The risk function of the goodness-of-fit tests for tail models

Ingo Hoffmann · Christoph J. Börner

Received: 10 August 2018 / Revised: 2 August 2019 / Published online: 23 January 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract
This paper contributes to answering a question that is of crucial importance in risk management and extreme value theory: How to select the threshold above which one assumes that the tail of a distribution follows a generalized Pareto distribution. This question has gained increasing attention, particularly in finance institutions, as the recent regulative norms require the assessment of risk at high quantiles. Recent methods answer this question by multiple uses of the standard goodness-of-fit tests. These tests are based on a particular choice of symmetric weighting of the mean square error between the empirical and the fitted tail distributions. Assuming an asymmetric weighting, which rates high quantiles more than small ones, we propose new goodness-of-fit tests and automated threshold selection procedures. We consider a parameterized family of asymmetric weight functions and calculate the corresponding mean square error as a loss function. Then we explicitly determine the risk function as the expected value of the loss function for finite sample. Finally, the risk function can be used to discuss whether a symmetric or asymmetric weight function should be chosen. With this the goodness-of-fit test which should be used in a new method for determining the threshold value is specified.

Keywords Decision theory · Risk function · Goodness-of-fit tests · Tail model

Mathematics Subject Classification 62C99 · 62E17

1 Introduction
In many disciplines, there is often a need to adapt a statistical model to the existing data to make statements about uncertain future outcomes. In particular, when assessing
risks, an estimate of the major losses must be based on events that, although they have a low probability of occurrence, have a high impact. Especially in the financial sector, with its tightening regulatory requirements, models will be in demand that allow very good, qualitative and quantitative statements at high quantiles in the tail range of an underlying unknown distribution function. There already is a high and growing interest in significant statements at high quantiles in various application areas; see, e.g., Wagner and Marsh (2004), Kotz and Seier (2009) and the references cited therein.

Since the actual distribution of the data is often unknown, statisticians begin with a guess about the underlying statistical model for the entire value range or, more specifically, for the considered tail. They often use various distribution functions to choose the most suitable one later. In many cases, these models do not perfectly reflect the data. However, specific statistical tests can be applied to assess how good or bad a model fits the data, e.g., the Cramér-von Mises test (Cramér 1928; Von Mises 1931) or Anderson-Darling test (Anderson and Darling 1952, 1954). These goodness-of-fit tests are also used in automated procedures to determine the threshold value at which the tail of the underlying distribution can be modeled using the GPD (Bader et al. 2018).

All of these tests are based on the weighted mean square error \( \hat{R} \), which corresponds to the loss function in decision theory (Aggarwal 1955; Ferguson 1967). Evaluated for a specific sample of length \( n \), \( \hat{R}_n \) calculates the weighted deviation of the modeled data from the measured data and, thus, the individual loss of accuracy by the model. At this level, the derived statistical tests above are used to assess the quality of the model. However, to be able to judge how good a statistical test is, the question is, how large is the average loss when all possible time series of measured data with length \( n \) and unknown distribution functions are considered. An answer to this question is provided by the finite sample expectation value of the weighted mean square error \( E[\hat{R}_n] \), which corresponds to the risk function in decision theory.

If the error is not squared but has an initially free exponent, Aggarwal (1955) was able to explicitly calculate the risk function for this deviation error. Different weight functions were considered for only two specific cases: the Cramér-von Mises and Anderson-Darling tests. In particular, when evaluating models for the upper or lower tail, weight functions are important, which enable only a stronger weighting of deviations in these areas of the distribution. These weighting functions define the families of special tail statistics on which we focus here. The question remains; of which statistics of this parameterized family should be used for the given task to establish a suitable goodness-of-fit test in an automated method for determining the threshold value.

As the main result of this analysis, we calculate the risk function for that family of tail statistics, which allows us to compare different statistics in terms of their average loss. Thus, the question of a suitable statistic for a tail-oriented goodness-of-fit test can be discussed. Our result shows that some statistics diverge and cannot be used. The results further suggest that from theoretical and practical viewpoints, the statistics first suggested by Ahmad et al. (1988) should be chosen as a goodness-of-fit test for analyzing the tail and evaluating a tail model. This statistic should be further investigated and used as the origin of an automated method for determining the threshold to
separate the tail from the distribution. In a small example, we show a way how this can be implemented.

Note that peculiarities of censored data or contaminations of these due to superimposed distribution functions were not taken into account, when these straightforward tests were derived. For such cases, there are recently developed new methods based on depth (Denecke and Müller 2012, 2014). Instead, in our further considerations, we exclude censoring and contamination of the data and assume that the data set consists of independently, identically distributed realizations.

The remainder of the paper is structured as follows: After defining the family of tail statistics in Sect. 2, the corresponding risk function is explicitly calculated in Sect. 3. Section 4 summarizes some corollaries that follow from the theorem of the previous section. As an interesting side result, we define a one-parameter discrete distribution function over a finite support of non-negative integers and determine all moments of this distribution. This may be useful in a decision-theoretic problem where probabilities are to be assigned to a limited number of environmental states. Section 5 shows an example of how an automated modeling method of the tail can be implemented. The final section discusses the results and summarizes the key points.

2 Definition of tail statistics

Let \( X_1, X_2, \ldots, X_n \) be a sample of random variables with a common unknown continuous distribution function \( F(x) \) and density function \( f(x) \). The corresponding empirical distribution function for \( n \) observations is defined as

\[
F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \leq x),
\]

where \( 1 \) is the indicator function; \( 1(X_i \leq x) \) is equal to one if \( X_i \leq x \) and zero otherwise. Thus, \( F_n(x) = \frac{k}{n} \) if \( k \) observations are less than or equal to \( x \) for \( k = 0, 1, \ldots, n \) (Kolmogorov 1933).

As a convenient measure of the discrepancy or ”distance” between the distribution functions \( F_n(x) \) and \( F(x) \), we consider the weighted mean square error

\[
\hat{R}_n = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 w(F(x)) \, dF(x),
\]

introduced in the context of statistical test procedures by Cramér (1928), Von Mises (1931) and Smirnov (1936). The non-negative weight function \( w(t) \) in Eq. (2) is a suitable preassigned function for accentuating the difference between the distribution functions in the range where the test procedure is desired to be sensitive. Consider the weight function

\[
w(t) = \frac{1}{t^a (1-t)^b}
\]
for free real-valued stress parameters \(a, b \geq 0\) and \(t\) as indeterminate for the later integration variable. Here, \(a\) affects the weight at the lower tail, and \(b\) affects the weight at the upper tail. These stress parameters, at a certain position of the distribution function, allow one to change the strength with which the deviations from the empirical distribution function at that position are weighted. Put simply, by using the stress parameters, the magnification is adjusted, with which the deviation between the distributions at a fixed position is considered.

Then, for \(a = b = 0\), Eq. (2) provides the Cramér-von Mises statistic (Cramér 1928; Von Mises 1931), and when both tails are heavily weighted \((a = b = 1)\), it is equal to the Anderson-Darling statistic (Anderson and Darling 1952, 1954). The Anderson-Darling statistic simultaneously weights the difference between the distributions more heavily at both ends of the distribution \(F(x)\).

Mixed weight functions can hinder the individual study of one tail or the other of the distribution function. In particular, in the construction of goodness-of-fit tests that focus on a tail, pure functions, which weight one side of the distribution function strongly, are beneficial. As the regulatory requirements become more stringent, statistics which weight the differences in either the upper or lower tail of the distribution function more strongly may become increasingly interesting. Therefore, the following weight functions should gain importance.

The weight function for the lower tail \((a \geq 0, b = 0)\) is

\[
w(t) = \frac{1}{t^a}. \tag{4}\]

The weight function for the upper tail \((a = 0, b \geq 0)\) is

\[
w(t) = \frac{1}{(1-t)^b}. \tag{5}\]

If we initially leave the stress parameters indeterminate in the calculation of the weighted mean square error Eq. (2), two families of statistics can be derived: one family is for the lower tail, and the other is for the upper tail. For \(a = b\), these two families can be transformed into each other using coordinate transformation \(Z = -X\) of the random variable. Therefore, in the following, we only treat the statistics family for stress parameter \(a\). The derived results then apply to the second family for parameter \(b\).

With Eq. (4), the weighted mean square error Eq. (2) reduces to

\[
\hat{R}_{n,a} = n \int_{-\infty}^{+\infty} \frac{(F_n(x) - F(x))^2}{(F(x))^a} \, dF(x). \tag{6}\]

The computing formulae for this family of lower-tail statistics can be obtained by following the method given in Anderson and Darling (1954).

Let \(x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}\) be the sample values (in ascending order) obtained by ordering each realization \(x_1, x_2, \ldots, x_n\) of \(X_1, X_2, \ldots, X_n\). Then, we can summarize the following calculation rules for the statistics:

\[\text{Springer}\]
The risk function of the goodness-of-fit tests for tail models

- $a < 3$ and $a \neq 1, 2$

$$\hat{R}_{n,a} = \frac{2}{(1-a)(2-a)(3-a)} n + \sum_{i=1}^{n} \left[ \frac{2}{2-a} \left( F(x(i)) \right)^{2-a} - \frac{2i-1}{n} \frac{1}{1-a} \left( F(x(i)) \right)^{1-a} \right]$$

(7)

Note: In the special case where $a = 0$, Eq. (7) reduces to the statistics $W_n^2 (= \hat{R}_{n,0})$ proposed by Cramér (1928) and Von Mises (1931):

$$W_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left[ \frac{2i-1}{2n} - F(x(i)) \right]^2$$

(8)

- $a = 1$

$$\hat{R}_{n,1} = -\frac{3}{2n} + \sum_{i=1}^{n} \left[ 2F(x(i)) - \frac{2i-1}{n} \ln \left( F(x(i)) \right) \right]$$

(9)

To obtain an appropriate goodness-of-fit test specifically for the tail of a distribution, the computation formulae Eq. (9) were first described by Ahmad et al. (1988) and later examined more formally by the same authors with regard to the distribution of their test statistics $AL_n^2$ (Sinclair et al. 1990).

- $a = 2$

$$\hat{R}_{n,2} = \sum_{i=1}^{n} \left[ \frac{2i-1}{n} \frac{1}{F(x(i))} + 2 \ln \left( F(x(i)) \right) \right]$$

(10)

- $a \geq 3$

For the stress parameter $a \geq 3$, no feasible solution can be calculated because $\hat{R}_{n,a}$ approaches infinity.

### 3 Risk function

In decision theory, the weighted mean square error $\hat{R}_n$, which is defined in the previous section (see Eq. (2)), is generally referred to as the loss function, and the expected value of the loss function is called the risk function (Aggarwal 1955; Ferguson 1967):

$$R_n = \mathbb{E} \left[ \hat{R}_n \right].$$

(11)

For the case considered here, the risk function $R_{n,a} = \mathbb{E} \left[ \hat{R}_{n,a} \right]$ can be calculated explicitly depending on parameter $a$. Our main result summarizes the following theorem and the complementary corollaries in Sect. 4.
Theorem 1 Let $\hat{R}_{n,a}$ be the weighted mean square error defined by Eq. (6). Then, for $0 \leq a < 2$, the risk function is given by

$$R_{n,a} = \frac{1}{(2-a)(3-a)}.$$ (12)

Proof Using the transformation $u = F(x)$, the lower tail statistics can be expressed in terms of $u \in [0, 1]$, and $u(1) \leq u(2) \leq \cdots \leq u(n)$ is an ordered sample of size $n$ from a continuous uniform distribution over the interval $[0, 1]$. The expectation in Eq. (11) must be taken with respect to this distribution. Since the distribution of the $i$th-order statistic $U(i)$ in a random sample of size $n$ from the uniform distribution over the interval $[0, 1]$ is a beta distribution with the following probability density

$$p(u) = \frac{1}{B(i, n-i+1)} u^{i-1} (1-u)^{n-i}$$ (13)

the expectation value for $\hat{R}_{n,a}$ can be calculated as follows:

- $a < 2$ and $a \neq 1$

$$R_{n,a} = E\left[\hat{R}_{n,a}\right]$$

$$= \frac{2n}{(1-a)(2-a)(3-a)}$$

$$+ \sum_{i=1}^{n} \frac{2}{2-a} E\left[u_{(i)}^{2-a}\right]$$

$$- \sum_{i=1}^{n} \frac{2i-1}{n} \frac{1}{1-a} E\left[u_{(i)}^{1-a}\right]$$

$$= \frac{2n}{(1-a)(2-a)(3-a)}$$

$$+ \sum_{i=1}^{n} \frac{2}{2-a} \int_{0}^{1} u_{(i)}^{1+a} (1-u)^{n-i} \, du$$

$$- \sum_{i=1}^{n} \frac{2i-1}{n} \frac{1}{1-a} \int_{0}^{1} u_{(i)}^{1-a} (1-u)^{n-i} \, du$$

$$= \frac{2n}{(1-a)(2-a)(3-a)}$$

$$+ \sum_{i=1}^{n} \frac{2}{2-a} \frac{B(i+2-a, n-i+1)}{B(i, n-i+1)}$$

$$- \sum_{i=1}^{n} \frac{2i-1}{n} \frac{1}{1-a} \frac{B(i+1-a, n-i+1)}{B(i, n-i+1)}$$ (14)
Note that for \( a \geq 2 \) no feasible solution can be calculated, as \( \text{E}[u_{(i)}^{1-a}] \) becomes infinite when evaluated for \( u \in [0, 1] \). Therefore, we have to limit the parameter further to \( a < 2 \).

To evaluate the remaining sums in Eq. (14), let

\[
H_k(v) = \sum_{i=1}^{n} i^k \frac{B(i + v, n - i + 1)}{B(i, n - i + 1)}. \tag{15}
\]

Due to Lemma 1, see Appendix A Eq. (32), this is equal to

\[
H_k(v) = \frac{n}{v + 1} \sum_{l=0}^{k} S_{k+1,l+1} \frac{v + 1}{v + 1 + l} (n - 1)_{(l)} \tag{16}
\]

with \( v > -1, k \in \mathbb{N}, (n - 1)_{(l)} \) is the Pochhammer notation for falling factorials and \( S_{k,l} \) are the Stirling numbers of the second kind (Abramowitz and Stegun 2014).

Now, Eq. (14) becomes

\[
R_{n,a} = \frac{2n}{(1-a)(2-a)(3-a)} + \frac{2}{2-a} H_0(2-a) + \frac{1}{n} \frac{1}{1-a} H_0(1-a) - \frac{2}{n} \frac{1}{1-a} H_1(1-a). \tag{17}
\]

After a few algebraic transformations, this equation reduces to Eq. (12). Note, the risk function is independent of \( n \).

Finally, let us examine the special case:

1. \( a = 1 \)

\[
R_{n,1} = E\left[ \hat{R}_{n,1} \right] = -\frac{3n}{2} + \sum_{i=1}^{n} 2 E[u_{(i)}] - \sum_{i=1}^{n} \frac{2i - 1}{n} \ E[\ln u_{(i)}]
= -\frac{3n}{2} + \sum_{i=1}^{n} 2 \frac{B(i + 1, n - i + 1)}{B(i, n - i + 1)} - \sum_{i=1}^{n} \frac{2i - 1}{n} \left( \psi(i) - \psi(n + 1) \right)
= -\frac{3n}{2} + 2 H_0(1) - \sum_{i=1}^{n} \frac{2i - 1}{n} \left( \psi(i) - \psi(n + 1) \right) \tag{18}
\]

For the last sum, we use the computation formulae presented by Aggarwal (1955, Eq. (59) therein), with \( \psi(i) \) being the digamma function, cf. Abramowitz and Stegun (2014). Then,
The risk function depending on the stress parameter. For integer values, the corresponding risk is marked with bullets. The pole is marked with a thin line.

\[ R_{n,1} = -\frac{3n}{2} + n + \frac{n}{2} + \frac{1}{2} \]

\[ = \frac{1}{2} \quad (19) \]

This result is also yielded by Eq. (12) for \( a = 1 \).

To summarize the results, Fig. 1 shows the dependence of the risk function on the stress parameter, within the interval of definition \( 0 \leq a < 2 \).

An interesting result is that the risk function increases rapidly for \( a > 1 \). In our numerical studies we found that this property makes it difficult to compare different statistics with \( a > 1 \), especially for small samples. In our applications, see Sect. 5, we will take that into account.

4 Corollaries

Because of the symmetry of the two families of statistics from Sect. 2, we note the following:

**Corollary 1** Let \( \hat{R}_{n,b} \) be the weighted mean square error defined by Eq. (2) with the weight function in Eq. (5). Then, for \( 0 \leq b < 2 \), the risk function is

\[ R_{n,b} = \frac{1}{(2 - b)(3 - b)}. \quad (20) \]

**Proof** By direct calculation with regard to the coordinate transformation \( Z = -X \) of the random variable.

In the special cases of the Cramér-von Mises statistic and the Anderson-Darling statistic, the following two corollaries hold:
The risk function of the goodness-of-fit tests for tail models

Corollary 2 Let \( W_n^2 = \hat{R}_{n,CM} \) (Eq. (2)) with weight functions Eq. (4) for \( a = 0 \) be the Cramér-von Mises statistic Eq. (8). Then, the risk function is

\[
R_{n,CM} = E \left[ W_n^2 \right] = \frac{1}{6}.
\]  \hspace{1cm} (21)

Proof We have \( \hat{R}_{n,CM} = \hat{R}_{n,a=0} \). Hence,

\[
R_{n,CM} = E \left[ W_n^2 \right] = E \left[ \hat{R}_{n,CM} \right] = E \left[ \hat{R}_{n,a=0} \right] \overset{\text{Eq. (12)}}{=} \frac{1}{6}.
\]  \hspace{1cm} (22)

Note: The risk function calculated here is in accordance with the result of Aggarwal (1955, p. 453 below). Note that Aggarwal (1955) defined the loss function without multiplication by \( n \).

Corollary 3 Let \( A_n^2 = \hat{R}_{n,AD} \) be the Anderson-Darling statistic (Eq. (2)) with the weight function in Eq. (3) for \( a = b = 1 \). Then, the risk function is

\[
R_{n,AD} = E \left[ A_n^2 \right] = 1.
\]  \hspace{1cm} (23)

Proof Because of the identity \( w(t) = \frac{1}{i(1-t)} = \frac{1}{t} + \frac{1}{1-t} \), the expression \( \hat{R}_{n,AD} = \hat{R}_{n,a=1} + \hat{R}_{n,b=1} \) is equal to the well-known Anderson-Darling statistic \( A_n^2 \) (Anderson and Darling 1952, 1954). Hence,

\[
R_{n,AD} = E \left[ A_n^2 \right] = E \left[ \hat{R}_{n,AD} \right] = E \left[ \hat{R}_{n,a=1} \right] + E \left[ \hat{R}_{n,b=1} \right] \overset{\text{Eqn. (12, 20)}}{=} 1.
\]  \hspace{1cm} (24)

Note: The risk function calculated here is in accordance with the result of Aggarwal (1955, p. 461 above).

For further considerations in the context of decision theory, the following corollary and the discrete probability distribution defined in it may be of interest because they can be used as a parametric probability model in a decision problem for a finite number of environmental states.

Corollary 4 Let \( X \) be a discrete random variable with finite range \( i = 1, \ldots, n \). Then,

\[
p(i; \nu) = \frac{\nu + 1}{n} \frac{B(i + \nu, n - i + 1)}{B(\nu + 1, n - i + 1)}.
\]  \hspace{1cm} (25)

where \( \nu > -1 \) and integer \( n > 0 \), assigns a probability to each value in the range of \( X \). The tuple \( p(\nu) \in [0, 1]^n \) with entries \( p(i; \nu) \in [0, 1] \) is a probability vector that, depending on \( \nu \), forms a family of discrete probability distributions on the finite support of non-negative integers with a cumulative distribution function.
\[ F(s; \nu) = \frac{B(n, \nu + 1)}{B(s, \nu + 1)} \]  

where \( s = 1, \ldots, n \).

For the discrete random variable, the following applies:

1. **k-Moment**

\[ m_k(\nu) = \sum_{l=0}^{k} S_{k+1,l+1} \frac{\nu + 1}{\nu + 1 + l} (n - 1)(l) \]  

2. **Expectation**

\[ E[X] = 1 + \frac{\nu + 1}{\nu + 2}(n - 1) \]  

3. **Variance**

\[ \text{Var}[X] = \frac{(\nu + 1)(\nu + n + 1)(n - 1)}{(\nu + 2)^2(\nu + 3)} \]  

**Proof** Eq. (32) of Lemma 1 corresponds to the \( k \)-moments of the discrete probability distribution in Eq. (27) and proves for \( k = 0 \) that \( \sum_{i} p(i; \nu) = 1 \). The alternative form

\[ p(i; \nu) = \frac{\Gamma(n)}{\Gamma(i)} \frac{\nu + 1}{(\nu + i)^{(n-i+1)}} \]  

(cf. Eq. (35) with \( k = 0 \)) shows that \( p(i; \nu) \geq 0 \) for \( \nu > -1 \). Eq. (32) for \( k = 1 \) gives the first moment of the discrete probability distribution and is equal to the expectation value in Eq. (28). With \( \text{Var}[X] = E[X^2] - (E[X])^2 \), only the calculation of the second moment \( E[X^2] \) remains. This calculation can be performed quickly using equation (32). After summarizing the terms, Eq. (29) follows.

The cumulative distribution function is obtained when all \( p(i; \nu) \) are summed for \( i = 1, \ldots, s \); where \( s \leq n \). Beginning with the alternative form of \( p(i; \nu) \) (cf. Eq. (34)) and using the techniques described in the proof of Lemma 1, the following holds

\[ F(s; \nu) = \frac{\Gamma(n)}{\Gamma(s)} \frac{(\nu + 1)}{(\nu + v + 1)} \sum_{i=1}^{s} \frac{\Gamma(i + v)}{\Gamma(i)} \]  

\[ = \frac{\Gamma(n)}{\Gamma(s)} \frac{(\nu + 1)}{(\nu + v + 1)} \sum_{i=1}^{s} \frac{\Gamma(s)}{\Gamma(i)} (v)^{(i)} \Gamma(v) \]  

\[ = \frac{\Gamma(n)}{\Gamma(s)} \frac{(\nu + 1)}{(\nu + v + 1)} \sum_{i=1}^{s} \frac{(s - 1)}{i - 1} (1)^{(s-i)} (v)^{(i)} \]  

\[ = \frac{\Gamma(n)}{\Gamma(s)} \frac{(\nu + 2)^{(s-1)}}{(\nu + v + 1)} \nu \Gamma(v) \]  

\[ \frac{\Gamma(n)}{\Gamma(s)} \frac{(\nu + v + 1)}{(\nu + v + 1)} \]
The risk function of the goodness-of-fit tests for tail models 1863

\[ \frac{\Gamma(n)}{\Gamma(s)} \frac{\Gamma(s + v + 1)}{\Gamma(n + v + 1)} \]  

(31)

The last expression is an alternative representation of Eq. (26).

The above defined \textit{discrete integrated beta distribution} can weight either the left edge, with \( v \in ]-1, 0[ \), or the right edge, with \( v \in ]0, +\infty[ \), of the interval \( i \in [1, n] \) more strongly. For \( v = 0 \), the discrete uniform distribution is reproduced.

The general distribution Eq. (25) can be used in models that occur in the context of decisions under risk. The limited number of environmental states is sorted in ascending or descending order according to their probability of occurrence. If the discrete probability distribution is to have certain properties, for example, for a given expectation value between 1 and \( n \), this distribution can be modeled with the parameter \( v \).

Note that the discrete integrated beta distribution is similar but not equal to the beta-binomial distribution \( \text{BeB}(n; \alpha, \beta) \) for definition cf. (Abramowitz and Stegun 2014). The pivotal difference is that the origin of the above distribution is based on the distribution of order statistics and not on the binomial distribution. Only for \( \alpha = v + 1 + \frac{1}{n} \) and \( \beta = 1 - \frac{1}{n} \) at very large \( n \) both distributions do have nearly the same properties.

Eugene et al. (2002) replaced the variable \( x \) of the standard beta distribution with an arbitrary distribution function \( G(x) \). The nested beta-G distribution \( \text{BeG}(x; \alpha, \beta) = I_{G(x)}(\alpha, \beta) = \frac{B(G(x); \alpha, \beta)}{B(\alpha, \beta)} \) thus represents a generalization of the standard beta distribution (where \( G(x) = x \)); see, e.g. Cordeiro et al. (2013) and the references cited therein. It is obvious, that the distribution function introduced with Eq. (26) can not be derived as a discrete analogue from the beta-G distribution.

\section{5 Application}

For a very large class of parent distribution functions, the GPD can be used as a model for the tail, cf., e.g., Embrechts et al. (2003). This class of distributions includes all common parent distributions that play a role in the financial sector. Because of that, almost no uncertainty exists regarding the model selection for the tail of the unknown parent distribution. The required quantiles can then be determined to high confidence levels with sufficient certainty. A certain threshold thus divides the parent distribution into two areas: a body and a tail region. This approach is already common practice for calculating high quantiles more accurately, as indicated in Basel Commitee (2009).

Various authors have proposed methods for determining the appropriate threshold and subsequently the GPD as model for the tail from empirical data (\textit{iid}). Most methods require the setting of parameters, which often requires experience and hinders full automation of the modeling process. Based on ideas from Huisman et al. (2001), Longin and Solnik (2001), Chapelle et al. (2005) and Crama et al. (2007) we propose the following method for the determination of the tail model.

Given a suitable distance measure \( \hat{R}_n = \hat{R}_n(F_n, \hat{F}) \) as a function of the estimated GPD as a tail model \( \hat{F}(x) \) and the empirical distribution function \( F_n \), Eq. (1), an automated modeling process can be constructed using the following algorithm.
1. Sort the random sample taken from an unknown parent distribution in descending order: \( x^{(1)} \geq x^{(2)} \geq \cdots \geq x^{(n)} \).
2. Let \( k = 2, \ldots, n \), and find, for each \( k \), the estimates of the parameters of the GPD.
   Note: For numerical reasons, we start at \( k = 2 \).
3. Calculate the probabilities \( \hat{F}(x^{(i)}) \) for \( i = 1, \ldots, k \) with the estimated GPD, and determine the distance \( \hat{R}_k \) for \( k = 2, \ldots, n \).
4. Find the index \( k^* \) of the minimum of the distance \( \hat{R}_k \).

Then, the optimal threshold value is estimated by \( \hat{u} = x^{(k^*)} \), and the model of the tail of the unknown parent distribution is given by the estimated GPD \( \hat{F}(x) \), which itself is determined from the subset \( x^{(1)} \geq x^{(2)} \geq \cdots \geq x^{(k^*)} \).

Chapelle et al. (2005) suggest using the mean squared error (MSE) as distance measure \( \hat{R}_n \). Bader et al. (2018) implement a different algorithm based on sequences of goodness-of-fit tests. They use the Cramér-von Mises statistic and the Anderson-Darling statistic in combination with some beforehand set stopping rules. In contrast, we propose to use the distance measure \( AL_n^2 \) (lower tail, \( \hat{R}_{n,a=1} \), see Eq. (9)) or its counterpart \( AU_n^2 \) (upper tail, \( \hat{R}_{n,b=1} \)) in the algorithm above. This choice is justified as follows: The distance measure should take more account of the deviations between the measured data and the modeled data, especially in the tail area. This can be achieved with asymmetrical weight functions \( a, b > 0 \), cf. Eqs. (4) and (5). Figure 1 in Sect. 3 shows, that the risk function increases rapidly for values of \( a, b > 1 \) and numerical evaluations become difficult. If one focuses on integer exponents for practical reasons and for reasons of simplification, then \( a, b = 1 \) remains a suitable choice.

A small and simple example with the frequently occurring standard normal distribution shows that this choice can lead to better results if high quantiles are to be determined. In our example a Monte Carlo simulation generates 10,000 samples of a standard normal distribution as a known parent distribution. The sample length varies from 50 to 1000 data points. The modeling process then determines the beginning of the upper tail and estimates the GPD as a tail model. Finally, the quantile \( \hat{q}_\alpha \) is estimated at the given confidence level \( \alpha = 0.999 \). We are interested in how much the estimate of the quantile deviates on average from the true quantile, \( \langle \hat{q}_\alpha \rangle - q_\alpha \), and how much the estimated quantile scatters on average, \( \langle (\hat{q}_\alpha - \langle \hat{q}_\alpha \rangle)^2 \rangle \).

With \( q_{0.999} = 3.090232 \ldots \) for the standard normal distribution, Table 1 summarizes the results for the different statistics used as distance measure \( \hat{R}_n \). Here, \( AU_n^2 \) denotes the upper tail statistic, \( W^2 \) denotes the Cramér-von Mises statistic, \( A_n^2 \) denotes the Anderson-Darling statistic, and MSE denotes the mean square error.

The MSE-Distance uses a tail section on average too long for modeling. For small datasets, even more than half of the data is used; thus, in case of the standard normal distribution, data from the lower tail are already included in the modeling. This behavior was also found in the application of the method to other parent distributions commonly used in the finance and insurance industries (not shown here: lognormal distribution, generalized extreme value distribution, generalized Pareto distribution and exponential distribution). Further on, therefore, only the \( AU_n^2, W^2 \) and \( A_n^2 \) statistics are discussed.

The analyses also confirm that there is a reasonable lower bound on the sample length of approximately 50 data points to apply the method, regardless of the dis-
The risk function of the goodness-of-fit tests for tail models

Table 1 Comparison of different statistics as the distance measure

| Statistics | Sample length | Ψ Tail length | Ψ Quantil \(\hat{q}_α\) | Ψ Deviation \(\hat{q}_α - q_α\) | Ψ Scattering \(\sqrt{\langle (\hat{q}_α - q_α)^2 \rangle}\) |
|------------|---------------|---------------|----------------|----------------|----------------|
| \(AU^2\)  | 50            | 26.0          | > 10\(^1\)   | > 10\(^1\)   | > 10\(^3\)   |
| \(W^2\)   | 50            | 20.3          | > 10\(^1\)   | > 10\(^1\)   | > 10\(^3\)   |
| \(A^2\)   | 50            | 25.7          | > 10\(^1\)   | > 10\(^1\)   | > 10\(^3\)   |
| MSE       | 50            | 34.5          | 2.490         | -0.600        | 0.817         |
| \(AU^2\)  | 100           | 39.5          | 3.360         | 0.269         | 6.691         |
| \(W^2\)   | 100           | 32.1          | 3.228         | 0.138         | 11.897        |
| \(A^2\)   | 100           | 39.1          | 3.417         | 0.327         | 13.320        |
| MSE       | 100           | 62.3          | 2.686         | -0.404        | 0.717         |
| \(AU^2\)  | 150           | 48.6          | 3.059         | -0.032        | 1.565         |
| \(W^2\)   | 150           | 40.7          | 2.984         | -0.107        | 1.243         |
| \(A^2\)   | 150           | 47.9          | 3.045         | -0.045        | 2.069         |
| MSE       | 150           | 85.3          | 2.774         | -0.316        | 0.488         |
| \(AU^2\)  | 200           | 58.3          | 3.003         | -0.087        | 0.845         |
| \(W^2\)   | 200           | 48.6          | 2.953         | -0.137        | 0.759         |
| \(A^2\)   | 200           | 57.3          | 2.981         | -0.109        | 0.852         |
| MSE       | 200           | 106.7         | 2.824         | -0.266        | 0.434         |
| \(AU^2\)  | 300           | 70.2          | 2.991         | -0.099        | 0.495         |
| \(W^2\)   | 300           | 60.5          | 2.964         | -0.126        | 0.481         |
| \(A^2\)   | 300           | 69.7          | 2.981         | -0.110        | 0.487         |
| MSE       | 300           | 142.9         | 2.891         | -0.200        | 0.361         |
| \(AU^2\)  | 500           | 90.9          | 3.000         | -0.090        | 0.346         |
| \(W^2\)   | 500           | 78.6          | 2.989         | -0.101        | 0.343         |
| \(A^2\)   | 500           | 89.8          | 2.996         | -0.094        | 0.342         |
| MSE       | 500           | 204.4         | 2.953         | -0.137        | 0.295         |
| \(AU^2\)  | 1000          | 131.9         | 3.021         | -0.069        | 0.235         |
| \(W^2\)   | 1000          | 117.0         | 3.023         | -0.067        | 0.234         |
| \(A^2\)   | 1000          | 130.1         | 3.023         | -0.067        | 0.235         |
| MSE       | 1000          | 329.8         | 3.012         | -0.078        | 0.220         |

tance measure \((AU^2, W^2 \text{ or } A^2)\) used. In the table, for data lengths of approximately 50 points, a sharp increase in the standard deviation \((\Psi \text{ Scattering})\) is observed due to the strong influence of statistical errors. Above 50 data points, the three methods show a similar behavior in terms of sample standard deviation and the average deviation from the true value of the quantile. However, the upper tail statistic \(AU^2\) \((= R_{n,b=1})\) shows a slightly smaller deviation on average from the true quantile. For very large sample lengths, the results obtained for the three statistics are almost the same. These observations were also made when applying the method to other parent distributions.
The slight advantages, however, suggest to use the lower or upper tail statistics \((AL^2, AU^2)\) as distance measure in an automated algorithm to determine the GPD as tail model from empirical data, especially in case of small samples.

### 6 Discussion and conclusion

In the present work, we have defined a family of tail statistics based on the use of asymmetric weight functions in a distance measure. As distance measure we used the weighted mean square error, which also serves the standard statistics according to Cramér-von Mises (1928, 1931) and Anderson-Darling (1952, 1954) as a starting point. The tail statistics can be parameterized by selecting a stress parameter – \(a\) (lower tail statistics) or \(b\) (upper tail statistics). The greater the stress parameter chosen, the greater the penalty for discrepancies between the measured data and the modeled data, especially in the tail area.

To investigate further properties of the tail statistics, we calculated the risk function of the statistics and found that the allowable range for the stress parameters is constrained: \(0 \leq a, b < 2\). Focusing only on integer values for the stress parameter, the result \(a, b \neq 2\) is surprising. Since the risk function for these values approaches infinity, the associated weight functions and the corresponding statistics, see e.g. Eq. (10), should not be used. In fact, during our research, we occasionally found excerpts of anonymous scripts that propose these statistics.

Further, for exponents \(1 < a, b < 2\) and small samples of financial data we found, that the evaluation of the corresponding statistics became difficult numerically. From the practical point of view we focus on integer values for the stress parameter, therefore, we suggest using statistics with \(a, b = 1\) proposed by Ahmad et al. (1988) as the basis for a goodness-of-fit test, in particular, for the tail of a distribution.

If the proposed statistic is used for \(a, b = 1\), the average loss \(R_{n,1} = \frac{1}{2}\) is slightly larger than the one for the Cramér-von Mises statistic \((R_{n,CM} = \frac{1}{5})\) but is much smaller than the one for the Anderson-Darling statistic \((R_{n,AD} = 1)\).

As an application of the tail statistics, with \(a, b = 1\), we defined an automated algorithm for finding the threshold within a sorted dataset that marks the beginning of the tail. Above the threshold, the tail of the distribution can then be modeled using the GPD to calculate the required high quantiles for risk assessment. Compared to standard methods, there are slight advantages in using these tail statistics, especially for small samples. Further studies will show how well the procedure works when examining real data in practice.

Outlook: The average losses detected by the risk function can be further minimized if the empirical distribution function Eq. (1) is not used to determine the weighted mean square error. Instead, in order to improve the results, the empirical probability may be evaluated depending on the selected weight function, cf. e.g., Ferguson (1967), or replaced by a weighted empirical characteristic function, cf. e.g., Meintanis et al. (2016), which again leads to new families of statistics and risk functions and opens another field of research.
Acknowledgements The authors would like to thank the editor and the two anonymous referees for their helpful comments.

Appendix A: Auxiliary calculations

Lemma 1 Let

\[ m_k(\nu) = \frac{\nu + 1}{n} \sum_{i=1}^{n} i^k \frac{B(i + \nu, n - i + 1)}{B(i, n - i + 1)} \]  

then

\[ m_k(\nu) = \sum_{l=0}^{k} S_{k+1, l+1} \frac{\nu + 1}{\nu + 1 + l} (n - 1)_{(l)}, \]  

where \( \nu \in \mathbb{R}, k \in \mathbb{N}, (n - 1)_{(l)} \) is the Pochhammer notation for falling factorials and \( S_{k,l} \) are the Stirling numbers of the second kind (Abramowitz and Stegun 2014).

Proof (Lemma 1) To begin, the beta functions \( B(\cdot, \cdot) \) of Eq. (32) are expressed in terms of the gamma function \( \Gamma(\cdot) \) (Abramowitz and Stegun 2014). Simplifying the fraction yields

\[ m_k(\nu) = (\nu + 1) \sum_{i=1}^{n} i^k \frac{\Gamma(n)}{\Gamma(i)} \frac{\Gamma(i - 1 + \nu + 1)}{\Gamma(n + \nu + 1)}. \]  

Depending on \( \nu \), the possibly resulting poles must be considered and the gamma function should be considered in its analytic continuation \( \Gamma(x + \alpha) = (x)^{({\alpha})} \Gamma(x) \), where \( (x)^{({\alpha})} \) is the Pochhammer notation for rising factorials (Abramowitz and Stegun 2014). After simplifying the fraction, we receive

\[ m_k(\nu) = (\nu + 1) \sum_{i=1}^{n} i^k \frac{\Gamma(n)}{\Gamma(i)} \frac{1}{(v + 1 + [i - 1])(n - [i - 1])}. \]  

Using the identity \( (x)^{(\alpha)} = (x)^{(\beta)}(x + \beta)^{(\alpha - \beta)} \) results in

\[ m_k(\nu) = (\nu + 1) \sum_{i=1}^{n} i^k \frac{\Gamma(n)}{\Gamma(i)} \frac{(v + 1)(i - 1)}{(v + 1)(n - i)}. \]  

Remember that \( \frac{\Gamma(n)}{\Gamma(i)} = \binom{n}{i} \frac{n}{n - i} = \binom{n-1}{i-1} (1)^{(n-i)}. \) Then,

\[ m_k(\nu) = \frac{(\nu + 1)}{(v + 1)(n)} \sum_{i=1}^{n} i^k \binom{n-1}{i-1} (1)^{(n-i)}(v + 1)^{(i-1)}. \]
Now, $i^k$ is decomposed into a sum of the falling factorials, where the coefficients consist of Stirling numbers of the second kind $i^k = \sum_{l=0}^{k} S_k,l(i)(l)$. Using the appropriate numbering of the sum with an appropriate extension of the terms gives

$$m_k(v) = \frac{(v + 1)}{(v + 1)^n} \times \sum_{i=1}^{n} \sum_{l=1}^{k+1} S_{k+1,l}(n - 1)(l-1) \frac{(i - 1)(l-1)}{(n - 1)(l-1)} \frac{(n - l)}{(i - l)} (1)^{(n-l)} (v + 1)^{(i-1)}.$$  (38)

By changing the order of the sums and truncating the binomial coefficient, the above equation reduces to

$$m_k(v) = \frac{(v + 1)}{(v + 1)^n} \times \sum_{l=1}^{k+1} S_{k+1,l}(n - 1)(l-1) \sum_{i=1}^{n} \frac{(n - l)}{(i - l)} (1)^{(n-l)} (v + 1)^{(i-1)}.$$  (39)

By renumbering the last sum, splitting the term $(v + 1)^{(l-1)}$ and using Chu-Vandermonde theorem (Oldham et al. 2009, Ch. 18), this equation reduces to

$$m_k(v) = \frac{(v + 1)}{(v + 1)^n} \times \sum_{l=1}^{k+1} S_{k+1,l}(n - 1)(l-1) (v + 1)^{(l-1)} (v + 1 + l)^{(n-l)}.$$  (40)

After renumbering the remaining sum and multiplying the rising factorials, Eq. (32) follows immediately. $\square$

References

Abramowitz M, Stegun IA (eds) (2014) Handbook of mathematical functions with formulas, graphs, and mathematical tables, National Bureau of Standards (Washington DC), Martino Publishing, reprint of 1964 Edition, Mansfield Centre
Aggarwal OP (1955) Some minimax invariant procedures for estimating a cumulative distribution function. Ann Math Stat 26:450–463
Ahmad MI, Sinclair CD, Spurr BD (1988) Assessment of flood frequency models using empirical distribution function statistics. Water Resour Res 24:1323–1328
Anderson TW, Darling DA (1952) Asymptotic theory of certain ‘Goodness of Fit’ criteria based on stochastic processes. Ann Math Stat 23:193–212
Anderson TW, Darling DA (1954) A test of goodness of fit. J Am Stat Assoc 49:765–769
Bader B, Yan J, Zhang X (2018) Automated threshold selection for extreme value analysis via Goodness-of-Fit tests with application to batched return level mapping. Ann Appl Stat 12:310–329
Basel Commitee on Banking Supervision (2009) Observed range of practice in key elements of advanced measurement approaches (AMA). Bank for International Settlements, Basel
Chapelle A, Crama Y, Hübner G, Peters JP (2005) Measuring and managing operational risk in financial sector: an integrated framework. Soc Sci Res Netw Electron J. https://doi.org/10.2139/ssrn.675186
Cordeiro GM, Christino CT, Hashimoto EM, Ortega EMM (2013) The beta generalized Rayleigh distribution with application to lifetime data. Stat Pap 54:133–161
Crama Y, Hübner G, Peters JP (2007) Impact of the collection threshold on the determination of the capital charge for operational risk. In: Greg N (ed) Adv Risk Manag. Gregoriou, Palgrave Macmillan, New, York, pp 1–16
Cramér H (1928) On the composition of elementary errors, second paper: statistical applications. Scand Actuar J 1:141–180
Denecke L, Müller ChH (2012) Consistency and robustness of tests and estimators based on depth. J Stat Plan Inference 142:2501–2517
Denecke L, Müller ChH (2014) New robust tests for the parameters of the Weibull distribution for complete and censored data. Metrika 77:585–607
Embrechts P, Klüppelberg C, Mikosch T (2003) Modelling extremal events: for insurance and finance (stochastic modelling and applied probability). Springer, corrected 4th printing, Berlin, Germany
Eugene N, Lee C, Famoye F (2002) Beta-normal distribution and its applications. Commun Stat 31:497–512
Ferguson TS (1967) Math Stat A. Academic Press, New York
Huisman R, Koedijk KG, Kool CJM, Palm F (2001) Tail-Index estimates in small samples. J Bus Econ Stat 19:208–216
Kolmogorov AN (1933) Sulla determinazione empirica di una legge di distribuzione. Giornale dell’Istituto Italiano degli Attuari 4:83–91
Kotz S, Seier E (2009) An analysis of quantile measures of kurtosis: center and tails. Stat Pap 50:553–568
Longin F, Solnik B (2001) Extreme correlation of international equity markets. J Financ 56:649–676
Meintanis SG, Allison J, Santana L (2016) Goodness-of-fit tests for semiparametric and parametric hypotheses based on the probability weighted empirical characteristic function. Stat Pap 57:957–976
Oldham K, Myland J, Spanier J (2009) An atlas of functions, 2nd edn. Springer, New York
Sinclair CD, Spurr BD, Ahmad MI (1990) Modified Anderson Darling test. Commun Stat 19:3677–3686
Smirnov NV (1936) Sur la distribution de ω²-criterion (critérion de R. von Mises). C R l’Académie Sci Paris 202:449–452
Von Mises RE (1931) Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik. Franz Deuticke, Leipzig und Wien
Wagner N, Marsh TA (2004) Tail index estimation in small samples—simulation results for independent and ARCH-type financial return models. Stat Pap 45:545–562

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.