),
\]

(E.13)

\[
P \left\{ \sup_x |F_{n_2}(x) - F(x)| > \overline{c_3} \right\} \leq 2 \exp \left( -2n_2 \overline{c_3}^2 \right).
\]

(E.14)

By the to statistical regularization theory (see p. 184) the regularized estimates \( f_{n_2}^\gamma(x) \) and \( y_{n_1}^\gamma(x) \) converge to the true PDFs \( f(x) \) and \( y(x) \) with probability one under the conditions (7.32) on \( \gamma(n) \) assumed in the theorem for \( \mu = \min(\mu_1, \mu_2) \). Hence, from (7.6) we get

\[
P \left\{ \sup_x |y_{n_1}^\gamma(x) - y(x)| > \nu_1 \right\} \leq 2 \exp \left( -n_1 \mu_1 \gamma \right),
\]

(E.15)

\[
P \left\{ \sup_x |f_{n_2}^\gamma(x) - f(x)| > \nu_2 \right\} \leq 2 \exp \left( -n_2 \mu_2 \gamma \right),
\]

(E.16)

for some numbers \( N_1 = N_1(\nu_1, \mu_1) \), \( N_2 = N_2(\nu_2, \mu_2) \) and all \( n_1 > N_1, n_2 > N_2 \). Here \( \nu_1, \nu_2, \mu_1 \) and \( \mu_2 \) are any positive numbers.

Let \( \nu_1 = c_1 (1 - a) C^* \sqrt{\gamma \Omega_{\min}} / 4 \) and \( \nu_2 = c \sqrt{\gamma} \). It follows from (7.7) and (E.12)–(E.16) that

\[
P \left\{ \| g^\gamma - g \|_C > \varepsilon \right\} \leq P \left\{ \frac{\| y_{n_1}^\gamma(x) - y(x) \|_C}{\sqrt{\gamma}} > c \right\} + P \left\{ \frac{\| A_n - A \|}{\sqrt{\gamma}} > c_1 + c_2 + c_3 \right\}
\]

\[
\leq 2 \exp \left( -n_1 \mu_1 \gamma \right) + \exp \left( -n_2 \mu_2 \gamma \right) + \exp \left( -2n_1 \overline{c_2}^2 \right)
\]

\[
+ \exp \left( -2n_2 \overline{c_3}^2 \right).
\]

Hence, for the chosen sequence \( \gamma = \gamma(n) \) we get the assertion of the theorem.

**Proof of Theorem 21.** For brevity, we denote \( g^\gamma = g^\gamma(x), g = g(x), h^\gamma = h^\gamma(x), h = h(x), G^\gamma = G^\gamma(x), G = G(x) \). It follows from \( G^\gamma \leq C \) and \( G(x) = b < 1 \) that
\[
\|h^\gamma - h\|_C = \left\| \frac{g^\gamma}{1 - G^\gamma} - \frac{g}{1 - G} \right\|_C = \sup_{x \in [0, a]} \left| \frac{g^\gamma - g + g(G^\gamma - G) - G(g^\gamma - g)}{(1 - G^\gamma)(1 - G)} \right| \\
\leq \frac{1}{1 - C} \left[ \sup_x |g^\gamma - g| + \frac{\sup_x (g)}{1 - b} \sup_x |G^\gamma - G| \right].
\]

(E.17)

Additionally,

\[
\|G^\gamma - G\|_C = \sup_x \left| \int_0^x (g^\gamma(\tau) - g(\tau)) \, d\tau \right| \leq \sup_x |g^\gamma(\tau) - g(\tau)|_a.
\]

The theorem follows from this fact, (7.33), and (E.17) by virtue of the fact that \(g(x)\) is bounded.
Appendix F

Proofs of Chapter 8

Proof of Theorem 23. By (8.5) we get for $0 \leq t \leq t_{\text{max}}(k)$:

$$\sup_{t} |H(t) - \tilde{H}(t, k, l)| \leq \sup_{t} \sum_{n=k+1}^{\infty} P \{ t_n < t \} + k \max_{1 \leq n \leq k} \sup_{t} |P \{ t_n < t \} - F_{t_n}(t)|.$$

Under the conditions of the theorem it follows from well-known results that

$$P \left\{ \frac{t_n - n\mu}{\sigma \sqrt{n}} < t \right\} = \Phi(t) + \sum_{i=1}^{m-2} \frac{Q_i(t)}{n^{i/2}} + o \left( n^{-(m-2)/2} \right) \quad (F.1)$$

uniformly in $t \in (-\infty, \infty)$, where $Q_i$ are expressions involving the PDF $\varphi(x) = (2\pi)^{-1/2} \exp\{-x^2/2\}$ of the standard normal DF $\Phi(x)$, the Hermite polynomials and semi-invariants of $\tau_i$ (see Embrechts et al. 1997, Theorem 2.3.2, p. 85; Petrov, 1975). In particular,

$$Q_1(x) = \varphi(x) \frac{1 - x^2}{6} \frac{E(\tau_1 - \mu)^3}{\sigma^2}$$

and, for $m = 3$,

$$P \left\{ \frac{t_n - n\mu}{\sigma \sqrt{n}} < t \right\} = \Phi(t) + \frac{E(\tau_1 - \mu)^3}{6\sigma^3 \sqrt{n}} (1 - t^2) \frac{1}{\sqrt{2\pi}} \exp\{-t^2/2\} + o \left( \frac{1}{\sqrt{n}} \right)$$
Hence, \( P \{ t_n < t \} \) is defined like the right-hand side of (F.1) with the replacement of \( t \) by \((t - n\mu)/(\sigma \sqrt{n})\). Then \( \sum_{n=1}^{\infty} P \{ t_n < t \} \) converges for \( m \geq 3 \) and
\[
\sum_{n=k+1}^{\infty} P \{ t_n < t \} \leq c, \quad c > 0
\] (F.2)
holds. If
\[
k \max_{1 \leq n \leq k} \sup_t |P \{ t_n < t \} - F_{t_n}(t)| \leq \eta
\]
holds for any constant \( \eta > 0 \), then at \( t \in [0, t_{\max}(k)] \),
\[
\sup_t |H(t) - \tilde{H}(t, k, l)| \leq c + \eta
\]
follows. Hence,
\[
P \left\{ \sup_t |H(t) - \tilde{H}(t, k, l)| > c + \eta \right\} < P \left\{ k \max_{1 \leq n \leq k} \sup_t |P \{ t_n < t \} - F_{t_n}(t)| > \eta \right\}.
\]
The right-hand side may be estimated using the asymptotical estimate of the convergence rate of the empirical DF to the true DF (Prakasa Rao, 1983),
\[
P \left\{ \sup_t |P \{ t_n < t \} - F_{t_n}(t)| > \eta \right\} \leq 2 \exp \left( -2l_\eta \eta^2 \right), \quad (F.3)
\]
which is satisfied for sufficiently large \( l \). Then it follows that
\[
P \left\{ \sup_t |H(t) - \tilde{H}(t, k, l)| > c + \eta \right\} < 2 \exp \left( -2l_\eta \eta^2 \right) = P(\eta, l, k)
\]
Since \( k \sim l^\rho, 0 < \rho < 1/3 \), the series \( \sum_{l=1}^{\infty} P(\eta, l, k) \) converges at least for one \( \eta > 0 \), and by the Borel–Cantelli, the assertion of the theorem follows.

**Proof of Theorem 24.** Using (8.8) we have, for \( t \in [0, 1] \),
\[
\sup_t |H(t) - \tilde{H}(t, k, l)| \leq (1 - \exp(-\nu))^{k+1} \exp(\nu)
\]
\[+ k \max_{1 \leq n \leq k} \sup_t |P \{ t_n < t \} - F_{t_n}(t)|\]
and, for \( \alpha > 0 \),
\[
l^\alpha \sup_t |H(t) - \tilde{H}(t, k, l)| \leq l^\alpha (1 - \exp(-\nu))^{k+1} \exp(\nu)
\]
\[+ l^\alpha k \max_{1 \leq n \leq k} \sup_t |P \{ t_n < t \} - F_{t_n}(t)|.\]
Since \( k = c(\nu) \cdot l^\rho \), where \( \rho < 1/3 - (2/3)\alpha, 0 < \alpha < 0.5, \rho > 0 \), for sufficiently large \( l \) and the corresponding \( c(\nu) \) we get
\[
l^\alpha (1 - \exp(-\nu))^{k+1} \exp(\nu) \leq 1;
\]
therefore, if

\[ l^\alpha k \max_{1 \leq n \leq k} \sup_t |P \{ t_n < t \} - F_{t_n}(t) | \leq \eta \]

for any constant \( \eta > 0 \), then it follows that

\[ l^\alpha \sup_t |H(t) - \tilde{H}(t, k, l)| \leq 1 + \eta. \]

Hence,

\[ P \left\{ l^\alpha \sup_t |H(t) - \tilde{H}(t, k, l)| > 1 + \eta \right\} < P \left\{ l^\alpha k \max_{1 \leq n \leq k} \sup_t |P \{ t_n < t \} - F_{t_n}(t) | > \eta \right\} . \]

Using (F.3) we have

\[ P \left\{ l^\alpha \sup_t |H(t) - \tilde{H}(t, k, l)| > 1 + \eta \right\} < 2 \exp \left( -2 \frac{l}{k^3} \left( \frac{\eta}{l^\alpha} \right)^2 \right) = P(\eta, l, k) \quad (F.4) \]

Since \( k = c(\nu) \cdot l^\rho \) and \( \alpha + 1.5\rho < 0.5 \), the series \( \sum_{l=1}^{\infty} P(\eta, l, k) \) converges at least for one \( \eta > 0 \), and by the Borel–Cantelli lemma, the assertion of the theorem holds.

**Proof of Corollary 2.** Let the right-hand side of (F.4) be equal to \( 0 < \chi < 1: \)

\[ 2 \exp \left( -2 \frac{l}{k^3} \left( \frac{\eta}{l^\alpha} \right)^2 \right) = \chi \]

Hence, we have

\[ \eta = kl^\alpha \sqrt{-\frac{k \ln (\chi/2)}{2l}}. \]

This gives the level of the confidence interval \( D = (1 + \eta)l^{-\alpha}. \)

**Proof of Theorem 25.** Since, for \( t \in (0, \frac{a}{h_k}) \), expression (8.10) is valid for sufficiently large \( n \),

\[ \sum_{n=k+1}^{\infty} P \{ t_n < t \} \sim \sum_{n=k+1}^{\infty} \Phi \left( \frac{t}{\sqrt{n}} \right) \]

holds. The expansion on the right-hand side converges. Therefore, \( \sum_{n=k+1}^{\infty} P \{ t_n < t \} < c \) follows, where \( c > 0 \) is a constant. The rest of the proof is similar to the proof of Theorem 23.

**Proof of Theorem 26.** For \( t \in [a, t_{\max}(k)] \), \( a > 0 \) we have, from (8.11),

\[ \sup_t \left( \frac{(F(t))^{k+1}}{1 - F(t)} \right) = \sup_t \left( \frac{(1 - \ell(t))^{-\alpha}}{\ell(t)^{-\alpha}} \right) = \sup_t \left( \frac{(1 - t^{-\alpha}c(t) \exp(\int_{x_0}^{t} \frac{\varepsilon(y)}{y} dy))^{k+1}}{t^{-\alpha}c(t) \exp(\int_{x_0}^{t} \frac{\varepsilon(y)}{y} dy)} \right) . \]

The mean value theorem implies

\[ \exp \left\{ \int_{x_0}^{t} \frac{\varepsilon(y)}{y} dy \right\} = \exp \left\{ \varepsilon(\theta) \ln \left( \frac{t}{x_0} \right) \right\} = \left( \frac{t}{x_0} \right)^{\varepsilon(\theta)}, \]

where \( \theta \) is a number between \( x_0 \) and \( t \).
for some \( \theta \in [x_0, t] \). Hence,

\[
\sup_t \frac{(F(t))^{k+1}}{1 - F(t)} = \sup_t \frac{\left(1 - c(t) t^{-\alpha + \varepsilon(\theta)} x_0^{-\varepsilon(\theta)}\right)^{k+1}}{c(t) t^{-\alpha + \varepsilon(\theta)} x_0^{-\varepsilon(\theta)}}.
\]

Since \( \varepsilon(x) \) is nonpositive, \( -\alpha + \varepsilon(\theta) < 0 \). Then we have

\[
\sup_t \frac{(F(t))^{k+1}}{1 - F(t)} = \frac{\left(1 - c_{\inf}(t_{\max}(k))^{-\alpha + \varepsilon(\theta)} x_0^{-\varepsilon(\theta)}\right)^{k+1}}{c_{\inf}(t_{\max}(k))^{-\alpha + \varepsilon(\theta)} x_0^{-\varepsilon(\theta)}},
\]

where

\[
c_{\inf} = \inf_t c(t) = \begin{cases} 
  c(t_{\max}(k)), & \text{if } c(t) \text{ is a monotone decreasing function} \\
  c(a), & \text{if } c(t) \text{ is a monotone increasing function}.
\end{cases}
\]

Since \( c(t_{\max}(k)) > c_0 \), \( c_{\inf} \geq \min(c_0, c(a)) = c^* \) and the right-hand side of (F.5) is less than

\[
\frac{\left(1 - c^*(t_{\max}(k))^{-\alpha + \varepsilon(\theta)} x_0^{-\varepsilon(\theta)}\right)^{k+1}}{c^*(t_{\max}(k))^{-\alpha + \varepsilon(\theta)} x_0^{-\varepsilon(\theta)}}.
\]

Assume that

\[
\max_i \tau_i \leq l^\eta,
\]

where \( \eta > 1 / (\alpha - c^*) \). Then

\[
t_{\max}(k) \leq l \max_{i=1, \ldots, l} \tau_i < l^{1+\eta}.
\]

This implies that (F.6) is less than or equal to

\[
\frac{\left(1 - c^* l^{(1+\eta)(-\alpha + \varepsilon(\theta))} x_0^{-\varepsilon(\theta)}\right)^{k+1}}{c^* l^{(1+\eta)(-\alpha + \varepsilon(\theta))} x_0^{-\varepsilon(\theta)}}.
\]

Then from (F.5) we have

\[
\sup_t \beta \left(\frac{F(t)^{k+1}}{1 - F(t)}\right)^{k+1} \leq \beta \frac{\left(1 - c^* l^{(1+\eta)(-\alpha + \varepsilon(\theta))} x_0^{-\varepsilon(\theta)}\right)^{k+1}}{c^* l^{(1+\eta)(-\alpha + \varepsilon(\theta))} x_0^{-\varepsilon(\theta)}}.
\]

Since \( (1 + \eta)(-\alpha + \varepsilon(\theta)) < 0 \), for sufficiently large \( l \) the right-hand side of the latter inequality is less than or equal to 1 for \( k \geq -A \cdot l^\rho \) and \( \rho \geq 0 \). Note that, for \( \beta > 0 \) and sufficiently large \( l \), \( A < 0 \) holds. Therefore, if

\[
l^\beta k \max_{1 \leq n \leq k} \left| \sup_t P\{t_n < t\} - F_{\lambda_n}(t) \right| \leq \chi
\]

\[\text{ (F.8)}\]

1 One can obtain this result by taking the left-hand side of the latter inequality to be less than or equal to 1.
for any constant $\chi > 0$, it follows from (8.5) and (8.6) that
\[ l^\beta \sup_t |H(t) - \tilde{H}(t, k, l)| \leq 1 + \chi. \]
Hence, from (F.7) and (F.8),
\[
P[l^\beta \sup_t |H(t) - \tilde{H}(t, k, l)| > 1 + \chi] < P[l^\beta k \max_{1 \leq n \leq k} |P(t_n < t) - F_{t_n}(t)| > \chi] \\
+ P[\max_i \tau_i > l^\eta]
\]
follows. By the global property of the regularly varying r.v.s (see Embrechts et al., 1997, p. 38) we have
\[
P[\max_i \tau_i > x] \sim lP[\tau_i > x] = lx^{-\alpha} \ell(x), \quad \text{as } x \to \infty.
\]
From the representation theorem (8.11) the following property of slowly varying functions follows (Mikosch, 1999): for every $\varepsilon^* > 0$,
\[ x^{-\varepsilon^*} \ell(x) \to 0 \quad \text{and} \quad x^{\varepsilon^*} \ell(x) \to \infty \quad \text{as } x \to \infty. \]
This implies that there exists $T > 0$ such that, for $x > T$,
\[ x^{-\varepsilon^*} \leq \ell(x) \leq x^{\varepsilon^*}. \]
Since $\varepsilon^* < \alpha$ it follows that
\[
P[\max_i \tau_i > l^\eta] \sim l^{1+\eta(\varepsilon^*-\alpha)} \quad \text{as } l \to \infty.
\]
Using (F.3), we finally have
\[
P[l^\beta \sup_t |H(t) - \tilde{H}(t, k, l)| > 1 + \chi] < l^{1+\eta(\varepsilon^*-\alpha)} \\
+ 2 \exp\left(-2\chi^2 l^{1-2\beta}/k^3\right) = P(\eta, l, k). \quad \text{(F.9)}
\]
Since $k = dl^\rho$ and $0 < \rho < (1 - 2\beta)/3$, $\eta > 1/(\alpha - \varepsilon^*)$ hold, the series $\sum_{l=1}^\infty P(\eta, l, k)$ converges at least for one $\eta > 0$, and by the Borel–Cantelli lemma, the assertion of the theorem holds.

**Proof of Corollary 3.** Let the right-hand side of (F.9) be equal to $0 < \nu < 1$:
\[ l^{1-\eta(\alpha-\varepsilon^*)} + 2 \exp\left(-2\chi^2 l^{1-2\beta}/k^3\right) = \nu. \]
Hence, we have
\[
\chi = kl^\beta \sqrt{-\frac{k}{2l} \ln \left( \frac{\nu - l^{1-\eta(\alpha-\varepsilon^*)}}{2} \right)}.
\]
This gives the level of the confidence interval $D = (1 + \chi)l^{-\beta}$. 
List of Main Symbols and Abbreviations

General guidelines
In general, Greek letters represent parameters.

$X^n = \{X_1, X_2, \ldots, X_n\}$ data sample
$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ order statistics
$R$ real line
$R_+$ nonnegative real line
$Z$ set of integers
$n$ sample size
$h$ bandwidth in kernel estimators
$A^T$ transpose of matrix $A$
$\text{rank } A$ rank of matrix $A$
$\alpha$ and $\gamma = 1/\alpha$ tail index and extreme value index
$\gamma = \gamma(n)$ regularization parameter

Let the functions $f(x)$ and $g(x)$ be defined on some set $M$ and let $a$ be a limit point of $M$.

$f(x) \sim g(x) (x \to a, x \in M)$ denotes a function $f(x)$ that satisfies
\[ \lim_{x \to a, x \in M} \frac{f(x)}{g(x)} = 1 \]

$f(x) = o(g(x)) (x \to a, x \in M)$ denotes a function $f(x)$ that satisfies
\[ \lim_{x \to a, x \in M} \frac{f(x)}{g(x)} = 0 \]

$f(x) = O(g(x)) (x \in M)$ denotes, for some constant $C > 0$, the inequality
\[ |f(x)| \leq C|g(x)| \text{ for all } x \in M \]
**Probabilities**

- **$P(\cdot)$**: probability measure
- **$E(\cdot)$**: expectation
- **$\text{var}(\cdot)$**: variance
- **$\text{bias}(\cdot)$**: bias
- **$\text{cov}(X, Y)$**: covariance between the r.v.s $X$ and $Y$
- **$N(\mu, \sigma)$**: normal (or Gaussian) density with mean $\mu$ and variance $\sigma^2$
- **$\Phi(x)$**: the distribution function of the standard normal distribution
- **$f(x), g(x)$**: probability density functions
- **$H(x)$**: renewal function
- **$h(x)$**: hazard rate function

**Functions**

- **$A(x)$**: Pickands’ dependence function
- **$\Gamma(x)$**: gamma function
- **$\ell(x)$**: slowly varying function
- **$K(x)$**: kernel function
- **$T(x)$**: transformation function of the data
- **$F^{-1}(x)$**: inverse function

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

\[
\theta(t) = \begin{cases} 
  1, & t \geq 0, \\ 
  0, & t < 0
\end{cases}
\]

\[
1(A) = 1(x \in A) = \begin{cases} 
  1, & x \in A, \\ 
  0, & \text{otherwise}
\end{cases}
\]

**Spaces**

- **$C[a, b]$**: space of all continuous real-valued functions defined on the closed interval $[a, b]$ with norm

\[
\|x\|_C = \max_{a \leq t \leq b} |x(t)|
\]

- **$\ell_p$**: space of sequences $x = (x_1, x_2, \ldots, x_n, \ldots)$ of real numbers such that $\|x\|_{\ell_p} = \left(\sum_{k=1}^{\infty} |x_k|^p\right)^{1/p} < \infty, p \geq 1$

- **$\ell_2$**: Hilbert space $\ell_2^p, p = 2$, with a scalar product $(x, y) = \sum_{k=1}^{\infty} x_k y_k$

- **$L_p[a, b], p \geq 1$**: space of functions with the integral $\int_{a}^{b} |x(t)|^p dt < \infty$ and norm $\|x\|_{L_p} = \left(\int_{a}^{b} |x(t)|^p dt\right)^{1/p}$

- **$H^\mu[a, b]$**: Hölder space with norm

\[
\|x\|_{H^\mu} = \sup_{t \in [a, b]} |x(t)| + \sup_{t_1, t_2 \in [a, b], t_1 \neq t_2} \frac{|x(t_1) - x(t_2)|}{|t_1 - t_2|^\mu}
\]
Signs

≃ asymptotic equality

Abbreviations

ACF autocorrelation function
AMISE asymptotical mean integrated squared error
ARMA autoregressive moving average process
BISDN Broadband integrated services digital network
DF distribution function
EPM elemental percentile method
EVI extreme value index
EVD extreme value distribution
GARCH generalized autoregressive conditionally heteroscedastic process
GE Gilbert–Elliott model
GEV generalized extreme value
GPD generalized Pareto distribution
HTML hypertext markup language
i.i.d. independent and identically distributed
ITD inter-arrival time distribution
MA moving average process
MDA maximum domain of attraction
MIAE mean integrated absolute error
MISE mean integrated squared error
ML maximum likelihood
MSE mean squared error
PDF probability density function
POT peaks over threshold
PWM method of probability-weighted moments
RF renewal function
r.v. random variable
SMIL Synchronized Multimedia Integration Language
TCP transmission control protocol
UMTS Universal Mobile Telecommunications System
WWW World Wide Web
References


REFERENCES


Teugels, J.L. (1968) Renewal theorems when the first or the second moment is infinite. *Annals of Statistics* 39, 1210–1219.
REFERENCES

Index

algorithm
- Bayesian classification 117
- for boundary kernel selection 139
- of density estimation based on adaptive transformation 128
- of density estimation based on fixed transformation 125
- of double bootstrap 12
- of sequential procedure 12
asymptotical mean integrated squared error 115

bandwidth of kernel 70
Bartlett’s formula 44
Bayesian risk of misclassification 155
bin width of histogram 76
BISDN 208
bivariate quantile curve 55
bootstrap
  - classical 10, 369
  - estimate of renewal function 228
  - non-classical 369
  - re-sample 9, 229
boundary effect of kernel estimates 73
censoring 94
characteristic number 192
classifier 152
  - Bayesian 152
  - empirical Bayesian 152
component-wise maxima 50

condition
- Cramér’s 5, 225
- Hall’s 8
- Hölder (or Lipschitz) 68, 190
- mixing 42
- von Mises 180
confidence interval
  - of group estimate 18
  - of high quantile 172
  - of renewal function 226, 227, 228
consistency
  - strong 73
  - weak 73
convex hull 50
copula 33, 49
cross-validation 74, 77
  - for dependent data 91
  - integrated squared error 78, 115
  - weighted integrated squared error 115
density estimation approach
  - $L_1$ 62
  - $L_2$ 63
  - $\chi^2$ 63
dependence
  - index sequence 86
  - long range 45, 86, 89
  - short range 45, 89
distribution
  - bivariate extreme value 49
  - exponential 234

Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice  N. Markovich
distribution (Continued)
  extreme value xiii
  fitted 119, 129
  Fréchet 23
  function 2
  gamma 234
  generalized extreme value 3, 54
  generalized hyperbolic 92
  generalized Pareto xiii, 14, 129
  heavy-tailed 3
  interarrival-time 220
  isosceles triangular 118
  light-tailed 3
  normal inverse Gaussian 92
  Pareto 239, 241
  Poisson 208
  regularly varying 4, 168, 227
  stable 18
  subexponential 3
  target 119, 129
  Weibull 172, 188, 234
dose–effect dependence 204
eigenfunction 192
equation
  Fredholm’s 66
  Volterra’s 185, 198, 211
estimate
  maximum likelihood 167
  re-transformed 118
estimator
  Barron 112
  based on exponential regression model 14
  combined parametric-nonparametric 101, 165
  EVI kernel 13
  Frees’ 223, 233
  group 19
  Hill’s 6
  histogram-type of renewal function 224
  intensity of nonhomogeneous Poisson process 209
  Kaplan–Meier 206
  modified Weissman’s 166
  moment 13
  on-line 20
  Parzen–Rosenblatt kernel 68, 70
  Pickands’ 13
  polygram 76, 126, 131
  POT 14, 165
  projection 63, 74
  ratio 7
  smoothed projection 69
  UH 13
  variable bandwidth kernel 113
  weighted quantile 164
  Weissman’s 165
  Euler’s constant 231
  expected shortfall 92
failure time 181, 199
Fourier coefficients 193
frailty 201, 202
function
  autocorrelation 43
  auto-covariance 45
  covariance 42
  empirical distribution 67
  empirical mean excess 8, 28
  hazard rate 180
  Laplace’s 25
  leave-out- l cross-validation 91
  mean excess 28
  moment generating 5, 225
  Pickands dependence 49
  ratio of the hazard rates 181, 198
  renewal 220, 221
  sample autocorrelation 43
  sample heavy-tailed autocorrelation 43
  slowly varying 4
  survival 94, 181
functional
  regularization 68
  stabilizing 67
high quantile 163
Hill plot 39
hormesis 200
independent random
  variables 2
index
  extreme value 3
  tail 3
inequality
  Hoeffding’s 260, 262, 266
  Hölder 269
intensity of nonhomogeneous Poisson process 208
inter-arrival times 221

kernel
  boundary 132
  Epanechnikov’s 71
  Gaussian 71
  modified bi-weight 136
  triangular 132
Kullback’s metric 62

Laplace–Stieltjes transform 221
leave-out sequence 91

lemma
  Borel–Cantelli 256, 258, 261, 266, 286, 287, 289
  of inverse operator 183
lifetime 181
likelihood ratio 198

maximum domain of attraction 3
mean integrated squared error 135
mean risk 104
mean squared error 9, 72, 88

method
  block maxima xiii
  $D$ 82, 127
  discrepancy 80, 115, 119, 127
  elemental percentile 15, 167
  exponent 64
  Lagrange 183
  least-squares 81
  maximum likelihood 53, 62, 167
  of mismatch 184
  of moments 15
  POT xiii, 176
  of probability-weighted moments 15, 54, 167
  regularization 67, 183
  structural risk-minimization 104
  Xie’s RS 234
  $\omega^2$ 82, 127, 147
model
  Gilbert–Elliott Markov 210
  Pareto-type 165
  retrial 213
  semi-Markov 182, 193
  Monte-Carlo study 23
  mortality
    risk 94, 180
    table 94
  operator
    adjoint 184
    self-adjoint (Hermitian) 192
  operator equation 182
  orbit 212
  order of kernel 70
  order statistics 6
  orthonormalized system 192
  outliers xii, 117
  over-smoothing bandwidth selector 77

parameter
  Hurst 45
  smoothing 76
Parseval equality 277, 278
pattern recognition 151
plot of histogram-type estimate 232
probability
  density function 2
  of misclassification 152
  space 1, 182
problem
  correct by Hadamard 67
  ill-posed 67, 182
  inverse 181
process
  ARMA 43
  exactly second-order self similar 47
  GARCH 44
  log-return 91
  MA 45
  second order stationary 45
QQ plot 28, 165
quantile 163
random variable 2
regularization parameter 67, 183
representation
  Jenkinson–von Mises 3
  Karamata of slowly varying function 227
  Rényi 272
retrial call
  definition A 213
  definition B 213
retrial queues 212
right endpoint of distribution 5, 28

scheme
  Fisher’s 62
  transform–retransform 117, 128
second-order asymptotic relation 18
selection of $k$ in Hill’s estimate
  bootstrap 9
  double bootstrap 12
  exceedance plot 8
  Hill’s plot 8
  sequential procedure 12
space
  Hölder 190
  Sobolev 65
statistic
  Kolmogorov–Smirnov 80
  Mises–Smirnov 80, 148
  Rényi 191
  Stieltjes convolution 221
  Taylor’s expansion 255
  TCP-flow 55
  theorem
    Glivenko–Cantelli 198, 211
    Lebesgue’s 61
    Pickands’ 5
    Scheffé’s 63
    Sklar’s 33, 49
    Smith’s 221
  total variation 63
  transformation
    adaptive 128
    fixed 118, 124
    function 117
  transmission control protocol 30
  u-statistic 223
  value at risk 92
  Vapnik–Chervonenkis dimension 104
warranty control 221
wavelet basis 75
Web prefetching 161
Web traffic 30
The Wiley Series in Probability and Statistics is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

ABRAHAM and LEDOLTER · Statistical Methods for Forecasting
AGRESTI · Analysis of Ordinal Categorical Data
AGRESTI · An Introduction to Categorical Data Analysis
AGRESTI · Categorical Data Analysis, Second Edition
ALTMAN, GILL, and MCDONALD · Numerical Issues in Statistical Computing for the Social Scientist
AMARATUNGA and CABRERA · Exploration and Analysis of DNA Microarray and Protein Array Data
ANDÉL · Mathematics of Chance
ANDERSON · An Introduction to Multivariate Statistical Analysis, Third Edition
*ANDERSON · The statistical Analysis of Time Series
ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE, and WEISBERG · Statistical Methods for Comparative Studies
ANDERSON and LOYNES · The Teaching of Practical Statistics
ARMITAGE and DAVID · (EDITORS) Advances in Biometry
ARNOLD, BALAKRISHNAN, and NAGARAJA · Records
*ARTHANARI and DODGE · Mathematical Programming in Statistics
*BAILEY · The Elements of Stochastic Processes with Applications to the Natural Sciences
BALAKRISHNAN and KOUTRAS · Runs and Scans with Applications
BALAKRISHNAN AND NG · Precedence-Type Tests and Applications
BARNETT · Comparative Statistical Inference, Third Edition
BARNETT · Environmental Statistics: Methods & Applications
BARNETT and LEWIS · Outliers in Statistical Data, Third Edition
BARTOSZYNSKI and NIEWIADOMSKA-BUGAJ · Probability and Statistical Inference
BASILEVSKY · Statistical Factor Analysis and Related Methods: Theory and Applications
BASU and WATTS · Nonlinear Regression Analysis and Its Applications
BECHHOFER, SANTNER, and GOLDSMAN · Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons
BELSLEY · Conditioning Diagnostics: Collinearity and Weak Data in Regression
BELSLEY, KUH, and WELSCH · Regression Diagnostics: Identifying Influential Data and Sources of Collinearity
BENDAT and PIERSOL · Random Data: Analysis and Measurement Procedures, Third Edition
BERNARDO and SMITH · Bayesian Theory
BERRY, CHALONER, and GEWEKE · Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner
BHAT and MILLER · Elements of Applied Stochastic Processes, Third Edition
BHATTACHARYA and JOHNSON · statistical Concepts and Methods

*N. Markovich
CSÖRGÖ and HORVÁTH · Limit Theorems in Change Point Analysis
DANIEL · Applications of Statistics to Industrial Experimentation
DANIEL · Biostatistics: A Foundation for Analysis in the Health Sciences, Sixth Edition
*DANIEL · Fitting Equations to Data: Computer Analysis of Multifactor Data, Second Edition
DASU and JOHNSON · Exploratory Data Mining and Data Cleaning
DAVID and NAGARAJA · Order Statistics, Third Edition
*DEGROOT, FIENBERG, and KADANE · Statistics and the Law
DEL CASTILLO · Statistical Process Adjustment for Quality Control
DEMARISS · Regression with Social Data: Modeling Continuous and Limited Response Variables
DEMIDENKO · Mixed Models: Theory and Applications
DENISON, HOLMES, MALLICK, and SMITH · Bayesian Methods for Nonlinear Classification and Regression
DETTETE and STUDDEN · The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis
DEY and MUKERJEE · Fractional Factorial Plans
DILLON and GOLDSTEIN · Multivariate Analysis: Methods and Applications
DODGE · Alternative Methods of Regression
*DODGE and ROMIG · Sampling Inspection Tables, Second Edition
*DOOB · Stochastic Processes
DOWDY, WEARDEN, and CHILKO · Statistics for Research, Third Edition
DRAPER and SMITH · Applied Regression Analysis, Third Edition
DREYDEN and MARDIA · Statistical Shape Analysis
DUDEWICZ and MISHRA · Modern Mathematical Statistics
DUNN and CLARK · Applied Statistics: Analysis of Variance and Regression, Second Edition
DUNN and CLARK · Basic Statistics: A Primer for the Biomedical Sciences, Third Edition
DUPUIS and ELLIS · A Weak Convergence Approach to the Theory of Large Deviations
EDLER and KITSOS (editors) Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment
*ELANDT-JOHNSON and JOHNSON · Survival Models and Data Analysis
ENDERS · Applied Econometric Time Series
ETHIER and KURTZ Markov · Processes: Characterization and Convergence
EVANS, HASTINGS, and PEACOCK · Statistical Distribution, Third Edition
FISHER and VAN BELLE · Biostatistics: A Methodology for the Health Sciences
FITZMAURICE, LAIRD, and WARE · Applied Longitudinal Analysis
*FLEISS · The Design and Analysis of Clinical Experiments
FLEISS · Statistical Methods for Rates and Proportions, Second Edition
FLEISS and HARRINGTON · Counting Processes and Survival Analysis
FULLER · Introduction to Statistical Time Series, Second Edition
FULLER · Measurement Error Models
GALLANT · Nonlinear Statistical Models
GEISSER · Modes of Parametric Statistical Inference
GELMAN and MENG (editors) Applied Bayesian Modeling and Casual Inference from Incomplete-data Perspectives
GEWEKE · Contemporary Bayesian Econometrics and Statistics
GHOSH, MUKHOPADHYAY, and SEN · Sequential Estimation
GIESBRECHT and GUMPERTZ · Planning, Construction, and Statistical Analysis of Comparative Experiments
GIFI · Nonlinear Multivariate Analysis
GIVENS and HOETING · Computational Statistics
GLASSERMAN and YAO · Monotone Structure in Discrete-Event Systems
GNANADESIKAN · Methods for Statistical Data Analysis of Multivariate Observations, Second Edition
GOLDSTEIN and LEWIS · Assessment: Problems, Development, and Statistical Issues
GREENWOOD and NIKULIN · A Guide to Chi-Squared Testing

*Now available in a lower priced paperback edition in the Wiley Classics Library
ROSSI, ALLENBY, and MCCULLOCH · Bayesian Statistics and Marketing
ROUSSEEUW and LEROY · Robust Regression and Outline Detection
RUBIN · Multiple Imputation for Nonresponse in Surveys
RUBINSTEIN · Simulation and the Monte Carlo Method
RUBINSTEIN and MELAMED · Modern Simulation and Modeling
RYAN · Modern Regression Methods
RYAN · Statistical Methods for Quality Improvement, Second Edition
SALEH · Theory of Preliminary Test and Stein-Type Estimation with Applications
SALTELLI, CHAN, and SCOTT · (editors) Sensitivity Analysis
*SCHIEFF · The Analysis of Variance
SCHIMIEK · Smoothing and Regression: Approaches, Computation, and Application
SCHOTT Matrix Analysis for Statistics
SCHOUTENS · Levy Processes in Finance: Pricing Financial Derivative
SCHUSS · Theory and Applications of Stochastic Differential Equations
SCOTT · Multivariate Density Estimation: Theory, Practice, and Visualization
*SPEAR · Linear Models
SEARLE · Linear Models for Unbalanced Data
SEARLE · Matrix Algebra Useful for Statistics
SEARLE and WILLETT · Matrix Algebra for Applied Economics
SEBER · Multivariate Observations
SEBER and LEE · Linear Regression Analysis, Second Edition
SEBER and WILD · Nonlinear Regression
SENNOTT · Stochastic Dynamic Programming and the Control of Queueing Systems
*SERRLING · Approximation Theorems of Mathematical Statistics
SFAER and VOYK · Probability and Finance: Its Only a Game!
SILVAPOULUL and SEN · Constrained Statistical Inference: Inequality, Order, and Shape Restrictions
SINGPURWALLA · Reliability and Risk: A Bayesian Perspective
SMALL and MCLEISH · Hilbert Space Methods in Probability and Statistical Inference
SRIVASTAVA · Methods of Multivariate Statistics
STAPLETON · Linear Statistical Models
STAUDT and SHEATHER · Robust Estimation and Testing
STOYAN, KENDALL, and MECKE · Stochastic Geometry and Its Applications, Second Edition
STOYAN and STOYAN · Fractals, Random and Point Fields: Methods of Geometrical Statistics
SUTTON, ABRAMS, JONES, SHELDON, and SONG · Methods for Meta-Analysis in Medical Research
TANAKA · Time Series Analysis: Nonstationary and Noninvertible Distribution Theory
THOMPSON · Empirical Model Building
THOMPSON · Sampling, Second Edition
THOMPSON · Simulation: A Modeler’s Approach
THOMPSON and SEBER · Adaptive Sampling
THOMPSON, WILLIAMS, and FINDLAY · Models for Investors in Real World Markets
TIAO, BISGAARD, HILL, PEÑA, and STIGLER · (editors) Box on Quality and Discovery: with Design, Control, and Robustness
TIERNEY · LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics
TSAY · Analysis of Financial Time Series
UPTON and FINGLETON · Spatial Data Analysis by Example, Volume II: Categorical and Directional Data
VAN BELLE · Statistical Rules of Thumb
VAN BELLE, FISHER, HEAGERTY, and LUMLEY · Biostatistics: A Methodology for the Health Sciences, Second Edition
VESTRUP · The Theory of Measures and Integration
VIDAKOVIC · Statistical Modeling by Wavelets
VINOD and REAGLE · Preparing for the Worst: Incorporating Downside Risk in Stock Market Investments
WALLER and GOTWAY · Applied Spatial Statistics for Public Health Data

*Now available in a lower priced paperback edition in the Wiley Classics Library
WEERAHANDI - Generalized Inference in Repeated Measures: Exact Methods in MANOVA and Mixed Models
WEISBERG - Applied Linear Regression, Second Edition
WELISH - Aspects of Statistical Inference
WESTFALL and YOUNG - Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment
WHITTAKER - Graphical Models in Applied Multivariate Statistics
WINKER - Optimization Heuristics in Economics: Applications of Threshold Accepting
WONNACOTT and WONNACOTT - Econometrics, Second Edition
WOODING - Planning Pharmaceutical Clinical Trials: Basic Statistical Principles
WOOLSON and CLARKE - Statistical Methods for the Analysis of Biomedical Data, Second Edition
WU and HAMADA - Experiments: Planning, Analysis, and Parameter Design Optimization
WU and ZHANG - Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches
YANG - The Construction Theory of Denumerable Markov Processes
*ZELLMER - An introduction to Bayesian Inference in Econometrics
ZELTERMAN - Discrete Distributions: Applications in the Health Sciences
ZHOU, OBUCHOWSKI and McCLISH - Statistical Methods in Diagnostic Medicine

*Now available in a lower priced paperback edition in the Wiley Classics Library