Tail Distribution and Extreme Quantile Estimation Using Non-parametric Approaches

Imen Rached[1](✉) and Elisabeth Larsson[2]

[1] Allianstic Research Laboratory, EFREI Paris, Ecole d'Ingénieurs Généraliste Informatique et Technologies du Numérique,
30–32 Avenue de la République, 94800 Villejuif, France
imen.rached@efrei.fr

[2] Scientific Computing, Department of Information Technology, Uppsala University,
Box 337, 751 05 Uppsala, Sweden
elisabeth.larsson@it.uu.se

Abstract. Estimation of tail distributions and extreme quantiles is important in areas such as risk management in finance and insurance in relation to extreme or catastrophic events. The main difficulty from the statistical perspective is that the available data to base the estimates on is very sparse, which calls for tailored estimation methods. In this chapter, we provide a survey of currently used parametric and non-parametric methods, and provide some perspectives on how to move forward with non-parametric kernel-based estimation.

Keywords: Risk measures · Extreme value theory · Kernel estimation · Bandwidth selection

1 Introduction

This chapter presents a position survey on the overall objectives and specific challenges encompassing the state of the art in tail distribution and extreme quantile estimation of currently used parametric and non-parametric approaches and their application to Financial Risk Measurement. What is envisioned, is an enhanced non-parametric estimation method based on the Extreme Value Theory approach. The compounding perspectives of current challenges are addressed, like the threshold level of excess data to be chosen for extreme values and the bandwidth selection from a bias reduction perspective. The application of the kernel estimation approach and the use of Expected Shortfall as a coherent risk measure instead of the Value at Risk are presented. The extension to multivariate data is addressed and its challenges identified.

Overview of the Following Sections. In the following sections, Financial risk measures are presented in Sect. 2. Section 3, covers Extreme Value Theory, Sect. 4,

J. Kołodziej and H. González-Vélez (Eds.): cHiPSet, LNCS 11400, pp. 69–87, 2019.
https://doi.org/10.1007/978-3-030-16272-6_3

Parametric estimation and Semi-parametric estimation methods, Sect. 5, Non-Parametric estimation methods and Sect. 6, the perspectives identified by the addressed challenges when estimating the presented financial risk measures.

2 Financial Risk Measures

The Long Term Capital Management collapse and the 1998 Russian debt crisis, the Latin American and Asian currency crises and more recently, the U.S. mortgage credit market turmoil, followed by the bankruptcy of Lehman Brothers and the world's biggest-ever trading loss at Société Générale are some examples of financial disasters during the last twenty years. In response to the serious financial crises, like the recent global financial crisis (2007–2008), regulators have become more concerned about the protection of financial institutions against catastrophic market risks. We recall that market risk is the risk that the value of an investment will decrease due to movements in market factors. The difficulty of modelling these rare but extreme events has been greatly reduced by recent advances in Extreme Value Theory (EVT). Value at Risk (VaR) and the related concept of Expected Shortfall (ES) have been the primary tools for measuring risk exposure in the financial services industry for over two decades. Additional literature can be found in [39] for Quantitative Risk Management and in [42] or in [25] for the application of EVT in insurance, finance and other fields.

2.1 Value at Risk

Consider the loss X of a portfolio over a given time period δ, then VaR is a risk statistic that measures the risk of holding the portfolio for the time period δ. Assume that X has a cumulative distribution function (cdf), F_X, then we define VaR at level $\alpha \in (0,1)$ as

$$\mathrm{VaR}_{\alpha}^{\delta}(X) = \inf\{x \in \mathbb{R} : P(X > x) \leq 1 - \alpha\} = \inf\{x \in \mathbb{R}, F_X(x) \geq \alpha\} = F_X^{\leftarrow}(\alpha), \quad (1)$$

$F_X^{\leftarrow}$ is the generalized inverse of the cdf F_X. Typical values of α are 0.95 and 0.99, while δ usually is 1 day or 10 days. Value-at-Risk (VaR) has become a standard measure for risk management and is also recommended in the Basel II accord. For an overview on VaR in a more economic setting we refer to [37] and [23]. Despite its widespread use, VaR has received criticism for failing to distinguish between light and heavy losses beyond the VaR. Additionally, the traditional VaR method has been criticized for violating the requirement of sub-additivity [4]. Artzner et al. analysed risk measures and stated a set of properties/axioms that should be desirable for any risk measure. The four axioms they stated are:

- Monotonicity: Higher losses mean higher risk.
- Translation Equivariance: Increasing (or decreasing) the loss increases (decreases) the risk by the same amount.
- Subadditivity: Diversification decreases risk.
- Positive Homogeneity: Doubling the portfolio size doubles the risk.

Any risk measure which satisfies these axioms is said to be coherent. A related concept to VaR, which accounts for the tail mass is the conditional tail expectation (CVaR), or Expected Shortfall (ES). ES is the average loss conditional on the VaR being exceeded and gives risk managers additional valuable information about the tail risk of the distribution. Due to its usefulness as a risk measure, in 2013 the Basel Committee on Bank Supervision has even proposed replacing VaR with ES to measure market risk exposure.

2.2 Conditional Value at Risk or Expected Shortfall

$$\mathrm{CVaR}_\alpha(X) := E[X|X \geq \mathrm{VaR}_\alpha(X)] \tag{2}$$

Acerbi and Tasche proved in [1] that CVaR satisfies the above axioms and is therefore a coherent risk measure.

Conditional Value-at-Risk can be derived from VaR in the case of a continuous random variable and another possibility to calculate CVaR is to use Acerbi's Integral Formula:

$$\mathrm{CVaR}_\alpha(X) = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_\beta(X)d\beta \tag{3}$$

Estimating ES from the empirical distribution is generally more difficult than estimating VaR due to the scarcity of observations in the tail. As in most risk applications, we do not need to focus on the entire distribution. Extreme value theory is then a practical and useful tool for modeling and quantifying risk. Value at Risk and Extreme value theory is covered well in most books on risk management and VaR in particular (also ES with much less extent), see for example [33,37,39], and [22]. Vice versa, VaR is treated in some Extreme value theory literature, such as [26] and [17].

3 Extreme Value Theory: Two Main Approaches

Extreme value theory (EVT) is the theory of modelling and measuring events which occur with very small probability: More precisely, having an $X_1, ..., X_n$ sample of n random variables independently and identically following a distribution function $F(\cdot)$, we want to estimate the real x_{p_n} defined by

$$x_{p_n} = \bar{F}^{\leftarrow}(p_n), \quad \text{with} \quad p_n < 1/n. \tag{4}$$

where p_n is a known sequence and $\bar{F}^{\leftarrow}(u) = \inf\{x \in \mathbb{R}, \bar{F}(x) \leq u\}$. $\bar{F}^{\leftarrow}$ is the generalized inverse of the survival function $\bar{F}(\cdot) = 1 - F(\cdot)$. Note that x_{p_n} is the order quantile $1 - p_n$ of the cumulative distribution function F.

A similar problem to the estimate of x_{p_n} is the estimate of "small probabilities" p_n or the estimation of the tail distribution. In other words, for a series of fixed (c_n) reals, we want to estimate the probability p_n defined by

$$p_n = P(X > c_n), \quad \text{with} \quad c_n > x_{n,n}. \tag{5}$$

The main result of Extreme Value Theory states that the tails of all distributions fall into one of three categories, regardless of the overall shape of the distribution. Two main approaches are used for implementing EVT in practice: Block maxima approach and Peaks Over Thresholds (POT).

3.1 Block Maxima Approach

The Fisher and Tippett [29] and Gnedenko [30] theorems are the fundamental results in EVT. The theorems state that the maximum of a sample of properly normalized independent and identically distributed random variables converges in distribution to one of the three possible distributions: the Weibull, Gumbel or the Fréchet.

Theorem 1 (Fisher, Tippett, Gnedenko).
Let $X_1, ..., X_n \sim^{i.i.d.} F$ and $X_{1,n} \leq ... \leq X_{n,n}$. If there exist two sequences a_n and b_n and a real γ such that

$$P\left(\frac{X_{n,n} - a_n}{b_n} \leq x\right) \longrightarrow H_\gamma(x), \tag{6}$$

when $n \longrightarrow +\infty$, then

$$H_\gamma(x) = \begin{cases} \exp\left(-(1+\gamma x)^{-1/\gamma}\right) & \textit{if } \gamma \neq 0, \quad 1+\gamma x > 0. \\ \exp(-\exp(-x)) & \textit{if } \gamma = 0, \quad x \in \mathbb{R}. \end{cases} \tag{7}$$

We say that F is in the domain of attraction of H_γ and denote this by $F \in \mathrm{DA}(H_\gamma)$. The distribution function $H_\gamma(\cdot)$ is called the Generalized Extreme Value distribution (GEV).

This law depends only on the parameter called the tail index. The density associated is shown in Fig. 1 for different values of γ. According to the sign of γ, we define three areas of attraction:

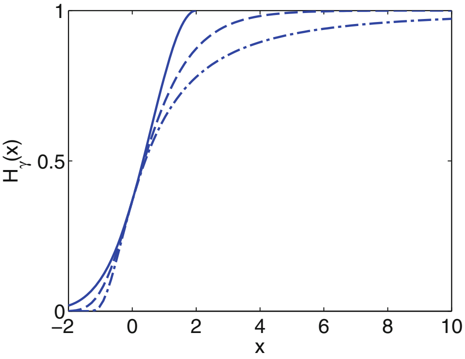

Fig. 1. The GEV distribution for $\gamma = -0.5$ (solid line), $\gamma = 0$ (dashed line), and $\gamma = 0.5$ (dash-dot line).

- If $\gamma > 0$, $F \in$ DA (Fréchet): This domain contains the laws for which the survival function decreases as a power function. Such tails are know as "fat tails" or "heavy tails". In this area of attraction, we find the laws of Pareto, Student, Cauchy, etc.
- If $\gamma = 0$, $F \in$ DA (Gumbel): This domain groups laws for which the survival function declines exponentially. This is the case of normal, gamma, log-normal, exponential, etc.
- if $\gamma < 0$, $F \in$ DA (Weibull): This domain corresponds to thin tails where the distribution has a finite endpoint. Examples in this class are the uniform and reverse Burr distributions.

The Weibull distribution clearly has a finite endpoint ($s_+(F) = \sup\{x, \, F(x) < 1\}$). This is usually the case of the distribution of mortality and insurance/reinsurance claims for example, see [20]. The Fréchet tail is thicker than the Gumbel's. Yet, it is well known that the distributions of the return series in most financial markets are heavy tailed (fat tails). The term "fat tails" can have several meanings, the most common being "extreme outcomes occur more frequently than predicted by the normal distribution".

The block Maxima approach is based on the utilization of maximum or minimum values of these observations within a certain sequence of constant length. For a sufficiently large number k of established blocks, the resulting peak values of these k blocks of equal length can be used for estimation. The procedure is rather wasteful of data and a relatively large sample is needed for accurate estimate.

3.2 Peaks Over Threshold (POT) Approach

The POT (Peaks-Over-Threshold) approach consists of using the generalized Pareto distribution (GPD) to approximate the distribution of excesses over a threshold. This approach has been suggested originally by hydrologists. This approach is generally preferred and forms the basis of our approach below. Both EVT approaches are equivalent by the Pickands-Balkema-de Haan theorem presented in [5,40].

Theorem 2 (Pickands-Balkema-de Haan). *For a large class of underlying distribution functions F,*

$$F \in DA(H_\gamma) \qquad \Longleftrightarrow \qquad \sup_{x \in [0,\, s_+(F)[} |F_u(x) - G_{\gamma,\,\sigma(u)}(x)| \overset{u \to s_+(F)}{\longrightarrow} 0, \tag{8}$$

where $s_+(F) = \sup\{x, \, F(x) < 1\}$ is the end point of the distribution, $F_u(x) = \mathbf{P}(X - u \leq x | X > u)$ is the distribution of excess, and $G_{\gamma,\,\sigma}$ is the Generalized Pareto Distribution (GPD) defined as

$$G_{\gamma,\,\sigma}(x) = \begin{cases} 1 - (1 + \frac{\gamma}{\sigma}x)^{-1/\gamma} & \textit{if } \gamma \neq 0, \ \sigma > 0, \\ 1 - \exp(-x/\sigma) & \textit{if } \gamma = 0, \ \sigma > 0. \end{cases} \tag{9}$$

This means that the conditional excess distribution function F_u, for u large, is well approximated by a Generalized Pareto Distribution. Note that the tail index γ is the same for both the GPD and GEV distributions. The tail shape parameter σ and the tail index are the fundamental parameters governing the extreme behavior of the distribution, and the effectiveness of EVT in forecasting depends upon their reliable and accurate estimation. By incorporating information about the tail through our estimates of γ and σ, we can obtain VaR and ES estimates, even beyond the reach of the empirical distribution.

4 Parametric and Semi-parametric Estimation Methods

The problem of estimating the tail index γ has been widely studied in the literature. The most standard methods are of course the method of moments and maximum likelihood. Unfortunately, there is no explicit form for the parameters, but numerical methods provide good estimates. More generally, the two common approaches to estimate the tail index are:

- Semi-parametric models (e.g., the Hill estimator).
- Fully parametric models (e.g., the Generalized Pareto distribution or GPD).

4.1 Semi-parametric Estimation

The most known estimator for the tail index $\gamma > 0$ of fat tails distribution is without contest the Hill estimator [31]. The formal definition of fat tail distributions comes from regular variation. The cumulative distribution is in the Fréchet domain if and only if as $x \to \infty$, the tails are asymptotically Pareto-distributed:

$$F(x) \approx 1 - Ax^{-\tau}, \tag{10}$$

where $A > 0$ and $\tau = 1/\gamma$. Based on this approximation, the Hill estimator is written according to the statistics order $X_{1,n} \leq ... \leq X_{n,n}$ as follows:

$$H_{k_n,n} = 1/k_n \sum_{j=1}^{k_n} \ln X_{n-j+1,n} - \ln X_{n-k_n,n}. \tag{11}$$

where k_n is a sequence so that $1 \leq k_n \leq n$. Other estimators of this index have have been proposed by Beirlant et al. [6,7] using a regression exponential model to reduce the Hill estimator bias and by [28] that introduce a least squares estimator. The use of a kernel in the Hill estimator has been studied by Csörgő et al. [18]. An effective estimator of the extreme value index has been proposed by Falk and Marohn in [27]. A more detailed list of the different works on the estimation of the index of extreme values is found in [19]. Note that the Hill estimator is sensitive to the choice of threshold $u = X_{n-k_n,n}$ (or the number of excess k_n) and is only valid for fat-tailed data.

4.2 Parametric Estimation

The principle of POT is to approximate the survival function of the excess distribution by a GPD after estimating its parameters from the distribution of excess over a threshold u as explained in the following two steps:

- **First step—Tail distribution estimation**
  Let $X_1, ..., X_n$ follow a distribution F and let $Y_1, \ldots, Y_{N_n}, (Y_i = X_i - u_n)$ be the exceedances over a chosen threshold u_n. The distribution of excess F_{u_n} is given by:

  $$F_{u_n}(y) = P(X - u_n \leq y \,|\, X > u_n) \tag{12}$$

  and then, the distribution F, of the extreme observations, is given by:

  $$F(u_n + y) = F(u_n) + \bar{F}_{u_n}(y) \times \bar{F}(u_n) \tag{13}$$

  The distribution of excess F_{u_n} is approximated by $G_{\gamma,\sigma(u_n)}$ and the first step consists in estimating the parameters of this last distribution using the sample $(Y_1, \ldots, Y_{N_n})$. The parameter estimations can be done using MLE. Different methods have been proposed to estimate the parameters of the GPD. Other estimation methods are presented in [26]. The Probability Weighted Moments (PWM) method proposed by Hosking and Wallis [32] for $\gamma < 1/2$ was extended by Diebolt et al. [21] by a generalization of PWM estimators for $\gamma < 3/2$, as for many applications, e.g., in insurance, distributions are known to have a tail index larger than 1.
- **Second step—Quantile estimation**

  In order to estimate the extreme quantile x_p defined as

  $$x_{p_n} : \bar{F}(x_{p_n}) = 1 - F(x_{p_n}) = p_n, \quad np_n \to 0. \tag{14}$$

  We estimate $F(u)$ by its empirical counterpart N_u/n and we approximate F_{u_n} by the approximate Generalized Pareto Distribution $GPD(\hat{\gamma}_n, \hat{\sigma}_n)$ in the Eq. (1). Then, for the threshold $u = X_{n-k,n}$, the extreme quantile is estimated by

  $$\hat{x}_{p_n,k} = X_{n-k,n} + \hat{\sigma}_n \frac{\left(\frac{k}{np_n}\right)^{\hat{\gamma}_n} - 1}{\hat{\gamma}_n}. \tag{15}$$

The application of POT involves a number of challenges. The early stage of data analysis is very important in determining whether the data has the fat tail needed to apply the EVT results. Also, the parameter estimates of the limit GPD distributions depend on the number of extreme observations used. The choice of a threshold should be large enough to satisfy the conditions to permit its application (u tends towards infinity), while at the same time leaving sufficient observations for the estimation. A high threshold would generate few excesses, thereby inflating the variance of our parameter estimates. Lowering the threshold would necessitate using samples that are no longer considered as being in the tails which would entail an increase in the bias.

5 Non-parametric Estimation Methods

A main argument for using non-parametric estimation methods is that no specific assumptions on the distribution of the data is made *a priori*. That is, model specification bias can be avoided. This is relevant when there is limited information about the 'theoretical' data distribution, when the data can potentially contain a mix of variables with different underlying distributions, or when no suitable parametric model is available. In the context of extreme value distributions, the GPD and GEV distributions discussed in Sect. 3 are appropriate parametric models for the univariate case. However, for the multivariate case there is no general parametric form.

We restrict the discussion here to one particular form of non-parametric estimation, kernel density estimation [44]. Classical kernel estimation performs well when the data is symmetric, but has problems when there is significant skewness [9,24,41].

A common way to deal with skewness is transformation kernel estimation [45], which we will discuss with some details below. The idea is to transform the skew data set into another variable that has a more symmetric distribution, and allows for efficient classical kernel estimation.

Another issue for kernel density estimation is boundary bias. This arises because standard kernel estimates do not take knowledge of the domain of the data into account, and therefore the estimate does not reflect the actual behaviour close to the boundaries of the domain. We will also review a few bias correction techniques [34].

Even though kernel estimation is non-parametric with respect to the underlying distribution, there is a parameter that needs to be decided. This is the bandwidth (scale) of the kernel function, which determines the smoothness of the density estimate. We consider techniques intended for constant bandwidth [35], and also take a brief look at variable bandwidth kernel estimation [36]. In the latter case, the bandwidth and the location is allowed to vary such that bias can be reduced compared with using fixed parameters.

Kernel density estimation can be applied to any type of application and data, but some examples where it is used for extreme value distributions are given in [8,9]. A non parametric method to estimate the VaR in extreme quantiles, based on transformed kernel estimation (TKE) of the cdf of losses was proposed in [3]. A kernel estimator of conditional ES is proposed in [13,14,43].

In the following subsections, we start by defining the classical kernel estimator, then we describe a selection of measures that are used for evaluating the quality of an estimate, and are needed, e.g, in the algorithms for bandwidth selection. Finally, we go into the different subareas of kernel estimation mentioned above in more detail.

5.1 Classical Kernel Estimation

Expressed in words, a classical kernel estimator approximates the probability density function associated with a data set through a sum of identical, symmetric kernel density functions that are centered at each data point. Then the sum is normalized to have total probability mass one.

We formalize this in the following way: Let $k(\cdot)$ be a bounded and symmetric probability distribution function (pdf), such as the normal distribution pdf or the Epanechnikov pdf, which we refer to as the kernel function.

Given a sample of n independent and identically distributed observations $X_1, \ldots, X_n$ of a random variable X with pdf $f_X(x)$, the classical kernel estimator is given by

$$\hat{f}_X(x) = \frac{1}{n} \sum_{i=1}^{n} k_b(x - X_i), \tag{16}$$

where $k_b(\cdot) = \frac{1}{b} k(\frac{\cdot}{b})$ and b is the bandwidth. Similarly, the classical kernel estimator for the cumulative distribution function (cdf) is given by

$$\hat{F}_X(x) = \frac{1}{n} \sum_{i=1}^{n} K_b(x - X_i), \tag{17}$$

where $K_b(x) = \int_{-\infty}^{x} k_b(t)dt$. That is, $K(\cdot)$ is the cdf corresponding to the pdf $k(\cdot)$.

5.2 Selected Measures to Evaluate Kernel Estimates

A measure that we would like to minimize for the kernel estimate is the mean integrated square error (MISE). This is expressed as

$$\text{MISE}(b) = \mathbb{E}\left[\int_{\Omega} \left(\hat{f}_X(x) - f_X(x)\right)^2 dx\right], \tag{18}$$

where Ω is the domain of support for $f_X(x)$, and the argument b is included to show that minimizing MISE is one criterion for bandwidth selection. However, MISE can only be computed when the true density $f_X(x)$ is known. MISE can be decomposed into two terms. The integrated square bias

$$\int_{\Omega} \left(\mathbb{E}\left[\hat{f}_X(x)\right] - f_X(x)\right)^2 dx, \tag{19}$$

and the integrated variance

$$\int_{\Omega} \text{Var}\left[\hat{f}_X(x)\right] dx. \tag{20}$$

To understand these expressions, we first need to understand that $\hat{f}_X$ is a random variable that changes with each sample realization. To illustrate what it means, we work through an example.

Example 1. Let X be uniformly distributed on $\Omega = [0, 1]$, and let $k(\cdot)$ be the Gaussian kernel. Then

$$f_X(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

For each kernel function centered at some data point y we have that

$$\mathbb{E}_X\left[k_b(x-y)\right] = \int_0^1 k_b(x-y)\cdot 1dy = K_b(x) - K_b(x-1). \tag{22}$$

If we apply that to the kernel estimator (16), we get the integrated square bias

$$\int_0^1 \left(K_b(x) - K_b(x-1) - 1\right)^2 dx. \tag{23}$$

By first using that $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, we get the following expression for the integrated variance

$$\int_0^1 \left(\int_0^1 k_b(x-y)^2 \cdot 1dy - \left(K_b(x) - K_b(x-1)\right)^2\right) dx. \tag{24}$$

The integrals are evaluated for the Gaussian kernel, and the results are shown in Fig. 2. The bias, which is largest at the boundaries, is minimized when the bandwidth is very large, but a large bandwidth also leads to a large variance. Hence, MISE is minimized by a bandwidth that provides a trade-off between bias and variance.

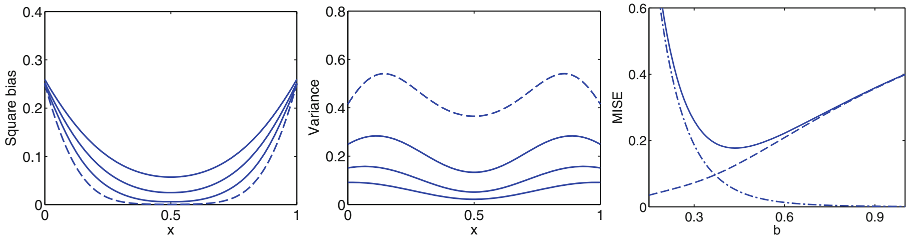

Fig. 2. The square bias (left) and the variance (middle) as a function of x for $b = 0.3$, 0.4, 0.5, 0.6. The curve for $b = 0.3$ is shown with a dashed line in both cases. MISE (right) is shown as a function of b (solid line) together with the integrated square bias (dashed line) and the integrated variance (dash-dot line).

To simplify the analysis, MISE is often replaced with the asymptotic MISE approximation (AMISE). This holds under certain conditions involving the sample size and the bandwidth. The bandwidth depends on the sample size, and we can write $b = b(n)$. We require $b(n) \downarrow 0$ as $n \longrightarrow \infty$, while $nb(n) \longrightarrow \infty$ as $n \longrightarrow \infty$. Furthermore, we need $f_X(x)$ to be twice continuously differentiable. We then have [44] the asymptotic approximation

$$\mathrm{AMISE}(b) = \frac{1}{nb}R(k) + \frac{1}{4}m_2(k)^2 b^4 R(f''), \tag{25}$$

where $R(g) = \int g(x)^2\,dx$ and $m_p(k) = \int x^p k(x)\,dx$. The bandwidth that minimizes AMISE can be analytically derived to be

$$b_{\text{opt}} = \left(\frac{R(k)}{m_2(k)^2 R(f'')} \frac{1}{n} \right)^{\frac{1}{5}}, \tag{26}$$

leading to

$$\text{AMISE}(b_{\text{opt}}) = \frac{5}{4} \left(\frac{m_2(k)^2 R(k)^4 R(f'')}{n^4} \right)^{\frac{1}{5}}. \tag{27}$$

The optimal bandwidth can then be calculate for different kernel functions. We have, e.g., for the Gaussian kernel [11]

$$b^G_{\text{opt}} = \left(\frac{1}{2\sqrt{\pi} R(f'')} \frac{1}{n} \right)^{\frac{1}{5}}. \tag{28}$$

The difficulty in using AMISE is that the norm of second derivative of the unknown density needs to be estimated. This will be further discussed under the subsection on bandwidth selectors.

We also mention the skewness γ_X of the data, which is a measure that can be used to see if the data is suitable for classical kernel estimation. We estimate it as

$$\hat{\gamma}_X = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^3}{\left(\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2\right)^{\frac{3}{2}}}. \tag{29}$$

It was shown in [44], see also [41], that minimizing the square integrated error (ISE) for a specific sample is equivalent to minimizing the cross-validation function

$$\text{CV}(b) = \int_{\Omega} \left(\hat{f}_X(x)\right)^2 dx - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_i(x_i), \tag{30}$$

where $\hat{f}_i(\cdot)$ is the kernel estimator obtained when leaving the observation x_i out. Other useful measures of the goodness of fit are also discussed in [41].

5.3 Bias-Corrected Kernel Estimation

As was illustrated in Example 1, boundaries where the density does not go to zero generate bias. This happens because the kernel functions cross the boundary, and some of the mass ends up outside the domain. We want from the kernel method that $\mathbb{E}[\hat{f}_X(x)] = f_X(x)$ in all of the support Ω of the density function, but this condition does not hold at boundaries, unless we also have that the density is zero there. An overview of the topic, and of simple boundary correction methods, is given in [34]. By employing a linear bias correction method, we can make the moments of order 0 and 1 satisfy the consistency requirements $m_0 = 1$ (total probability mass) and $m_1 = 0$, such that the expectation is consistent to order

b^2 at the boundary. A general linear correction method for a density supported on $\Omega = [x_0, \infty]$ that is shown to perform well in [34] has the form

$$\tilde{k}_b(y) = \frac{(a_2 - a_1 y)k_b(y)}{a_0 a_2 - a_1^2}, \tag{31}$$

for a kernel function centered at the data location x. The coefficients $a_j = a_j(b, x)$ are computed as

$$a_j(b, x) = \int_{x_0}^{z} y^j k_b(x - y)dy, \tag{32}$$

where z is the end point of the support of the kernel function. An example with modified Gaussian kernels close to a boundary is shown in Fig. 3. At the boundary, the amplitude of the kernels becomes higher to compensate for the mass loss, while away from the boundary they resume the normal shape and size. The kernel functions closest to the boundary become negative in a small region, but this does not affect the consistency of the estimate.

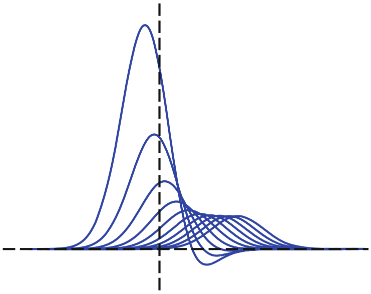

Fig. 3. Bias corrected kernel functions using the linear correction approach near a boundary (dashed line).

A more recent bias correction method is derived in [38], based on ideas from [15]. This type of correction is applied to the kernel estimator for the cdf, and can be seen as a Taylor expansion. It also improves the capturing of valleys and peaks in the distribution function, compared with the classical kernel estimator. It requires that the density is four times differentiable, that the kernel is symmetric, and at least for the theoretical derivations, that the kernel is compactly supported on $\Omega = [-1, 1]$. The overall bias of the estimator is $\mathcal{O}(b^4)$ as compared with $\mathcal{O}(b^2)$ for the linear correction method, while the variance is similar to what is achieved with the uncorrected estimator. This boundary correction approach is used for estimating extreme value distributions in [9].

$$\tilde{F}_X(x) = \frac{\hat{F}_X(x) + \lambda(\hat{F}_1(x) + \hat{F}_2(x))}{1 + 2\lambda}, \tag{33}$$

where $\hat{F}_{1,2}(x) = \hat{F}_X(x \mp \ell b) \pm \ell b \hat{f}_X(x \mp \ell b)$, and

$$\ell(\lambda) = \sqrt{\frac{(1+2\lambda)m_2(k)}{2\lambda}}. \tag{34}$$

The parameter λ is kernel dependent, and should be chosen such that AMISE is minimized, but according to [15], the estimator is not that sensitive to the choice. An explicit expression for AMISE with this correction is derived in [38], and is also cited in [9], where the value $\lambda = 0.0799$ is also given as an (approximate) minimizer of the variance for the Epanechnikov kernel.

5.4 Transformation Kernel Estimation

The objective in transformation kernel estimation is to find a transformation of the random variable X, which for example has a right-skewed distribution into a symmetric random variable Y. Then classical kernel estimation can be successfully applied to Y.

The transformation function $T(\cdot)$ should be monotonic and increasing. For a right-skewed true density, it should also be concave. It also needs to have at least one continuous derivative. The transformation is applied to the original data to generate a transformed data sample

$$y_i = T(x_i), \quad i = 1, \ldots, n. \tag{35}$$

For the pdfs of the two random variables it holds that

$$f_X(x) = T'(x) f_Y(y), \tag{36}$$

and for the cdfs we have that $F_X(x) = F_Y(y)$. We apply the kernel density estimator to the transformed data, leading to the following estimator for the original density:

$$\hat{f}_X(x) = T'(x)\hat{f}_Y(y) = T'(x)\frac{1}{n}\sum_{i=1}^{n} k_b(T(x) - T(x_i)). \tag{37}$$

Several different transformation classes have been proposed for heavy tailed data. The shifted power transformation family was proposed in [45]

$$T(x) = \begin{cases} \operatorname{sign}(\lambda_2)(x+\lambda_1)^{\lambda_2}, & \lambda_2 \neq 0, \\ \ln(x+\lambda_1), & \lambda_2 = 0, \end{cases} \tag{38}$$

where $\lambda_1 > -\min(x_i)$ and $\lambda_2 \leq 1$. An algorithm for choosing the transformation parameters is given in [10]. First a restriction is made to parameters $\lambda_{1,2}$ that give close to zero skewness (29) for the transformed data. Then AMISE (27) of the classical kernel estimation for the density $f_Y(y)$ is minimized assuming an asymptotically optimal bandwidth. This is equivalent to minimizing $R(f_Y'')$. As we do not know the true density, an estimator is needed. The estimator suggested in [10] is

$$\hat{R}(f_Y'') = \frac{1}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=1+1}^{n} \frac{1}{c^5} k * k(y_i - y_j), \tag{39}$$

where the convolution $k * k(u) = \int k(u-s)k(s)ds$, and c is the bandwidth used in this estimate.

The Möbius-like mapping introduced in [16] takes data in $\Omega_X = [0, \infty)$ and maps it to $\Omega_Y = [-1,\, 1)$.

$$T(x) = \frac{x^\alpha - M^\alpha}{x^\alpha + M^\alpha}. \tag{40}$$

The scale M is determined by minimizing $R(\hat{f}_Y'')$. Given a scale M, α is determined such that no probability mass spills over at the right boundary. That is, the resulting density does not have mass at (or beyond) infinity.

A modified Champernowne distribution transformation is derived in [12], with transformation function

$$T(x) = \frac{\alpha(x+c)^{\alpha-1}((M+c)^\alpha) - c^\alpha}{((x+c)^\alpha + (M+c)^\alpha - 2c^\alpha)^2}, \tag{41}$$

where M can be chosen as the median of the data, and α and c are found by maximizing a log likelihood function, see [12].

So far, we have only considered the possibility of performing one transformation, but one can also transform the data iteratively, or perform two specific consecutive transformations. Doubly transformed kernel estimation is discussed, e.g., in [9]. The idea is to first transform the data to something close to uniform, and then to apply an inverse beta transformation. This makes the final distribution close to a beta distribution, and the optimal bandwidth can then easily be computed.

5.5 Bandwidth Selection

As briefly mentioned in Sects. 5.2 and 5.4, the choice of bandwidth b in kernel estimation has a significant impact on the quality of the estimator, but choosing the appropriate bandwidth requires the use of one estimator or another. The rule-of-thumb bandwidth estimator of Silverman [44],

$$b = \left(\frac{4}{3n}\right)^{\frac{1}{5}} \hat{\sigma}_X, \tag{42}$$

is often cited, but it assumes that the underlying density can be approximated by a normal density. This is hence not appropriate for heavy-tailed, right-skewed distributions.

Many bandwidth selection methods use a normal reference at some step in the process [11], but this introduces a parametric step in the non-parametric estimation. An interesting alternative, the Improved Sheather-Jones bandwidth selection algorithm, is also described in [11], where the normal reference is eliminated by formulating a non-linear equation for the optimal bandwidth.

We start from the point of how to estimate $R(f''_X) = \|f''_x\|^2$. There are at least two possible plug-in estimators, based on the equality

$$\|f_X^{(j)}\|^2 = (-1)^j \mathbb{E}[f^{(2j)}]. \tag{43}$$

The two types of estimators are

$$\widehat{\|f_X^{(j)}\|^2} = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{m=1}^{n} \int_{\mathbb{R}} k_b^{(j)}(x - x_k) k_b^{(j)}(x - x_m) dx \tag{44}$$

and

$$(-1)^j \widehat{\mathbb{E}[f^{(2j)}]} = \frac{(-1)^j}{n^2} \sum_{k=1}^{n} \sum_{m=1}^{n} k_b^{(2j)}(x_k - x_m). \tag{45}$$

By requiring the two estimators to have the same asymptotic mean square error, we get a condition on the bandwidth. A derivation for the Gaussian kernel is provided in [11], and we summarize these result here to illustrate the bandwidth selection algorithm. Requiring (44) and (45) to be asymptotically equivalent gives the following relation between the bandwidths for two consecutive derivatives:

$$\hat{b}_j^* = \left(\frac{1 + \frac{1}{2^{j+\frac{1}{2}}}}{3} \frac{(2j-1)!!}{N\sqrt{\frac{\pi}{2}} \widehat{\|f^{(j+1)}\|^2}} \right)^{\frac{1}{3+2j}} \equiv \gamma_j(\hat{b}_{j+1}^*), \tag{46}$$

and, for the Gaussian kernel, (44) for $j+1$ becomes

$$\widehat{\|f_X^{(j+1)}\|^2} = \frac{(-1)^j}{n^2} \sum_{k=1}^{n} \sum_{m=1}^{n} k_{\sqrt{2} b_{j+1}^*}^{(2j)}(x_k - x_m). \tag{47}$$

The two relations (46) and (47) together define the function $\gamma_j(b)$. We also have Eq. (28) for the bandwidth of the Gaussian kernel. Combining this with (46) for $j = 1$ allows us to eliminate $\|\widehat{f''}\|^2$ to get

$$b^* = \left(\frac{6\sqrt{2} - 3}{7} \right)^{\frac{1}{5}} b_1^* \equiv \xi b_1^*. \tag{48}$$

Now we can apply (46) recursively to get a relation

$$b^* = \xi \gamma^{[\ell]}(b_{\ell+1}^*), \tag{49}$$

where $\gamma^{[\ell]}(b) = \gamma_1(\cdots \gamma_{\ell-1}(\gamma_\ell(b)) \cdots)$, for $\ell \geq 1$. Here, it would be possible to assume a normal distribution to estimate $\|\widehat{f^{\ell+2}}\|^2$ in order to compute $b_{\ell+1}^*$, and then all the other bandwidths. However, this does not work well if the true distribution is far from normal [11]. In the improved Sheather-Jones algorithm, we instead assume that $b^*_{\ell+1} = b$ for some large enough ℓ. The experiments in [11] indicate that $\ell = 5$ should be enough. We then get a non-linear equation to solve for b

$$b = \xi \gamma^{[\ell]}(b). \tag{50}$$

Using this relation, no assumptions on the true distribution are made, and this bandwidth selector is shown to perform well also for non-normal distributions.

6 More Challenges in Estimating the Risk Measures—Financial Time Series and Multivariate Case

A Dynamic Approach. Highlighting the underlying assumptions is relevant for understanding model uncertainty when estimating rare or extreme events. The VaR and ES are estimated given that the distribution of asset returns does not change over time. In the last two sections, when applying the POT approach to the returns in order to calculate these risk measures, their distribution was assumed to be stationary. A dynamic model which captures current risk is then more realistic. EVT can also be used based on a stochastic time series model. These dynamic models use an ARCH/GARCH type process along with the POT to model VaR and ES which depend on and change due to the fluctuations of the market. This approach, studied in [2], reflects two stylized facts exhibited by most financial return series, namely stochastic volatility and the fat-tailedness of conditional return distributions over short time horizons.

The Multivariate Case for EVT. When estimating the VaR of a multi-asset portfolio, under financial crises, correlations between assets often become more positive and stronger. Assuming that the variables are independent and identically distributed is a strong hypothesis. Portfolio losses are the result not only of the individual asset's performance but also, and very importantly, the result of the interaction between assets. Hence, from the accuracy point of view, ideally we would prefer the multivariate approach.

An extension of the univariate EVT models using a dependence structure leads to a parametric model and is then expected to be less efficient for scarce data. A non-parametric approach should be preferred to estimate portfolio tail risk. Transformation kernel density estimation is used in [8] for studying multivariate extreme value distributions in temperature measurement data. Future directions involve to apply this type of methodology to real and simulated portfolio data.

References

1. Acerbi, C., Tasche, D.: On the coherence of expected shortfall. J. Bank. Finance **26**(7), 1487–1503 (2002)
2. McNeil, A.J., Frey, R.: Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. J. Empir. Finance **7**, 271–300 (2000)
3. Alemany, R., Bolancé, C., Guillén, M.: A nonparametric approach to calculating value-at-risk. Insur. Math. Econ. **52**(2), 255–262 (2013)
4. Artzner, P., Delbaen, F., Eber, J.M., Heath, D.: Coherent measures of risk. Math. Finance **9**(3), 203–228 (1999)
5. Balkema, A.A., de Haan, L.: Residual life time at great age. Ann. Probab. **2**(5), 792–804 (1974)
6. Beirlant, J., Dierckx, G., Goegebeur, Y., Matthys, G.: Tail index estimation and an exponential regression model. Extremes **2**(2), 177–200 (1999)
7. Beirlant, J., Dierckx, G., Guillou, A., Staăricaă, C.: On exponential representations of log-spacings of extreme order statistics. Extremes **5**(2), 157–180 (2002)
8. Beranger, B., Duong, T., Perkins-Kirkpatrick, S.E., Sisson, S.A.: Exploratory data analysis for moderate extreme values using non-parametric Kernel methods. arXiv:1602.08807 [stat.ME] (2016)
9. Bolancé, C., Bahraoui, Z., Alemany, R.: Estimating extreme value cumulative distribution functions using bias-corrected Kernel approaches. XREAP2015-01 (2015)
10. Bolancé, C., Guillén, M., Perch Nielsen, J.: Kernel density estimation of actuarial loss functions. Insur. Math. Econ. **32**(1), 19–36 (2003)
11. Botev, Z.I., Grotowski, J.F., Kroese, D.P.: Kernel density estimation via diffusion. Ann. Stat. **38**(5), 2916–2957 (2010)
12. Buch-larsen, T., Nielsen, J.P., Guillén, M., Bolancé, C.: Kernel density estimation for heavy-tailed distributions using the Champernowne transformation. Statistics **39**(6), 503–516 (2005)
13. Cai, Z., Wang, X.: Nonparametric estimation of conditional VaR and expected shortfall. J. Econ. **147**(1), 120–130 (2008)
14. Chen, S.: Non-parametric estimation of expected shortfall. J. Financ. Econ. **6**, 87–107 (2008)
15. Choi, E., Hall, P.: On bias reduction in local linear smoothing. Biometrika **85**(2), 333–345 (1998)
16. Clements, A., Hurn, S., Lindsay, K.: Mobius-like mappings and their use in Kernel density estimation. J. Am. Stat. Assoc. **98**(464), 993–1000 (2003)
17. Coles, S.: An Introduction to Statistical Modeling of Extreme Values. Springer Series in Statistics. Springer, London (2001). https://doi.org/10.1007/978-1-4471-3675-0
18. Csörgő, S., Deheuvels, P., Mason, D.: Kernel estimates of the tail index of a distribution. Ann. Stat. **13**(3), 1050–1077 (1985)
19. Csörgő, S., Viharos, L.: Estimating the tail index. In: Szyszkowicz, B. (ed.) Asymptotic Methods in Probability and Statistics, pp. 833–881. North-Holland, Amsterdam (1998)
20. Danielsson, J.: Financial Risk Forecasting. Wiley, Hoboken (2011)
21. Diebolt, J., Guillou, A., Rached, I.: Approximation of the distribution of excesses through a generalized probability-weighted moments method. J. Stat. Plan. Infer. **137**(3), 841–857 (2007)
22. Dowd, K.: Measuring Market Risk. Wiley, Hoboken (2005)

23. Duffie, D., Pan, J.: An overview of value at risk. J. Deriv. **4**(3), 7–49 (1997)
24. Eling, M.: Fitting insurance claims to skewed distributions: are the skew-normal and skew-student good models? Insur. Math. Econ. **51**(2), 239–248 (2012)
25. Embrechts, P. (ed.): Extremes and Integrated Risk Management. Risk Books, London (2000)
26. Embrechts, P., Klüppelberg, C., Mikosch, T.: Modelling Extremal Events: for Insurance and Finance. Springer, Heidelberg (1997). https://doi.org/10.1007/978-3-642-33483-2
27. Falk, M., Marohn, F.: Efficient estimation of the shape parameter in Pareto models with partially known scale. Stat. Decis. **15**, 229–239 (1997)
28. Feuerverger, A., Hall, P.: Estimating a tail exponent by modelling departure from a Pareto distribution. Ann. Stat. **27**(2), 760–781 (1999)
29. Fisher, R.A., Tippett, L.H.C.: Limiting forms of the frequency distribution of the largest or smallest member of a sample. Math. Proc. Camb. Philos. Soc. **24**(2), 180–190 (1928)
30. Gnedenko, B.: Sur la distribution limite du terme maximum d'une série aléatoire. Ann. Math. **44**(3), 423–453 (1943)
31. Hill, B.M.: A simple general approach to inference about the tail of a distribution. Ann. Stat. **3**, 1163–1174 (1975)
32. Hosking, J.R., Wallis, J.R.: Parameter and quantile estimation for the generalized Pareto distribution. Technometrics **29**(3), 339–349 (1987)
33. Hull, J.C.: Risk Management and Financial Institutions. Prentice Hall, Upper Saddle River (2006)
34. Jones, M.C.: Simple boundary correction for Kernel density estimation. Stat. Comput. **3**(3), 135–146 (1993)
35. Jones, M.C., Marron, J.S., Sheather, S.J.: A brief survey of bandwidth selection for density estimation. J. Am. Stat. Assoc. **91**(433), 401–407 (1996)
36. Jones, M.C., McKay, I.J., Hu, T.C.: Variable location and scale Kernel density estimation. Ann. Inst. Stat. Math. **46**(3), 521–535 (1994)
37. Jorion, P.: Value at Risk: The New Benchmark for Managing Financial Risk. McGraw-Hill, New York (2001)
38. Kim, C., Kim, S., Park, M., Lee, H.: A bias reducing technique in Kernel distribution function estimation. Comput. Stat. **21**(3–4), 589–601 (2006)
39. McNeil, A.J., Frey, R., Embrechts, P.: Quantitative Risk Management: Concepts, Techniques and Tools. Princeton Series in Finance. Princeton University Press, Princeton (2005)
40. Pickands, J.: Statistical inference using extreme order statistics. Ann. Stat. **3**(1), 119–131 (1975)
41. Pitt, D., Guillen, M., Bolancé, C.: Estimation of parametric and nonparametric models for univariate claim severity distributions - an approach using R. XREAP2011-06 (2011)
42. Reiss, R.D., Thomas, M.: Statistical Analysis of Extreme Values: with Applications to Insurance, Finance, Hydrology and Other Fields. Birkhäuser, Basel (2007)
43. Scaillet, O.: Nonparametric estimation of conditional expected shortfall. Insur. Risk Manag. J. **74**(1), 639–660 (2005)
44. Silverman, B.W.: Density Estimation for Statistics and Data Analysis, Monographs on Statistics and Applied Probability, vol. 26. Chapman & Hall/CRC, London (1986)
45. Wand, M.P., Marron, J.S., Ruppert, D.: Transformations in density estimation. J. Am. Stat. Assoc. **86**(414), 343–353 (1991)