Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Tail Analysis Without Parametric Models: A Worst-Case Perspective

Henry Lam, Clementine Mottet

To cite this article:
Henry Lam, Clementine Mottet (2017) Tail Analysis Without Parametric Models: A Worst-Case Perspective. Operations Research 65(6):1696-1711. https://doi.org/10.1287/opre.2017.1643

http://pubsonline.informs.org/journal/opre/

OPERATIONS RESEARCH
Vol. 65, No. 6, November–December 2017, pp. 1696–1711
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Tail Analysis Without Parametric Models: A Worst-Case Perspective

Henry Lam,[a] **Clementine Mottet**[b]

[a] Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027; [b] Department of Mathematics and Statistics, Boston University, Boston, Massachusetts 02215

Contact: khl2114@columbia.edu, http://orcid.org/0000-0002-3193-563X (HL); cmottet@bu.edu (CM)

Received: April 29, 2015
Revised: October 11, 2016
Accepted: March 31, 2017
Published Online in Articles in Advance: September 25, 2017

Subject Classifications: statistics: nonparametric; probability: distributions; programming: infinite dimensional, nonlinear
Area of Review: Stochastic Models

https://doi.org/10.1287/opre.2017.1643

Abstract. A common bottleneck in evaluating extremal performance measures is that, because of their very nature, tail data are often very limited. The conventional approach selects the best probability distribution from tail data using parametric fitting, but the validity of the parametric choice can be difficult to verify. This paper describes an alternative based on the computation of worst-case bounds under the geometric premise of tail convexity, a feature shared by all common parametric tail distributions. We characterize the optimality structure of the resulting optimization problem, and demonstrate that the worst-case convex tail behavior is in a sense either extremely light tailed or extremely heavy tailed. We develop low-dimensional nonlinear programs that distinguish between the two cases and compute the worst-case bound. We numerically illustrate how the proposed approach can give more reliable performances than conventional parametric methods.

Funding: The authors gratefully acknowledge support from the National Science Foundation [Grants CMMI-1400391/1542020 and CMMI-1436247/1523453].
Supplemental Material: The online appendix is available at https://doi.org/10.1287/opre.2017.1643.

Keywords: tail modeling • robust analysis • nonparametric

1. Introduction

Modeling extreme behaviors is a fundamental task in analyzing and managing risk. As the earliest applications, hydrologists and climatologists study historical data of sea levels and air pollutants to estimate the risk of flooding and pollution (Gumbel 2012). In nonlife or casualty insurance, insurers rely on accurate prediction of large losses to price and manage insurance policies (McNeil 1997, Beirlant and Teugels 1992, Embrechts et al. 1997). Relatedly, financial managers estimate risk measures of portfolios to safeguard losses (Glasserman and Li 2005; Glasserman et al. 2007, 2008). In engineering, the measurement of system reliability often involves modeling the tail behaviors of individual components' failure times (Nicola et al. 1993, Heidelberger 1995).

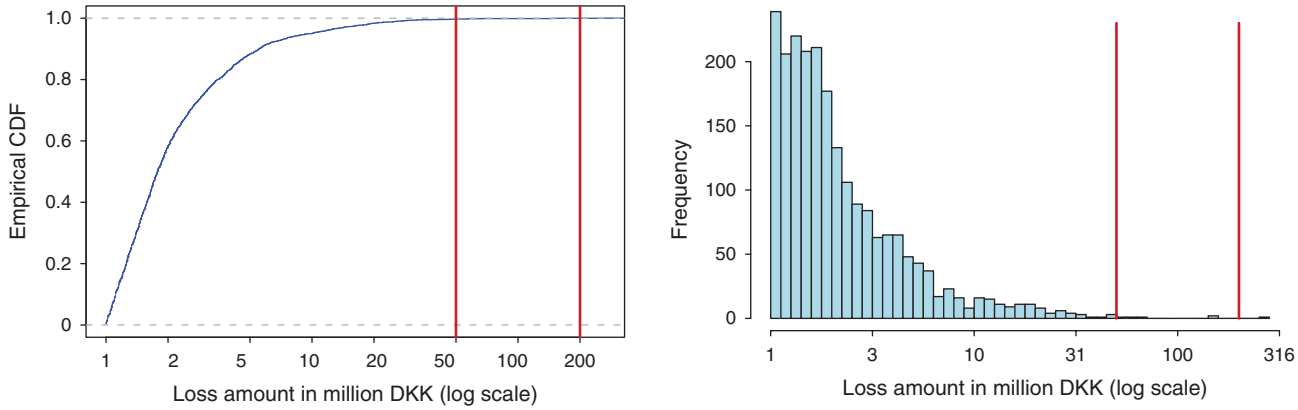
Despite its importance in various disciplines, tail modeling is an intrinsically difficult task because, by their own nature, tail data are often very limited. Consider these two examples:

Example 1 (Adopted from McNeil 1997)**.** There were 2,156 Danish fire losses amounting to over one million Danish Krone (DKK) from 1980 to 1990. The empirical cumulative distribution function (ECDF) and the histogram (in log scale) are plotted in Figure 1. For a concrete use of the data, an insurance company might be interested in pricing a high-excess contract with reinsurance, which has a payoff of $X - 50$ (in million DKK) when $50 < X \leqslant 200$, 150 when $X > 200$, and 0 when $X \leqslant 50$, where X is the loss amount (the marks 50 and 200 are labeled with vertical lines in Figure 1). Pricing this contract would require, among other information, $E[\text{payoff}]$. However, only seven data points are above 50 (the loss amount above which the payoff is nonzero).

Example 2. A more extreme situation is a synthetic data set of size 200 generated from an unknown distribution, whose histogram is shown in Figure 2. Suppose the quantity of interest is $P(4 < X < 5)$. This appears to be an ill-posed problem since the interval $[4, 5]$ has no data at all. This situation is not uncommon when in any application one tries to extrapolate the tail with a small sample size.

The purpose of this paper is to develop a theoretically justified methodology to estimate tail-related quantities of interest such as those depicted in the examples above. This requires drawing information properly from data not in the tail. We will illustrate how to do this and revisit the two examples later with numerical performance of our method.

Figure 1. (Color online) ECDF and Histogram for Danish Fire Losses from 1980 to 1990

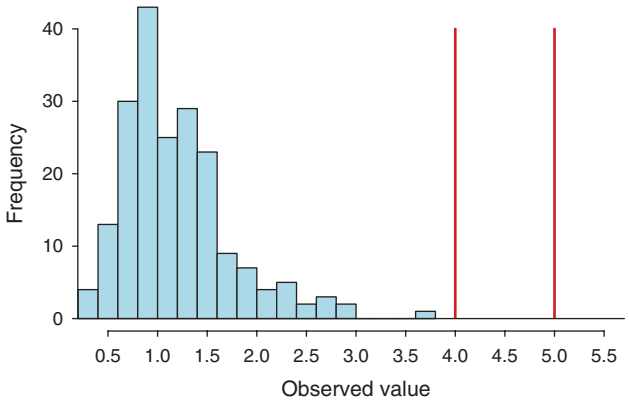


2. Our Approach and Main Contributions

We adopt a nonparametric approach. Rather than fitting a tail parametric curve when there can be few or zero observations in the tail region, we base our analysis on the geometric premise that the tail density is convex. We emphasize that this condition is satisfied by *all* common parametric distributions (e.g., normal, lognormal, exponential, gamma, Weibull, Pareto etc.). For this reason we believe it is a natural and minimal assumption to make.

In any given problem, there can be potentially infinitely many feasible candidates of convex tails. The central idea of our method is a worst-case characterization. Formally, given information on the nontail part of the distribution and a target quantity of interest (e.g., $P(4 < X < 5)$ in Example 2), we aim to find a convex tail, consistent with the nontail part, that gives rise to the worst-case value of the target (e.g., the largest possible value of $P(4 < X < 5)$). This value serves as a tight bound for the target that is robust with respect to the ambiguity of the tail, without using any particular tail knowledge other than our a priori assumption of convexity.

Figure 2. (Color online) Histogram of a Synthetic Data Set with Sample Size 200



Our proposed approach requires solving an optimization over a potentially infinite-dimensional space of convex tails. As our key contributions, we show that this problem has a very simple optimality structure, and find its solution via low-dimensional nonlinear programs. In particular, we have the following:

1. We characterize the worst-case tail behavior under the tail convexity condition. We show that the worst-case tail, for *any* bounded target quantity of interest, is in a sense either *extremely light tailed or extremely heavy tailed*. Both cases can be characterized by piecewise linear densities, the distinction being whether the pieces form a bounded support distribution or lead to probability masses that escape to infinity.

2. We provide efficient algorithms to distinguish between the two cases above, and to solve for the optimal distribution in each case. For a large class of objectives, the algorithm requires at most a two-dimensional nonlinear program.

Our approach outputs statistically valid worst-case bounds when integrating with confidence estimates drawn from the nontail portion of the data. This approach uses the convexity assumption to get around the difficulty faced by conventional parametric methods (discussed in detail in the next section) in directly estimating the tail curve, by effectively mitigating the estimation burden to the central part of the density curve where more data are available. However, we pay the price of conservativeness: our method can generate a worst-case bound that is overpessimistic. We therefore believe it is most suitable for a small sample size, when a price of conservativeness is unavoidable in trading with statistical validity.

The remainder of this paper is organized as follows. Section 3 discusses some previous techniques and reviews the relevant literature. Section 4 presents our formulation and results for an abstract setting. Section 5 studies the numerical solution algorithm. Section 6 focuses on integrating these results with data. Section 7 shows some numerical illustration. Section 8

concludes and discusses future work. Some auxiliary theorems and proofs are left to the online appendix.

3. Related Work

3.1. Overview of Common Tail-Fitting Techniques

As far as we know, all existing techniques for modeling extreme events are parametric based, in the sense that a "best" parametric curve is chosen and the parameters are fit to the tail data. The classic text of Hogg and Klugman (2009) provides a comprehensive discussion on the common choices of parametric tail densities. While exploratory data analysis, such as quantile plots and mean excess plots, can provide guidance regarding the class of parametric curves to use (such as heavy, middle, or light tail), this approach is limited by its reliance on a large amount of data in the tail and subjectivity in the choice of parametric curve.

Beyond the goodness-of-fit approach, there are two widely used results on the parametric choice that is provably suitable for extreme values. The Fisher-Tippett-Gnedenko Theorem (Fisher and Tippett 1928, Gnedenko 1943) postulates that the sample maxima, after suitable scaling, must converge to a generalized extreme value (GEV) distribution, given that it converges at all to some nondegenerate distribution. This result is useful if the data are known to derive from the maximum of some distributions. For instance, environmental data on sea level and river heights are often collected as annual maxima (Davison and Smith 1990), and in this scenario it is sensible to fit the GEV distribution. In other scenarios, the data have to be predivided into blocks and blockwise maxima have to be taken in order to apply GEV, but this blockwise approach is statistically wasteful (Embrechts et al. 2005).

The Pickands-Balkema-de Haan Theorem (Pickands 1975, Balkema and De Haan 1974) does not require data to come from maxima. Rather, the theorem states that the excess losses over thresholds converge to a generalized Pareto distribution (GPD) as the thresholds approach infinity, under the same conditions as the Fisher-Tippett-Gnedenko Theorem. The Pickands-Balkema-de Haan theorem provides a solid mathematical justification for using GPD to fit the tail portion of data (McNeil 1997, Embrechts et al. 2005). Fitting GPD can be done by well-studied procedures such as maximum likelihood estimation (Smith 1985), and the method of probability-weighted moments (Hosking and Wallis 1987). The Hill estimator (Hill 1975, Davis and Resnick 1984) is also a widely used alternative.

Despite the attraction and frequent usage, fitting GPD suffers from two pitfalls: First, there is no convergence rate result that tells how high a threshold should be for the GPD approximation to be valid (e.g., McNeil 1997). Hence, picking the threshold is an ad hoc task in practice. Second, and more importantly, even if the threshold chosen is sufficiently high for the approximation to hold, a large amount of data above it is needed to accurately estimate the parameters in GPD. In our two examples, especially Example 2, this is plainly impossible.

3.2. Related Literature on Our Methodology

Our mathematical formulation and techniques are related to two lines of literature. The use of convexity and other shape constraints (such as log-concavity) have appeared in density estimation (Cule et al. 2010, Seregin and Wellner 2010, Koenker and Mizera 2010) and convex regression (Seijo et al. 2011, Hannah and Dunson 2013, Lim and Glynn 2012) in statistics. A major reason for using convexity in these statistical problems is the removal of tuning parameters, such as bandwidth, as required by other methods such as the use of kernel.

The second line of related literature is optimization over probability distributions, which have appeared in decision analysis (Smith 1995, Bertsimas and Popescu 2005, Popescu 2005), robust control theory (Iyengar 2005, El Ghaoui and Nilim 2005, Petersen et al. 2000, Hansen and Sargent 2008), distributionally robust optimization (Delage and Ye 2010, Goh and Sim 2010), and stochastic programming (Birge and Wets 1987, Birge and Dulá 1991). The typical formulation involves optimization of some objective governed by a probability distribution that is partially specified via constraints like moments (Karr 1983, Winkler 1988) and statistical distances (Ben-Tal et al. 2013). Our formulation differs from these studies by its pertinence to tail modeling (i.e., knowledge of certain regions of the density, but none beyond it). Among all the previous works, only Popescu (2005) has considered convex density assumption as an instance of a proposed class of geometric conditions that are added to moment constraints. While the result bears similarity to ours in that a piecewise linearity structure shows up in the solution, our qualitative classification of the tail, the solution techniques, and the formulation in integrating with data all differ from the semidefinite programming approach in Popescu (2005).

4. Abstract Formulation and Results

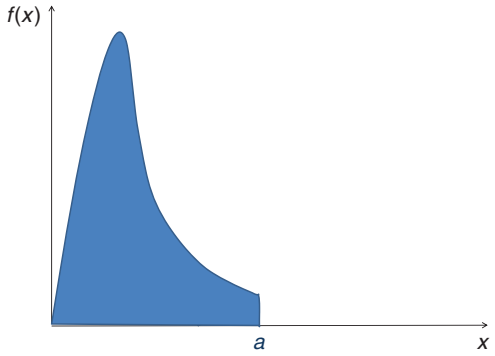
We begin by considering an abstract formulation assuming full information on the distribution up to some threshold, and no information beyond. The next subsections give the details.

4.1. Formulation

Consider a continuous probability distribution on $\mathbb{R}$ whose density exists and is denoted by $f(x)$. We assume that f is known up to a certain large threshold, say $a \in \mathbb{R}$. The goal is to extrapolate f.

We impose the assumption that $f(x)$, for $x \geqslant a$, is convex. Figure 3 shows an example of an $f(x)$ known

Figure 3. (Color online) A Probability Density $f(x)$ Known Up to a Threshold a



up to a, and Figures 4 and 5 each show an example of convex and nonconvex extrapolation. Observe that the convex tail assumption excludes any "surprising" bumps (and falls) in the density curve.

Now suppose we are given a target objective or performance measure $E[h(X)]$, where $E[\cdot]$ denotes the expectation under f, and $h\colon \mathbb{R} \to \mathbb{R}$ is a bounded function in X. The goal is to calculate the worst-case value of $E[h(X)]$ under the assumption that f is convex beyond a. That is, we want to obtain $\max E[h(X)] = \int_{-\infty}^{\infty} h(x)f(x)\,dx$ where the maximization is over all convex $f(x)$, $x \geqslant a$ such that it satisfies the properties of a probability density function. We assume that the density is left-differentiable at a, so that a convex extrapolation at a can be suitably defined. For the formulation, we need three constants extracted from $f(x)$, $x < a$, which we denote as $\eta, \nu, \beta > 0$, respectively:

1. η is the value of the density f at a, i.e., $f(a) = \eta$.
2. $-\nu$ is the left derivative of f at a, i.e., $f'_-(a) = -\nu$. We impose the condition that the right derivative $f'_+(a) \geqslant f'_-(a) = -\nu$. Note that, since f is convex (and bounded) on $[a, \infty)$, its one-sided derivative exists everywhere on $[a, \infty)$ (Rockafellar 1997, Theorem 23.1).
3. β is the tail probability at a. Since f is known up to a, $\int_{-\infty}^{a} f(x)$ is known to be equal to some number $1-\beta$, and $\int_a^{\infty} f(x)\,dx$ must equal β.

Figure 4. (Color online) An Example of Convex Tail Extrapolation

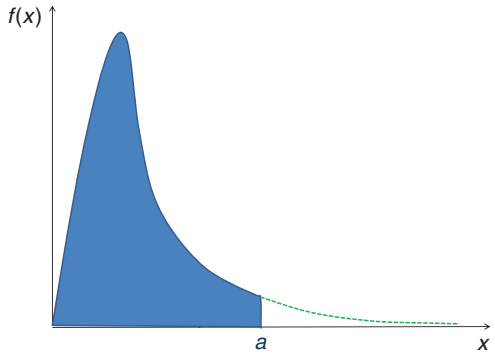


Figure 5. (Color online) An Example of Nonconvex Tail Extrapolation

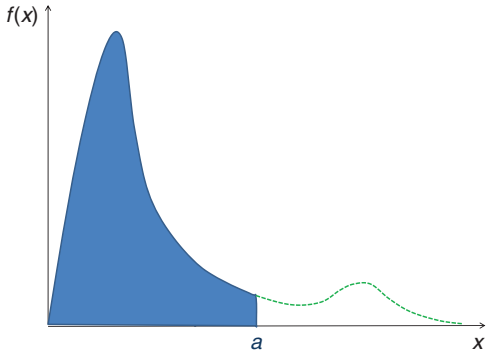


Figure 6 illustrates these quantities. For $\eta, \nu, \beta > 0$, our formulation can be written as

$$\max_f \int_a^{\infty} h(x)f(x)\,dx$$

$$\text{subject to} \int_a^{\infty} f(x)\,dx = \beta, \tag{1a}$$

$$f(a) = f(a+) = \eta, \tag{1b}$$

$$f'_+(a) \geqslant -\nu, \tag{1c}$$

$$f \text{ convex}, \quad \text{for } x \geqslant a, \tag{1d}$$

$$f(x) \geqslant 0, \quad \text{for } x \geqslant a. \tag{1e}$$

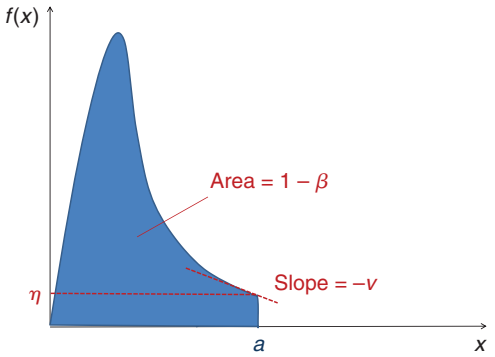
Note that we have set our objective to be $E[h(X); X \geqslant a]$, since $E[h(X); X < a]$ is completely known in this setting. Here $f(a+)$ denotes the right-limit at a, and $f(a) = f(a+)$ means that f is right-continuous at a, implying a continuous extrapolation at a.

4.2. Optimality Characterization

The solution structure of (1) turns out to be extremely simple and is characterized by either one of two closely related cases (focusing on the region $x \geqslant a$). Let $\mathscr{C}^+[a, \infty)$ denote the class of nonnegative continuous functions on $[a, \infty)$. Let

$$\mathscr{PL}_m^+[a, \infty) = \{f \in \mathscr{C}^+[a, \infty)\colon f(x) = c_j + d_j x \text{ for } x \in [y_{j-1}, y_j],\ j = 1, \ldots, m,$$

Figure 6. (Color online) The Parameters η, ν, β



$$\text{where } a = y_0 \leqslant y_1 \leqslant \cdots \leqslant y_m < \infty,$$
$$c_j, d_j \in \mathbb{R}, \text{ and } f(x) = 0 \text{ for } x > y_m\}$$

be the set of all nonnegative, continuous and piecewise linear functions on $[a,\infty)$ that have at most m line segments before vanishing. We have the following:

Theorem 1. *Suppose h is measurable and bounded. Consider optimization (1). If it is feasible, then we have either of the following*:

1. *An optimal solution f^* exists, where $f^* \in \mathscr{P}\mathscr{L}_3^+[a,\infty)$.*
2. *An optimal solution does not exist. There exists a sequence $\{f^{(k)} \in \mathscr{P}\mathscr{L}_3^+[a,\infty)\colon k \geqslant 1\}$, each $f^{(k)}$ feasible for (1), such that $\int_a^\infty h(x)f^{(k)}(x)\,dx \to Z^*$ as $k \to \infty$, where Z^* is the optimal value of (1). Moreover, let $\{c_3^{(k)} + d_3^{(k)}x\colon x \in [y_2^{(k)}, y_3^{(k)}]\}$ be the last line segment of $f^{(k)}$. We have $y_3^{(k)} \nearrow \infty$ and $d_3^{(k)} \searrow 0$ as $k \to \infty$.*

The proof of Theorem 1 is discussed in the next subsections. Note that f^* in the first case in Theorem 1 is a continuous piecewise linear density, and consequently has bounded support. In the second case, as $k \to \infty$, the sequence $\{f^{(k)}\colon k \geqslant 1\}$ has unboundedly increasing support endpoint ($y_3^{(k)} \nearrow \infty$), and its last line segment gets closer and more parallel to the horizontal axis ($d_3^{(k)} \searrow 0$). This sequence possesses a pointwise limit, but the limit is not a valid density and has a probability mass that "escapes" to positive infinity.

Figures 7 and 8 show the tail behaviors for the two cases above. A bounded support density in the first case possesses the lightest possible tail behavior. The second case, on the other hand, can be interpreted as an extreme heavy tail. Compare the sequence $f^{(k)}$ with a given arbitrary density. Given any fixed large enough x on the real line, as k grows, the decay rate of $f^{(k)}$ at the point x is eventually slower than that of the given density. Since a slower decay rate is the characteristic of a fatter tail, the behavior implied by $f^{(k)}$ in a sense captures the heaviest possible tail.

Figure 7. (Color online) Behavior of an Optimal Light-Tailed Extrapolation

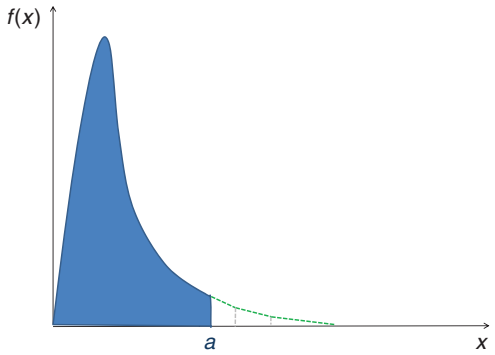
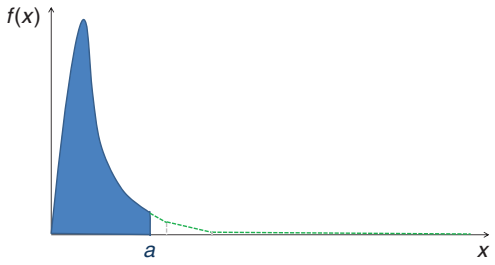


Figure 8. (Color online) Behavior of an Element in an Optimal Heavy-Tailed Extrapolation



4.3. Main Mathematical Developments

This section presents the mathematical argument for Theorem 1. This development will also help construct a solution algorithm in Section 5. We divide the argument into two parts. First we establish an equivalence of (1) to a moment-constrained optimization problem under a different probability space. Second, we characterize the solution of this moment-constrained problem, which can then be converted to the solution of (1).

We define some notations. Let $\mathbb{R}^+$ and $\mathbb{R}^-$ be the nonnegative and nonpositive real axis. Denote $\mathscr{P}(\mathscr{M})$ as the set of all probability measures on a measurable space $\mathscr{M}$ equipped with the Borel σ-field. Let $\mathscr{S}_l = \{(p_1,\ldots,p_l) \in (\mathbb{R}^+)^l\colon \sum_{i=1}^l p_i = 1\}$ be the l-dimensional probability simplex. Let $\delta(\cdot)$ be the Dirac measure. Denote $\mathscr{P}_n(\mathscr{M})$ as the set of all finite support distributions on $\mathscr{M}$ with at most n support points, i.e., each $\mathbb{P} \in \mathscr{P}_n(\mathscr{M})$ has masses $p_1, p_2, \ldots, p_n \in \mathscr{S}_n$ on points $x_1,\ldots,x_n \in \mathscr{M}$ defined such that $\mathbb{P} = \sum_{i=1}^n p_i\delta(x_i)$. For simplicity, since any $\mathbb{P} \in \mathscr{P}_n(\mathscr{M})$ can be represented by the support points $(x_1,\ldots,x_n) \in \mathscr{M}^n$ (some possibly identical) and $(p_1,\ldots,p_n) \in \mathscr{S}_n$, we sometimes write $\mathbb{P} \sim (x_1,\ldots,x_n,p_1,\ldots,p_n)$ for a given $\mathbb{P} \in \mathscr{P}_n(\mathscr{M})$. Moreover, we use the notation $\mathbb{E}[\cdot]$ to denote the associated expectation under $\mathbb{P}$.

For convenience, denote $\mathscr{P}^+ = \mathscr{P}(\mathbb{R}^+)$ as the set of all probability measures concentrated on $\mathbb{R}^+$, and $\mathscr{P}_n^+ = \mathscr{P}_n(\mathbb{R}^+)$ the corresponding set of measures with at most n support points. The measurability of h is assumed throughout the rest of the exposition.

4.3.1. Equivalence to Moment-Constrained Optimization.

We first reformulate (1) as follows:

Lemma 1. *Formulation (1) is equivalent to*

$$\max_f \int_a^\infty h(x)f(x)\,dx$$
$$\text{subject to } \int_a^\infty f(x)\,dx = \beta, \tag{2a}$$
$$f(a) = \eta, \tag{2b}$$
$$f'_+(x) \text{ exists and is nondecreasing and right-continuous, for } x \geqslant a, \tag{2c}$$
$$-\nu \leqslant f'_+(x) \leqslant 0, \quad \text{for } x \geqslant a, \tag{2d}$$

$$f'_+(x) \to 0, \quad \text{as } x \to \infty, \tag{2e}$$

$$f(x) = \int_a^x f'_+(t)\,dt + \eta, \quad \text{for } x \geqslant a. \tag{2f}$$

Proof of Lemma 1. The proof uses several elementary results from convex analysis. See Appendix EC.1 in the online appendix for details. □

As a key step, we show the equivalence of (2) to a moment-constrained program, by identifying the decision variable as $f'_+(x)$ via a one-to-one map with a probability distribution function. Let

$$H(x) = \int_0^x \int_0^u h(v+a)\,dv\,du \tag{3}$$

and

$$\mu = \frac{\eta}{\nu} \quad \text{and} \quad \sigma = \frac{2\beta}{\nu}, \tag{4}$$

where $\mu, \sigma > 0$ since we have assumed $\beta, \eta, \nu > 0$. Our result is as follows:

Theorem 2. *Suppose h is bounded. The optimal value of (2) is equal to that of*

$$\begin{aligned} \max_{\mathbb{P}} \quad & \nu\mathbb{E}[H(X)], \\ \text{subject to} \quad & \mathbb{E}[X] = \mu, \\ & \mathbb{E}[X^2] = \sigma, \\ & \mathbb{P} \in \mathscr{P}^+. \end{aligned} \tag{5}$$

Here the decision variable is a probability measure $\mathbb{P} \in \mathscr{P}^+$, and $\mathbb{E}[\,\cdot\,]$ is the corresponding expectation. Moreover, there is a one-to-one correspondence between the feasible solutions to (2) and (5), given by $f'_+(x+a) = \nu(p(x) - 1)$ for $x \in \mathbb{R}^+$, where f'_+ is the right derivative of a feasible solution f of (2) such that $f(x) = \int_a^x f'_+(t)\,dt + \eta$ for $x \geqslant a$, and p is a probability distribution function that is associated with a feasible probability measure over $\mathbb{R}^+$ in (5).

Proof of Theorem 2. The key step of the proof uses integration by parts and an explicit construction of a linear transformation between f'_+ and a probability distribution function p. See Appendix EC.1 in the online appendix for details. □

Note that ν appears in the objective function in (5) whose optimal value matches that of program (2).

4.3.2. Further Reduction and Optimality Characterization. Next we characterize the optimality structure for (5), a generalized moment problem in the form of an infinite-dimensional linear program. Using existing terminology, we call an optimization program *consistent* if there exists a feasible solution, and *solvable* if there exists an optimal solution.

For convenience, denote OPT($\mathscr{D}$) as the program

$$\begin{aligned} \max_{\mathbb{P}} \quad & \nu\mathbb{E}[H(X)] \\ \text{subject to} \quad & \mathbb{E}[X] = \mu, \\ & \mathbb{E}[X^2] = \sigma, \\ & \mathbb{P} \in \mathscr{D}, \end{aligned}$$

where H, μ, σ are defined in (3) and (4), and $\mathscr{D}$ is a collection of probability measures on $\mathbb{R}$. For example, program (5) is denoted as OPT($\mathscr{P}^+$). Moreover, let $Z(\mathbb{P}) = \nu\mathbb{E}[H(X)]$ be the objective function of OPT($\mathscr{D}$) in terms of $\mathbb{P}$. We have the following:

Theorem 3. *Program (5), or equivalently* OPT($\mathscr{P}^+$), *has the same optimal value as* OPT($\mathscr{P}_3^+$).

Proof of Theorem 3. Follows from a classical result on the extreme points of moment sets. See Appendix EC.1 in the online appendix. □

Next we derive some properties regarding the optimality of OPT($\mathscr{P}_3^+$):

Proposition 1. *Consider* OPT($\mathbb{P}_3^+$) *that is consistent. The optimal value Z^* is either achieved at some $\mathbb{P}^* \in \mathscr{P}_3^+$, or there exists a sequence of feasible $\mathbb{P}^{(k)} \in \mathscr{P}_3^+$ such that $Z(\mathbb{P}^{(k)}) \to Z^*$. In the second case, each $\mathbb{P}^{(k)} \sim (x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, p_1^{(k)}, p_2^{(k)}, p_3^{(k)})$, such that either $(x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, p_1^{(k)}, p_2^{(k)}, p_3^{(k)}) \to (x_1^*, x_2^*, \infty, p_1^*, p_2^*, 0)$ for some $x_1^*, x_2^* \in \mathbb{R}^+$ and $(p_1^*, p_2^*) \in \mathscr{S}_2$ (where x_1^* and x_2^* are possibly identical), or $(x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, p_1^{(k)}, p_2^{(k)}, p_3^{(k)}) \to (x_1^*, \infty, \infty, , ,)$ for some $x_1^* \in \mathbb{R}^+$.*

Proof of Proposition 1. See Appendix EC.1 in the online appendix. □

We are now ready to show Theorem 1:

Proof of Theorem 1. Convert the original optimization (1) into (5) by Lemma 1 and Theorem 2. If (5) is consistent, then, by Theorem 3, its optimal value is attained by the two cases in Proposition 1. Note that any solution $\mathbb{P} \in \mathscr{P}_3[0, \infty)$ represented by $(x_1, x_2, x_3, p_1, p_2, p_3)$ (where some of x_1, x_2 and x_3 are possibly identical) admits one-to-one correspondence with a solution f in (1), via $f'_+(x+a) = \nu(p(x) - 1)$ in Theorem 2, giving

$$f'_+(x) = \begin{cases} -\nu & \text{for } a \leqslant x < x_1 + a \\ -\nu(1 - p_1) & \text{for } x_1 + a \leqslant x < x_2 + a \\ -\nu(1 - p_1 - p_2) & \text{for } x_2 + a \leqslant x < x_3 + a \\ 0 & \text{for } x_3 + a \leqslant x \end{cases}$$

and hence

$$f(x) = \begin{cases} \eta - \nu(x - a) & \text{for } a \leqslant x \leqslant x_1 + a \\ \eta - \nu x_1 - \nu(1 - p_1)(x - a - x_1) & \text{for } x_1 + a \leqslant x \leqslant x_2 + a \\ \eta - \nu x_1 - \nu(1 - p_1)(x_2 - x_1) & \\ \quad - \nu(1 - p_1 - p_2)(x - a - x_2) & \text{for } x_2 + a \leqslant x \leqslant x_3 + a \\ 0 & \text{for } x_3 + a \leqslant x. \end{cases} \tag{6}$$

The first case in Proposition 1 thus concludes part 1 of Theorem 1. In the second case in Proposition 1, $x_3^{(k)} \to \infty$ and $p_3^{(k)} \to 0$ so that $1 - p_1^{(k)} - p_2^{(k)} \to 0$. Using (6), we conclude part 2 of Theorem 1. □

We close this section with two results. First is on the consistency of programs (1) and (5):

Lemma 2. *Program (5) is consistent if and only if $\sigma \geqslant \mu^2$. Correspondingly, program (1) is consistent if and only if*

$\eta^2 \leqslant 2\beta\nu$. When $\sigma = \mu^2$, (5) has only one feasible solution given by $\delta(\mu)$. Correspondingly, when $\eta^2 = 2\beta\nu$, (1) has only one feasible solution given by $f(x) = \eta - \nu(x-a)$ for $x \geqslant a$.

Proof of Lemma 2. See Appendix EC.2 in the online appendix. □

Graphically, $\eta^2 > 2\beta\nu$ implies that β is smaller than the area under the straight line starting from the point (a, η) down to the x-axis with slope $-\nu$. Hence no convex extrapolation can be drawn under this condition.

Next, we show that the boundedness assumption on h is nearly essential, in the sense that any polynomially growing h leads to an infinite optimal value for (1):

Proposition 2. *Suppose $\eta^2 < 2\beta\nu$ and $h(x) = \Omega(x^\epsilon)$ as $x \to \infty$ for some $\epsilon > 0$. The optimal value of (1) is ∞.*

Proof of Proposition 2. The proof explicitly constructs a sequence of feasible solutions that lead to exploding objective values. See Appendix EC.1 in the online appendix. □

5. Optimization Procedure for Quasi-Concave Objectives

This section develops a numerical solution algorithm for our worst-case optimization presented in Section 4. In building our algorithm, we focus on h that satisfies the following stronger assumption, which covers many natural scenarios including the two examples in the Introduction.

Assumption 1. *The function $h\colon \mathbb{R} \to \mathbb{R}^+$ is bounded, and is nondecreasing in $[a, c)$ and nonincreasing in (c, ∞) for some constant $a \leqslant c \leqslant \infty$ (i.e., c can possibly be ∞).*

Assumption 1 implies that h is quasi-concave. The nonnegativity of h is assumed without loss of generality when applied to optimization (1). Because h is bounded, one can always add a sufficiently large constant, say C, to make h nonnegative. Note that we have $E[h(X); X \geqslant a] = E[h(X)+C; X \geqslant a] - CP(X \geqslant a) = E[h(X)+C; X \geqslant a] - C\beta$, and so one can solve $E[h(X)+C; X \geqslant a]$ and recover $E[h(X); X \geqslant a]$.

We impose an additional mild regularity assumption:

Assumption 2. *The limit*

$$\lambda = \lim_{x\to\infty} \frac{H(x)}{x^2}, \tag{7}$$

where H is defined in (3), exists and is finite.

Note that when h is bounded, $H(x) = O(x^2)$ as $x \to \infty$, and $\limsup_{x\to\infty} H(x)/x^2 < \infty$. The essence of Assumption 2 is on the existence of the limit.

Under Assumption 2, denote

$$W(x_1) = \nu\left(\frac{\sigma - \mu^2}{\sigma - 2\mu x_1 + x_1^2} H(x_1) + \frac{(\mu - x_1)^2}{\sigma - 2\mu x_1 + x_1^2} H\left(\frac{\sigma - \mu x_1}{\mu - x_1}\right)\right) \tag{8}$$

for $x_1 \in [0, \mu)$ and $W(\mu) := \nu(H(\mu) + \lambda(\sigma - \mu^2))$, where μ and σ are defined in (4). We have the following strengthened version of Theorem 1:

Theorem 4. *Under Assumption 1, we have the following:*

1. *The conclusions of Theorem 1 hold with $\mathcal{PL}_3^+[a, \infty)$ replaced by $\mathcal{PL}_2^+[a, \infty)$.*
2. *Suppose $\eta^2 < 2\beta\nu$ and Assumption 2 holds additionally. The optimal value of (1) is given by $\max_{x_1 \in [0,\mu]} W(x_1)$.*
3. *Suppose $\eta^2 < 2\beta\nu$ and Assumption 2 holds additionally. If $\arg\max_{x_1\in[0,\mu]} W(x_1) \cap [0, \mu) \neq \varnothing$, then an optimal solution to (1) is given by*

$$f^*(x) = \begin{cases} \eta - \nu(x-a) & \text{for } a \leqslant x \leqslant x_1^* + a \\ \eta - \nu x_1^* - \nu \dfrac{(\mu - x_1^*)^2}{\sigma - 2\mu x_1^* + x_1^{*2}} (x - a - x_1^*) & \\ & \text{for } x_1^* + a \leqslant x \leqslant (\sigma - \mu x_1^*)/(\mu - x_1^*) + a \\ 0 & \text{for } (\sigma - \mu x_1^*)/(\mu - x_1^*) + a \leqslant x, \end{cases}$$

where $x_1^* \in \arg\max_{x_1\in[0,\mu]} W(x_1) \cap [0, \mu)$. Otherwise, we have $\arg\max_{x_1\in[0,\mu]} W(x_1) = \{\mu\}$, and there exists a sequence of feasible solutions $f^{(k)}$ with $\int_a^\infty h(x) f^{(k)}(x)\,dx \to Z^*$, where Z^* is the optimal value of (1). $f^{(k)} \to f^*$ pointwise where

$$f^*(x) = \begin{cases} \eta - \nu(x-a) & \text{for } a \leqslant x \leqslant \mu + a \\ 0 & \text{for } \mu + a \leqslant x. \end{cases}$$

The second case can occur only when $\lambda > 0$.

Part 1 of Theorem 4 simplifies the search space of densities in (1) from three to two linear segments. Because of this simplification, solving (1) reduces to finding the first kink of the optimal density (or sequence of densities), equivalently the first support point of the reformulation (5). This can be done by a one-dimensional line search $\max_{x_1\in[0,\mu]} W(x_1)$ in part 2 of the theorem.

Part 3 of Theorem 4 describes how to distinguish between the light- and heavy-tail cases in Theorem 1 by looking at the location of x_1^*. The former case occurs when there exists a x_1^* in $[0, \mu)$, and the latter occurs otherwise. Note that $f^*(x) = 0$, $x \geqslant \mu + a$ in the pointwise limit of $f^{(k)}$ in part 3 of Theorem 4 is a consequence of the last line segment of $f^{(k)}$ getting increasingly closer and more parallel to the x-axis.

Algorithm 1 summarizes the procedure for obtaining the optimal value of (1).

Algorithm 1 (Procedure for finding the optimal value of (1))
Inputs:
1. The function h that satisfies Assumptions 1 and 2.
2. The parameters $\beta, \eta, \nu > 0$.

Procedure:
1. If $\eta^2 > 2\beta\nu$, there is no feasible solution.
2. If $\eta^2 = 2\beta\nu$, the optimal value is $\nu H(\mu)$.
3. If $\eta^2 < 2\beta\nu$, the optimal value is given by $\max_{x_1\in[0,\mu]} W(x_1)$.

The rest of this section provides the developments for proving Theorem 4. First we introduce the following condition:

Assumption 3. *H is convex and H' satisfies a convex-concave property, i.e., $H'(x)$ is convex for $x \in (0, c)$ and concave for $x \in (c, \infty)$, for some $0 \leqslant c \leqslant \infty$.*

With Assumption 3, Theorem 3 can be strengthened to the following:

Proposition 3. *Under Assumption 3,* $\mathrm{OPT}(\mathscr{P}^+)$ *has the same optimal value as* $\mathrm{OPT}(\mathscr{P}_2^+)$.

Proof of Proposition 3. See Appendix EC.2 in the online appendix. □

This allows us to focus on one of the support points of $\mathrm{OPT}(\mathscr{P}_2^+)$ in the solution scheme, leading to the following proposition:

Proposition 4. *Under Assumptions 2 and 3, consider* $\mathrm{OPT}(\mathscr{P}_2^+)$ *with $\sigma > \mu^2$ and let Z^* be its optimal value.*

1. *If there exists an optimal solution in $\mathscr{P}_2^+$, then this solution has distinct support points and is represented by $(x_1^*, x_2^*, p_1^*, p_2^*)$, where $x_1^* \in \arg\max_{x_1\in[0,\mu)} W(x_1)$ and*

$$x_2^* = \frac{\sigma - \mu x_1^*}{\mu - x_1^*},\quad p_1^* = \frac{\sigma - \mu^2}{\sigma - 2\mu x_1^* + x_1^{*2}},\quad p_2^* = \frac{(\mu - x_1^*)^2}{\sigma - 2\mu x_1^* + x_1^{*2}}. \tag{9}$$

Moreover, $Z^* = \max_{x_1\in[0,\mu)} W(x_1)$.

2. *If there does not exist an optimal solution, then there must exist a sequence $\mathbb{P}^{(k)} \sim (x_1^{(k)}, x_2^{(k)}, p_1^{(k)}, p_2^{(k)}) \to (\mu, \infty, 1, 0)$. Moreover, $Z^* = \nu(H(\mu) + \lambda(\sigma - \mu^2))$.*

3. $Z^* = \max_{x_1\in[0,\mu]} W(x_1)$

Proof of Proposition 4. See Appendix EC.2 in the online appendix. □

The following corollary provides a simple sufficient condition for guaranteeing the light-tail case in the solution scheme:

Corollary 1. *Suppose Assumptions 1 and 2 hold and (1) is consistent. An optimal solution for (1) must exist if $\lambda = 0$.*

Proof of Corollary 1. By Lemma 2, consistency of (1) implies $\sigma \geqslant \mu^2$. By Theorem 2 and Proposition 3, it suffices to consider the equivalent program $\mathrm{OPT}(\mathscr{P}_2^+)$. Suppose $\lambda = 0$. If $\sigma = \mu^2$, then $\delta(\mu)$ is an optimal solution. If $\sigma > \mu^2$, then by Proposition 4, if there is no optimal solution, its optimal value must be $\nu(H(\mu) + \lambda(\sigma - \mu^2)) = \nu H(\mu)$, which is attained by $\delta(\mu)$ and leads to a contradiction (to both the hypotheses of no optimal solution and $\sigma > \mu^2$). □

We are now ready to show Theorem 4:

Proof of Theorem 4. *Proof of* 1. Assumption 1 implies Assumption 3. By Theorem 2 and Proposition 3, program (1) has the same optimal value as that of $\mathrm{OPT}(\mathscr{P}_2^+)$. Similar to the proof of Theorem 1, the result follows by noting that any $\mathbb{P} \in \mathscr{P}_2^+$ represented by (x_1, x_2, p_1, p_2) (with possibly identical x_i's) admits one-to-one correspondence with a solution f in (1), via $f'_+(x + a) = \nu(p(x) - 1)$ in Theorem 2, giving

$$f'_+(x) = \begin{cases} -\nu & \text{for } a \leqslant x < x_1 + a \\ -\nu p_2 & \text{for } x_1 + a \leqslant x < x_2 + a \\ 0 & \text{for } x_2 + a \leqslant x \end{cases}$$

and hence

$$f(x) = \begin{cases} \eta - \nu(x - a) & \text{for } a \leqslant x \leqslant x_1 + a \\ \eta - \nu x_1 - \nu p_2(x - a - x_1) & \text{for } x_1 + a \leqslant x \leqslant x_2 + a \\ 0 & \text{for } x_2 + a \leqslant x. \end{cases} \tag{10}$$

Proof of 2. The condition $\eta^2 < 2\beta\nu$ is equivalent to $\sigma > \mu^2$. The conclusion follows from part 3 in Proposition 4.

Proof of 3. The first case is obtained by substituting $x_1^* \in \arg\max_{x_1\in[0,\mu)} W(x_1)$ and x_2^*, p_2^* from (9), in part 1 in Proposition 4, into (10). The second case is obtained by substituting $(x_1^{(k)}, x_2^{(k)}, p_1^{(k)}, p_2^{(k)})$ in part 2 in Proposition 4 into (10) and taking the limit. The last conclusion follows from Corollary 1. □

6. Formulation and Procedure Under Data-Driven Environment

Sections 4 and 5 have discussed our worst-case approach in the abstract setting where the values of the needed parameters β, η, ν are completely known. In practice, these parameters are not directly specified. Instead, they are calibrated from data in the nontail region. Suppose we obtain confidence intervals (CIs) for $P(X > a)$ and $f(a)$ and a lower confidence bound for $f'_-(a)$, jointly with confidence level $1 - \alpha$. Denote them as $[\underline{\beta}, \bar{\beta}]$, $[\underline{\eta}, \bar{\eta}]$ and $-\bar{\nu}$. Suppose $\underline{\beta}, \bar{\beta}, \underline{\eta}, \bar{\eta}, \bar{\nu} > 0$. We substitute these estimates for the exact values of β, η, and $-\nu$ in our worst-case bound for $E[h(X); X \geqslant a]$:

$$\begin{aligned} \max_f \quad & \int_a^\infty h(x) f(x)\, dx \\ \text{subject to} \quad & \underline{\beta} \leqslant \int_a^\infty f(x)\, dx \leqslant \bar{\beta}, \\ & \underline{\eta} \leqslant f(a) = f(a+) \leqslant \bar{\eta}, \\ & f'_+(a) \geqslant -\bar{\nu}, \\ & f(x) \text{ convex}, \quad \text{for } x \geqslant a, \\ & f(x) \geqslant 0, \quad \text{for } x \geqslant a. \end{aligned} \tag{11}$$

It is immediate that the optimal value of (11) carries the following statistical guarantee:

Proposition 5. *Suppose that* $[\underline{\beta}, \bar{\beta}]$, $[\underline{\eta}, \bar{\eta}]$, *and* $-\bar{\nu}$ *are the joint* $(1-\alpha)$*-level CIs for* $P(X > a)$ *and* $f(a)$, *and lower confidence bound for* $f'(a)$. *Then with probability* $1-\alpha$ (*with respect to the data*) *optimization* (11) *gives an upper bound for* $E[h(X); X \geqslant a]$ *under the assumption that* $f(x)$ *is convex for* $x \geqslant a$ *and* $f(a) = f(a+)$.

Proof of Proposition 5. Let $f_{\text{true}}(x)$, $x \geqslant a$ be the ground-true density, and $Z_{\text{true}} = \int_a^\infty h(x) f_{\text{true}}(x)\,dx$. Let Z^* and $\mathscr{F}$ be the optimal value and feasible region of (11). If $f_{\text{true}} \in \mathscr{F}$, then $Z^* \geqslant Z_{\text{true}}$. Hence $P_{\text{data}}(Z^* \geqslant Z_{\text{true}}) \geqslant P_{\text{data}}(f_{\text{true}} \in \mathscr{F}) = 1-\alpha$, where P_{data} denotes the probability with respect to the data. □

For h that has support spanning across both $X < a$ and $X \geqslant a$, one approach is to estimate $E[h(X); X < a]$ separately from the computation of the worst-case bound from (11). The former can be done typically by using the empirical mean as the nontail region $X < a$ possesses more data to rely on. This segregated approach, however, only allows the conditions of valid probability density on the whole real line (e.g., $\int_{\mathbb{R}} f(x)\,dx = 1$) and the continuity at a to hold approximately but not exactly.

The following result presents the optimality structure for (11) in parallel to formulation (1).

Theorem 5. *Suppose* h *is bounded. Consider optimization* (11). *If it is feasible, then we have either one of the following*:

1. *An optimal solution* f^* *exists, where* $f^* \in \mathscr{PL}_3^+[a,\infty)$.
2. *An optimal solution does not exist. There exists a sequence* $\{f^{(k)} \in \mathscr{PL}_3^+[a,\infty): k \geqslant 1\}$, *each* $f^{(k)}$ *feasible for* (1), *such that* $\int_a^\infty h(x) f^{(k)}(x)\,dx \to Z^*$ *as* $k \to \infty$, *where* Z^* *is the optimal value of* (11). *Moreover, let* $\{c_3^{(k)} + d_3^{(k)} x: x \in [y_2^{(k)}, y_3^{(k)}]\}$ *be the last line segment of* $f^{(k)}$. *We have* $y_3^{(k)} \nearrow \infty$ *and* $d_3^{(k)} \searrow 0$ *as* $k \to \infty$.

Proof of Theorem 5. See Appendix EC.3 in the online appendix. □

Define

$$\underline{\mu} = \frac{\underline{\eta}}{\bar{\nu}}, \quad \bar{\mu} = \frac{\bar{\eta}}{\bar{\nu}}, \quad \underline{\sigma} = \frac{2\underline{\beta}}{\bar{\nu}}, \quad \bar{\sigma} = \frac{2\bar{\beta}}{\bar{\nu}}, \tag{12}$$

where $\underline{\mu}, \bar{\mu}, \underline{\sigma}, \bar{\sigma} > 0$ since we have assumed $\underline{\beta}, \bar{\beta}, \underline{\eta}, \bar{\eta}, \bar{\nu} > 0$. Define

$$\mathscr{W}(x,\omega,\rho) = \bar{\nu}\left(\frac{\rho-\omega^2}{\rho-2\omega x+x^2}H(x) + \frac{(\omega-x)^2}{\rho-2\omega x+x^2}H\left(\frac{\rho-\omega x}{\omega-x}\right)\right)$$

with $\mathscr{W}(\omega,\omega,\rho) := \bar{\nu}(H(\omega)+\lambda(\rho-\omega^2))$, where H and λ are defined as in (3) and (7).

For convenience, we also denote

$$\mathscr{K}(x;x_1,\omega,\rho) = \begin{cases} \bar{\nu}\omega - \bar{\nu}(x-a) & \text{for } a \leqslant x \leqslant x_1 + a \\ \bar{\nu}\omega - \bar{\nu}x_1 - \bar{\nu}\dfrac{(\omega-x_1)^2}{\rho-2\omega x_1+x_1^2}(x-a-x_1) & \\ \qquad \text{for } x_1 + a \leqslant x \leqslant (\rho-\omega x_1)/(\omega-x_1)+a & \\ 0 \quad \text{for } \dfrac{\rho-\omega x_1}{\omega-x_1} + a \leqslant x. & \end{cases}$$

Our data-integrated optimization (11) possesses the following consistency property in parallel to the fixed-parameter case in Lemma 2:

Lemma 3. *Program* (11) *is consistent if and only if* $\underline{\eta}^2 \leqslant 2\underline{\beta}\bar{\nu}$ *or equivalently* $\bar{\sigma} \geqslant \underline{\mu}^2$. *When* $\underline{\eta}^2 = 2\underline{\beta}\bar{\nu}$ *or equivalently* $\bar{\sigma} = \underline{\mu}^2$, (11) *has only one feasible solution given by* $f(x) = \underline{\eta} - \bar{\nu}(x-a)$ *for* $x \geqslant a$.

Proof of Lemma 3. The proof is similar to Lemma 2 and hence skipped. □

The following provides the solution scheme for our data-integrated optimization (11):

Theorem 6. *Under Assumption* 1, *we have the following*:

1. *The conclusions of Theorem* 5 *hold with* $\mathscr{PL}_3^+[a,\infty)$ *replaced by* $\mathscr{PL}_2^+[a,\infty)$.
2. *Suppose* $\underline{\eta}^2 < 2\underline{\beta}\bar{\nu}$ *and Assumption* 2 *holds additionally. The optimal value of* (11) *is given by*

$$\max\left\{\max_{\rho\in[\underline{\sigma}\vee\underline{\mu}^2,\bar{\sigma}],x_1\in[0,\bar{\mu}]} \mathscr{W}(x_1,\bar{\mu},\rho), \max_{\omega\in[\underline{\mu},\bar{\mu}\wedge\sqrt{\bar{\sigma}}],x_1\in[0,\omega]} \mathscr{W}(x_1,\omega,\bar{\sigma})\right\}. \tag{13}$$

3. *Suppose* $\underline{\eta}^2 < 2\underline{\beta}\bar{\nu}$ *and Assumption* 2 *holds additionally. Suppose*

$$\max_{\rho\in[\underline{\sigma}\vee\underline{\mu}^2,\bar{\sigma}],x_1\in[0,\bar{\mu}]} \mathscr{W}(x_1,\bar{\mu},\rho) \geqslant \max_{\omega\in[\underline{\mu},\bar{\mu}\wedge\sqrt{\bar{\sigma}}],x_1\in[0,\omega]} \mathscr{W}(x_1,\omega,\bar{\sigma}).$$

If there exists $(\rho^*,x_1^*) \in \operatorname{argmax}_{\rho\in[\underline{\sigma}\vee\underline{\mu}^2,\bar{\sigma}],x_1\in[0,\bar{\mu}]} \mathscr{W}(x_1,\bar{\mu},\rho)$ *such that* $x_1^* \in [0,\bar{\mu})$, *then an optimal solution to* (11) *is given by* $f^*(x) = \mathscr{K}(x;x_1^*,\bar{\mu},\rho^*)$. *Otherwise, there exists a sequence of feasible solutions* $f^{(k)}$ *with* $\int_a^\infty h(x) f^{(k)}(x)\,dx \to Z^*$, *the optimal value of* (11), *such that* $f^{(k)} \to f^*$ *pointwise, where*

$$f^*(x) = \begin{cases} \bar{\eta} - \bar{\nu}(x-a) & \text{for } a \leqslant x \leqslant \bar{\mu}+a \\ 0 & \text{for } \bar{\mu}+a \leqslant x, \end{cases}$$

which can occur only when $\lambda > 0$. *On the other hand, suppose*

$$\max_{\rho\in[\underline{\sigma}\vee\underline{\mu}^2,\bar{\sigma}],x_1\in[0,\bar{\mu}]} \mathscr{W}(x_1,\bar{\mu},\rho) < \max_{\omega\in[\underline{\mu},\bar{\mu}\wedge\sqrt{\bar{\sigma}}],x_1\in[0,\omega]} \mathscr{W}(x_1,\omega,\bar{\sigma}).$$

If there exists $(\omega^*,x_1^*) \in \operatorname{argmax}_{\omega\in[\underline{\mu},\bar{\mu}\wedge\sqrt{\bar{\sigma}}],x_1\in[0,\omega]} \mathscr{W}(x_1,\omega,\bar{\sigma})$ *such that* $x_1^* \in [0,\omega^*)$, *then an optimal solution to* (11) *is given by* $f^*(x) = \mathscr{K}(x;x_1^*,\omega^*,\bar{\sigma})$. *Otherwise, there exists a sequence of feasible solutions* $f^{(k)}$ *with* $\int_a^\infty h(x) f^{(k)}(x)\,dx \to Z^*$, *such that* $f^{(k)} \to f^*$ *pointwise, where*

$$f^*(x) = \begin{cases} \bar{\nu}\omega^* - \bar{\nu}(x-a) & \text{for } a \leqslant x \leqslant \omega^*+a \\ 0 & \text{for } \omega^*+a \leqslant x, \end{cases}$$

which again can occur only when $\lambda > 0$.

Proof of Theorem 6. Optimization (13) follows from a reduction of the inequality-based generalized moment problem converted from (11) into two subproblems. Appendix EC.3 in the online appendix provides the constituent propositions and further details. □

Note that in the current setting it is less straightforward to transform a problem with a general bounded h into one that has a nonnegative h than in Section 5 (see the discussion after Assumption 1), since the probability mass assigned to $[a, \infty)$ is now bounded between $\underline{\beta}$ and $\bar{\beta}$ instead of being a single specified value.

Algorithm 2 presents our procedure for solving (11).

Algorithm 2 (Procedure for finding the optimal value of (11))

Inputs:

1. The function h that satisfies Assumptions 1 and 2.
2. The parameters $\underline{\beta}, \bar{\beta}, \underline{\eta}, \bar{\eta}, \bar{\nu} > 0$.

Procedure:

1. If $\underline{\eta}^2 > 2\bar{\beta}\bar{\nu}$, there is no feasible solution.
2. If $\underline{\eta}^2 = 2\bar{\beta}\bar{\nu}$, the optimal value is $\bar{\nu}H(\underline{\mu})$.
3. If $\underline{\eta}^2 < 2\bar{\beta}\bar{\nu}$, the optimal value is

$$\max\left\{\max_{\rho\in[\underline{\sigma}\vee\underline{\mu}^2,\bar{\sigma}],x_1\in[0,\bar{\mu}]}\mathscr{W}(x_1,\bar{\mu},\rho),\ \max_{\omega\in[\underline{\mu},\bar{\mu}\wedge\sqrt{\bar{\sigma}}],x_1\in[0,\omega]}\mathscr{W}(x_1,\omega,\bar{\sigma})\right\}.$$

7. Numerical Examples

We present some numerical performance of our algorithm. We first consider several elementary examples, and then we will revisit the two examples in the introduction.

7.1. Elementary Examples

We consider three examples to demonstrate Algorithm 1.

Entropic Risk Measure. The entropic risk measure (e.g., Föllmer and Schied 2011) captures the risk aversion of users through the exponential utility function. It is defined as

$$\rho(X) = \frac{1}{\theta}\log(E[e^{-\theta X}]), \tag{14}$$

where $\theta > 0$ is the parameter of risk aversion. In the case when the distribution of the random variable X is known only up to some point a, we can find the worst-case value of the entropic risk measure subject to tail uncertainty by solving the optimization problem

$$\max_{P\in\mathscr{A}}\frac{1}{\theta}\log(E[e^{-\theta X}]) = \frac{1}{\theta}\log\left(E[e^{-\theta X};X\leqslant a]+\max_{P\in\mathscr{A}}E[e^{-\theta X};X>a]\right), \tag{15}$$

where $\mathscr{A}$ denotes the set of convex tails that match the given nontail region. Since the function $e^{-\theta X}$ satisfies Assumptions 1 and 2, we can apply Algorithm 1 to the second term of the RHS of (15). The thick line in Figure 9 represents the worst-case value of the entropic risk measure for different values of the parameter θ in the case when X is known to have a standard exponential distribution Exp(1) up to $a = -\log(0.7)$ (i.e., a is the 70-percentile and $\beta = \eta = \nu = 0.7$). For comparison, we also calculate and plot the entropic risk measure for several fitted probability distributions: Exp(1), two-segment continuous piecewise linear tail denoted as 2-PLT (two such instances in Figure 9), and mixtures of 2-PLT and shifted Pareto. Clearly, the worst-case values bound those calculated from the candidate parametric models, with the gap diminishing as θ increases.

The Newsvendor Problem. The classical newsvendor problem maximizes the profit of selling a perishable product by fulfilling demand using a stock level decision, i.e.,

$$\max_{q} E[p\min(q,D)] - cq, \tag{16}$$

where D is the demand random variable, p and c are the selling and purchase prices per product, and q is the stock quantity to be determined. We assume that $p > c$. The optimal solution to (16) is given by Littlewood's rule $q^* = F^{-1}((p-c)/p)$, where F^{-1} is the quantile function of D (Talluri and Van Ryzin 2006).

Suppose the distribution of D is only known to have the shape of a lognormal distribution with mean 50 and standard deviation 20 in the interval $[0, a)$, where a is

Figure 9. Optimal Upper Bound and Comparison with Parametric Extrapolations for the Entropic Risk Measure

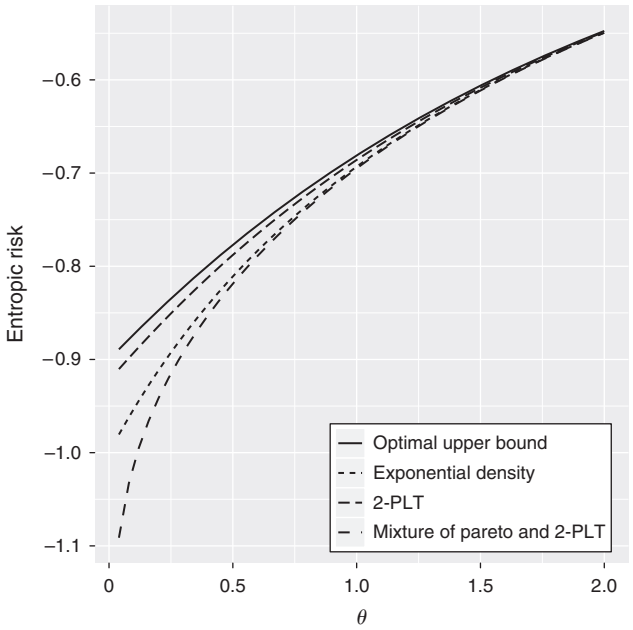
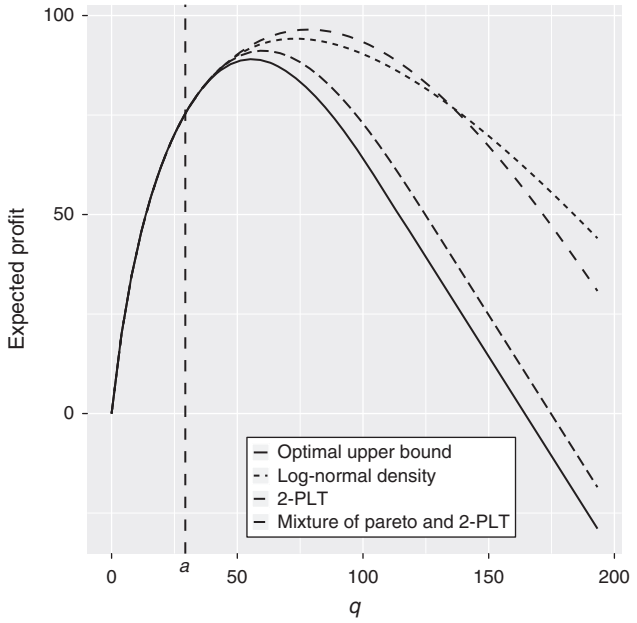


Figure 10. Optimal Objective Values of the Inner Optimization of the Robust Newsvendor Problem



the 70-percentile of the lognormal distribution. A robust optimization formulation for (16) is

$$\max_q \min_{P\in\mathcal{A}} E[p\min(q,D)] - cq$$
$$= \max_q \left\{ E[p\min(q,D); D \leqslant a] + \min_{P\in\mathcal{A}} E[p\min(q,D); D > a] - cq \right\}, \quad (17)$$

where $\mathcal{A}$ denotes the set of convex tails that match the given nontail region. The outer optimization in (17) is a concave program. We concentrate on the inner optimization. Since $p\min(q,D)$ is a nondecreasing function in D on $[0,\infty)$, its negation is nonincreasing, and Assumption 1 holds (note that minimization here can be achieved by merely maximizing the negation). Correspondingly, Assumption 2 can also be easily checked. We can therefore apply Algorithm 1 (with $\beta = 0.7$, $\eta \approx 0.007$, and $\nu \approx 0.0003$). Figure 10 shows the optimal lower bound of the inner optimization when $p = 7$, $c = 1$, and q varies between 0 and 193.26 (which is the 95-percentile of the lognormal distribution). The curve peaks at $q = 55.7$, which is the solution to problem (17). As a comparison, we also show different candidate values of the expectations that are obtained by fitting the tails of lognormal, 2-PLT (two instances), and a mixture of shifted Pareto and 2-PLT (see Figure 10).

Tail Interval Probability. Consider estimating probabilities of the type $P(c < X < d)$. We compare the bound provided by Algorithm 1 with the "truth" when X is realized from two distributions, a Pareto distribution with tail index 1, i.e., $P(X > x) = 1/x$ for all $x > 1$, and a gamma distribution with unit rate and shape parameter 2, i.e., $P(X > x) = (x+1)e^{-x}$ for all $x > 0$. Figures 11(a) and 11(b) give, for various thresholds a in percentile (shown as the x-value at the left end of each rectangle) and for various intervals (c,d) also in percentiles (shown as the y-values at the lower and upper ends of each rectangle), the ratio between the optimal upper bound and the true probability (represented by the color of each rectangle; the darker the bigger) for these two distributions, respectively. We can see that, for the Pareto case, when a is set to the 70th percentile and the interval (c,d) the (85th, 86th)-percentiles, the optimal bound given by Algorithm 1 is about twice the truth. For the same threshold a but the interval (c,d) associated with the (98th, 99th)-percentiles, the optimal bound is approximately eight times the truth. On the other hand, for the gamma case, at a equal to the 70th percentile and (c,d) the (98th, 99th)-percentiles, the bound is at most 2.1 times the truth. Figures 11(a) and 11(b) confirm the intuition that the smaller the distance between a and c, the less conservative is the bound. Moreover, the conservativeness level of our generated bound appears to depend on the true distribution. Among the two specifications, our bound is generally tighter when the truth is a gamma distribution than when it is a Pareto distribution.

7.2. Synthetic Data: Example 2 Revisited

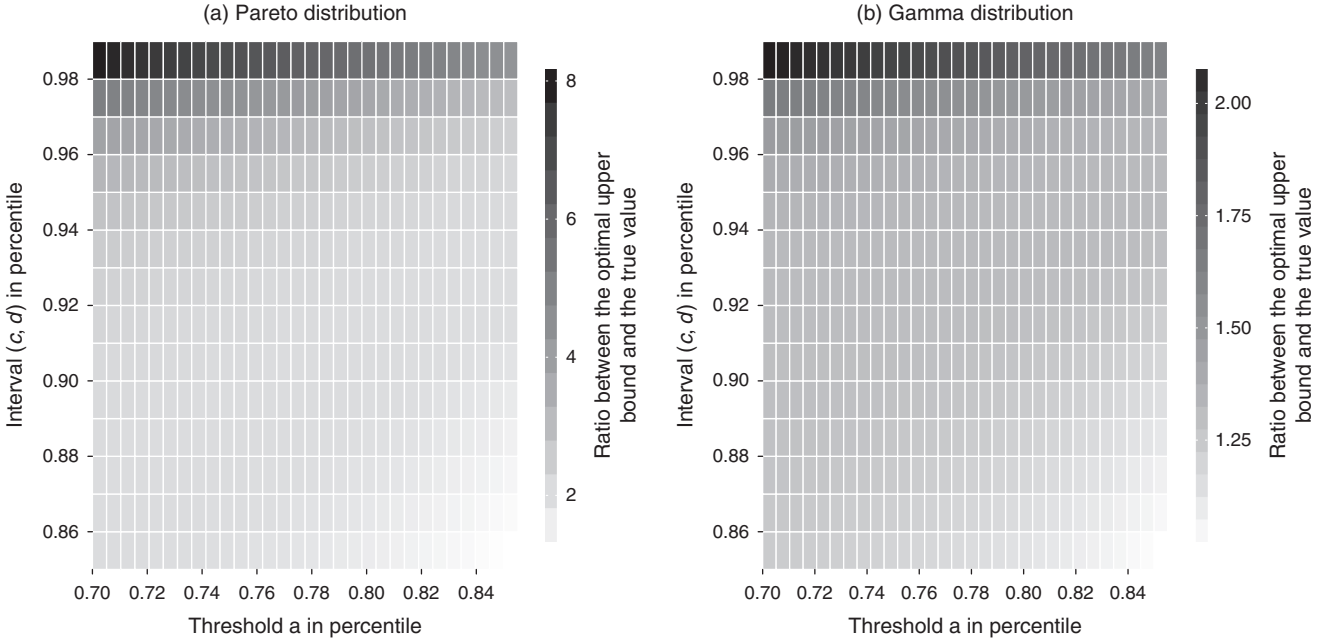
Consider the synthetic data set of size 200 in Example 2. This data set is actually generated from a lognormal distribution with parameter $(\mu,\sigma) = (0, 0.5)$, but we assume that only the data are available to us. We are interested in the quantity $P(4 < X < 5)$, and for this we will solve program (11) to generate an upper bound that is valid with 95% confidence.

We compute the interval estimates for β, η, and ν as follows. First, we obtain point estimates for these parameters through standard kernel density estimator (KDE) in the R statistical package. To obtain interval estimates, we run 1,000 bootstrap resamples and take the appropriate quantiles of the 1,000 resampled point estimates. To account for the fact that three parameters are estimated simultaneously, we apply a Bonferroni correction, so that the confidence level used for each individual estimator is $1 - 0.05/3$.

For a sense of how to choose a, Figure 12 shows the density and density derivative estimates and compares them to those of the lognormal distribution. The KDE suggests that convexity holds starting from around $x = 1.5$ (the point where the density derivative estimate starts to turn from a decreasing to an increasing function). Thus, it is reasonable to confine the choice of a to be larger than 1.5. In fact, this number is quite close to the true inflexion point 1.15.

Since the data become progressively scanter as x grows larger, and the KDE is designed to utilize neighborhood data, the interval estimators for the necessary

Figure 11. Ratio Between the Worst-Case Upper Bound and the Quantity $P(c < X < d)$, at Various Thresholds a and Intervals (c, d) in Percentiles, When X Follows Two Different Distributions



parameters β, η, and ν become less reliable for larger choices of a. For instance, Figure 12 shows that the bootstrapped KDE CI of the density derivative covers the truth only up to $x = 3.1$. In general, a good choice of a should be located at a point where there are some data in the neighborhood of a, such that the interval estimators for β, η, and ν are reliable, but as large as possible, because choosing a small a can make the tail extrapolation bound more conservative.

As a first attempt, we run Algorithm 2 using $a = 3.1$ to estimate an upper bound for the probability $P(4 < X < 5)$, which gives 8.8×10^{-3} while the truth is 2.1×10^{-3}. Thus, this estimated upper bound does cover the truth and also has the same order of magnitude. We perform the following two other procedures for comparison:

1. GPD approach: As discussed in Section 3.1, this is a common approach for tail modeling. Fit the data above a threshold u to the density function

$$(1 - \hat{F}(u))g_{\hat{\zeta},\hat{\beta}}(x - u),$$

where $\hat{F}(u)$ is the estimated ECDF at u, and $g_{\hat{\zeta},\hat{\beta}}(\cdot)$ is the GPD density, whose distribution function is defined as

$$G_{\zeta,\beta}(x) = \begin{cases} 1 - (1 + \zeta x/\beta)^{-1/\zeta} & \text{if } \zeta \neq 0 \\ 1 - \exp(-x/\beta) & \text{if } \zeta = 0, \end{cases}$$

for $x \geqslant 0$ if $\zeta \geqslant 0$ and $0 \leqslant x \leqslant -\beta/\zeta$ if $\zeta < 0$, and $\beta > 0$. Set the threshold u to be 1.8, the point at which a linear trend begins to be observed on the mean excess plot of the data, as recommended by McNeil (1997). Estimate $\hat{F}(u)$ by the sample mean of $I(X_i \leqslant u)$, where $I(\cdot)$ denotes the indicator function. Obtain the parameter estimates $\hat{\zeta}$ and $\hat{\beta}$ using the maximum likelihood estimator suggested by Smith (1987). Then use the delta method to obtain a 95% CI of the quantity $P(c < X < d)$.

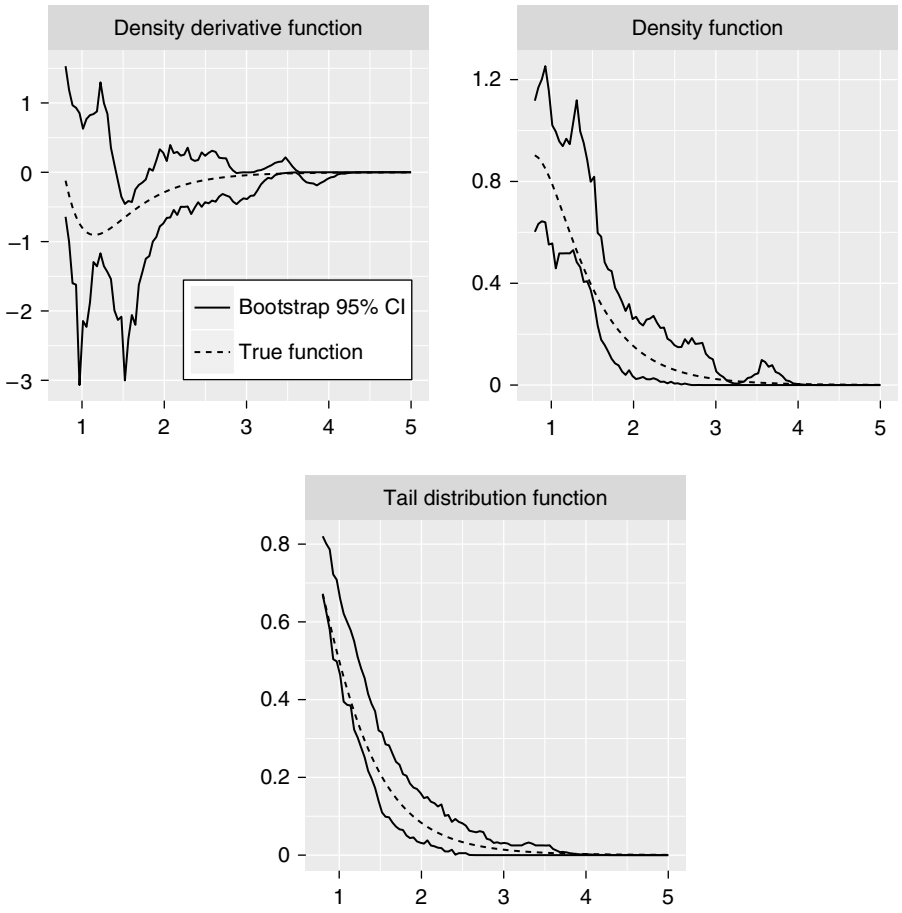
2. Worst-case approach with known parameter values: Assume β, η, and ν are known at $a = 3.1$. Then run Algorithm 1 to obtain the upper bound.

Table 1 shows the upper bounds obtained from the above approaches, and also shows the obvious fact that using ECDF alone for estimating $P(4 < X < 5)$ gives 0 since there are no data in the interval $[4, 5]$. The 95% CI output by GPD fit is $[-8.72 \times 10^{-4}, 1.10 \times 10^{-3}]$, which does not bound the truth (note that this is a two-sided interval, and the upper bound would be off even more if it had been one-sided). The worst-case approach with known parameters gives an upper bound of 3.16×10^{-3}, which is less conservative than the case when the parameters are estimated. The difference between these numbers can be interpreted as the price of estimation for β, η, and ν. For this particular setup, the

Table 1. Estimated Upper Bounds of the Probability $P(4 < X < 5)$ for the Synthetic Data in Example 2

Method	Estimated upper bound
Truth	2.14E−03
ECDF	0.00E+00
GPD	1.11E−03
Worst-case with known parameters	3.16E−03
Worst-case approach	8.80E−03

Figure 12. Bootstrapped Kernel Estimation of the Distribution, Density, and Density Derivative for the Synthetic Data



worst-case approach correctly covers the true value, whereas GPD fitting gives an invalid upper bound, thus showing that either the data size or the threshold level is insufficient to support a good fit of the GPD. This is an instance where the worst-case approach has outperformed GPD in terms of correctness.

Given that the worst-case approach with estimated parameters appears conceivably more conservative than with known parameters, we conduct a sensitivity study using only Algorithm 1. The first row in Table 2 shows the upper bound output by Algorithm 1 using the point estimates of the parameters β, η, ν. The other rows in Table 2 show the outputs of Algorithm 1 when some values of the parameters are changed to the upper estimates of the 95% CIs. Some scenarios are omitted in the table because they lead to infeasibility. We see that among all these scenarios, the most conservative upper bound occurs when β, η, ν are all set to be the upper estimates, giving to 8.67×10^{-3}, which is very close to using Algorithm 2. Note that some of these bounds do not cover the truth, which necessitates the use of the interval approach and Algorithm 2.

The above discussion focuses only on the realization of one data set, which raises the question of whether it holds more generally. Therefore, we obtain an empirical probability of coverage by repeating the following procedure 100 times:

1. Generate a lognormal sample of size 200 with parameters $(\mu, \sigma) = (0, 0.5)$.

Table 2. Sensitivity Analysis of the Worst-Case Upper Bound of $P(4 < X < 5)$ for the Synthetic Data in Example 2 Generated by Algorithm 1, When β, η, ν Are Changed from the Point Estimates to the Upper and Lower Estimates of the 95% CIs

β	η	ν	Worst-case upper bound
Estimated value	Estimated value	Estimated value	2.04E−03
Estimated value	Lower estimate	Estimated value	5.76E−06
Estimated value	Lower estimate	Upper estimate	5.76E−06
Upper estimate	Lower estimate	Estimated value	5.76E−06
Upper estimate	Lower estimate	Upper estimate	5.76E−06
Estimated value	Upper estimate	Estimated value	3.61E−04
Estimated value	Upper estimate	Upper estimate	1.62E−03
Estimated value	Estimated value	Upper estimate	2.05E−03
Upper estimate	Estimated value	Upper estimate	5.53E−03
Upper estimate	Estimated value	Estimated value	5.53E−03
Upper estimate	Upper estimate	Estimated value	8.30E−03
Upper estimate	Upper estimate	Upper estimate	8.67E−03

2. Estimate $\bar{\eta}$, $\underline{\eta}$, $\bar{\beta}$, $\underline{\beta}$, and $\bar{\nu}$ at a chosen point a (see below).

3. Use Algorithm 2 to compute the worst-case upper bound of $P(c < X < d)$.

We then estimate the coverage probability of our worst-case upper bound as the proportion of times that Algorithm 2 yields a bound that dominates the true probability $P(c < X < d)$. We repeat this procedure for different $[c, d]$ varying from $[4, 5]$ to $[9, 10]$, and for two different values of a given by 3.1 and 2.8. Tables 3 and 4 show the true probabilities, the mean upper bounds from the 100 experiments, and the empirical coverage probabilities.

The coverage probabilities in Tables 3 and 4 are mostly 1, which suggests that our procedure is conservative. For $a = 3.1$ and intervals that are close to a, i.e., $[c, d] = [4, 5]$ and $[5, 6]$, the coverage probability is not 1 but rather is close to the prescribed confidence level of 95%. Further investigation reveals that our procedure fails to cover the truth only in the case when the joint CI of the parameters η, β, and ν does not contain the true values, which is consistent with the rationale of our method. Although we have not tried lower values of a, it is very likely that in those settings the coverage probabilities will stay mostly 1, and the mean upper bounds will increase since the level of conservativeness increases.

As a comparison, Table 5 shows the results of GPD fit using the threshold $u = 1.8$. Here, all of the coverage probabilities are far from the prescribed level of 95%, which suggests that either GPD is the wrong parametric choice to use since the threshold is not high enough, or that the estimation error of its parameters is too large because of the lack of data. (Again, we have used a two-sided 95% CI for the GPD approach here; if we had used a one-sided upper confidence bound, then the upper bounding value would be even lower and the coverage probability would drop further). However, the mean upper bounds using GPD fit do cover the truth in all cases. Since the coverage probability is well below 95%, this suggests that the estimation of GPD parameters is highly sensitive to the realization of data.

Table 3. Mean Upper Bounds and Empirical Coverage Probabilities Using Worst-Case Approach with Threshold $a = 3.1$

c	d	Truth	Mean upper bound	Coverage probability
4	5	2.14E–03	1.03E–02	0.94
5	6	4.74E–04	6.12E–03	0.99
6	7	1.20E–04	4.33E–03	1.00
7	8	3.38E–05	3.35E–03	1.00
8	9	1.04E–05	2.74E–03	1.00
9	10	3.49E–06	2.31E–03	1.00

Table 4. Mean Upper Bounds and Empirical Coverage Probabilities Using Worst-Case Approach with Threshold $a = 2.8$

c	d	Truth	Mean upper bound	Coverage probability
4	5	2.14E–03	1.31E–02	1.00
5	6	4.74E–04	8.26E–03	1.00
6	7	1.20E–04	6.04E–03	1.00
7	8	3.38E–05	4.76E–03	1.00
8	9	1.04E–05	3.92E–03	1.00
9	10	3.49E–06	3.34E–03	1.00

Table 5. Mean Upper Bounds and Empirical Coverage Probabilities Using GPD

c	d	Truth	Mean upper bound	Coverage probability
4	5	2.14E–03	3.87E–03	0.62
5	6	4.74E–04	1.27E–03	0.53
6	7	1.20E–04	5.48E–04	0.51
7	8	3.38E–05	2.79E–04	0.43
8	9	1.04E–05	1.62E–04	0.40
9	10	3.49E–06	1.03E–04	0.37

In summary, Tables 3–5 show the pros and cons of our worst-case approach and GPD fitting. GPD is on average closer to the true target quantity, but its confidence upper bound can fall short of the prescribed coverage probability (in fact, only between 37% to 62% of the time it covers the truth in Table 5). On the other hand, our approach gives a reasonably tight upper bound when the interval in consideration (i.e., $[c, d]$) is close to the threshold a, and tends to be more conservative far out. This is a drawback, but sensibly so, given that the uncertainty of extrapolation increases as it gets farther away from what is known.

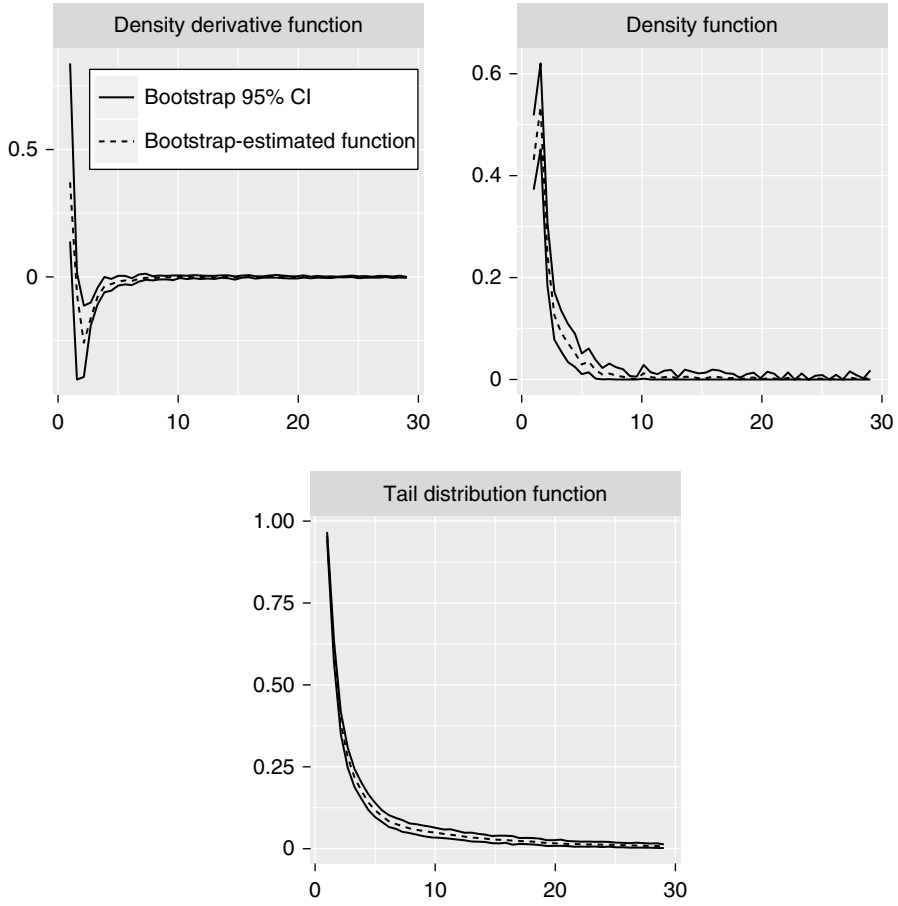
Both our worst-case approach and GPD fitting require choosing a threshold parameter. In GPD fitting, it is important to choose a threshold parameter high enough so that the GPD becomes a valid model. GPD fitting, however, is difficult for a small data set when the lack of data prohibits choosing a high threshold. On the other hand, the threshold in our worst-case approach can be chosen much higher, because our method relies on the data below or close to the threshold, but not those far above it.

7.3. Fire Insurance Data: Example 1 Revisited

Consider the fire insurance data in Example 1. The quantity of interest is the expected payoff of a high-excess policy with reinsurance, given by $h(x) = (x - 50)I(50 \leqslant x < 200) + 150I(x \geqslant 200)$. The data set has only seven observations above 50.

We apply our worst-case approach to estimate an upper bound for the expected payoff by using $a = 29.03$, the cutoff above which 15 observations are available. Similar to Section 7.2, we use the bootstrapped KDE to

Figure 13. Bootstrapped Kernel Estimation of the Distribution, Density, and Density Derivative for the Danish Fire Losses Data in Example 1



obtain CIs for β, η, and ν. The estimates in Figure 13 appear to be very stable for this example, thanks to the relatively large data size.

We run Algorithm 2 and obtain a 95% confidence upper bound of 1.99. For comparison, we fit a GPD using threshold $u = 10$, which follows McNeil (1997) as the choice that roughly balances the bias-variance trade-off. The 95% CI from GPD fit is $[-0.03, 0.23]$. Thus, the worst-case approach gives an upper bound that is one order of magnitude higher, a finding that resonates with that in Section 7.2. Our recommendation is that a modeler who cares only about the order of magnitude would be better off choosing GPD, whereas a more risk-averse modeler who wants a bound on the risk quantity with high probability guarantee would be better off choosing the worst-case approach.

8. Conclusion

This paper proposed a worst-case, nonparametric approach to bound tail quantities based on the tail convexity assumption. The approach relied on an optimization formulated over all possible tail densities. We characterized the optimality structure of this infinite-dimensional optimization problem by developing an equivalence to a moment-constrained problem. Under an additional quasi-concavity condition on the objective function, we constructed the numerical solution scheme by converting it into low-dimensional nonlinear programs. With the presence of data, this approach tractably generated statistically valid bounds via suitable relaxations of the optimization that took into account the estimation errors of the required parameters. We compared our proposed approach to existing tail-fitting techniques, and demonstrated its relative strength of outputting correct tail estimates under data-deficient environments. We also examined the level of conservativeness of our bounds, which was viewed as a limitation of the proposed approach.

We suggest two extensions of our research. First is to generalize the proposed method to multivariate distributions, perhaps through separate modeling on the marginal distributions and the dependency structure. Second is to study means to reduce the level of conservativeness. This can involve mathematical transformations of the variable and the addition of extra information (e.g., other constraints).

Acknowledgments

The authors thank the area editor Bert Zwart, the associate editor and the two referees for many valuable suggestions that have greatly improved the paper.

References

Balkema AA, De Haan L (1974) Residual life time at great age. *Ann. Probab.* 2(5):792–804.
Beirlant J, Teugels JL (1992) Modeling large claims in non-life insurance. *Insurance: Math. Econom.* 11(1):17–29.
Ben-Tal A, Den Hertog D, De Waegenaere A, Melenberg B, Rennen G (2013) Robust solutions of optimization problems affected by uncertain probabilities. *Management Sci.* 59(2):341–357.
Bertsimas D, Popescu I (2005) Optimal inequalities in probability theory: A convex optimization approach. *SIAM J. Optim.* 15(3): 780–804.
Birge JR, Dulá JH (1991) Bounding separable recourse functions with limited distribution information. *Ann. Oper. Res.* 30(1): 277–298.
Birge JR, Wets RJB (1987) Computing bounds for stochastic programming problems by means of a generalized moment problem. *Math. Oper. Res.* 12(1):149–162.
Cule M, Samworth R, Stewart M (2010) Maximum likelihood estimation of a multi-dimensional log-concave density. *J. Roy. Statist. Soc.: Ser. B* (*Statist. Methodology*) 72(5):545–607.
Davis R, Resnick S (1984) Tail estimates motivated by extreme value theory. *Ann. Statist.* 12(4):1467–1487.
Davison AC, Smith RL (1990) Models for exceedances over high thresholds. *J. Roy. Statist. Soc. Ser. B* (*Methodological*) 52(3): 393–442.
Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.* 58(3):595–612.
Nilim A, El Ghaoui L (2005) Robust control of Markov decision processes with uncertain transition matrices. *Oper. Res.* 53(5): 780–798.
McNeil A, Frey R, Embrechts P (2005) *Quantitative Risk Management* (Princeton University Press, Princeton, NJ).
Embrechts P, Klüppelberg C, Mikosch T (1997) *Modelling Extremal Events*, Vol. 33 (Springer, Berlin).
Fisher RA, Tippett LHC (1928) Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Math. Proc. Cambridge Philosophical Soc.* 24(2):180–190.
Föllmer H, Schied A (2011) *Stochastic Finance: An Introduction in Discrete Time* (Walter de Gruyter, Berlin).
Glasserman P, Li J (2005) Importance sampling for portfolio credit risk. *Management Sci.* 51(11):1643–1656.
Glasserman P, Kang W, Shahabuddin P (2007) Large deviations in multifactor portfolio credit risk. *Math. Finance* 17(3):345–379.
Glasserman P, Kang W, Shahabuddin P (2008) Fast simulation of multifactor portfolio credit risk. *Oper. Res.* 56(5):1200–1217.
Gnedenko B (1943) Sur la distribution limite du terme maximum d'une serie aleatoire. *Ann. Math.* 44(3):423–453.
Goh J, Sim M (2010) Distributionally robust optimization and its tractable approximations. *Oper. Res.* 58(4-part-1):902–917.
Gumbel EJ (2012) *Statistics of Extremes* (Dover Publications, Mineola, NY).
Hannah LA, Dunson DB (2013) Multivariate convex regression with adaptive partitioning. *J. Machine Learn. Res.* 14(1):3261–3294.
Hansen LP, Sargent TJ (2008) *Robustness* (Princeton University Press, Princeton, NJ).
Heidelberger P (1995) Fast simulation of rare events in queueing and reliability models. *ACM Trans. Modeling Comput. Simulation* (*TOMACS*) 5(1):43–85.
Hill BM (1975) A simple general approach to inference about the tail of a distribution. *Ann. Statist.* 3(5):1163–1174.
Hogg RV, Klugman SA (2009) *Loss Distributions*, Vol. 249 (John Wiley & Sons, Hoboken, NJ).
Hosking JR, Wallis JR (1987) Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics* 29(3):339–349.
Iyengar GN (2005) Robust dynamic programming. *Math. Oper. Res.* 30(2):257–280.
Karr AF (1983) Extreme points of certain sets of probability measures, with applications. *Math. Oper. Res.* 8(1):74–85.
Koenker R, Mizera I (2010) Quasi-concave density estimation. *Ann. Statist.* 38(5):2998–3027.
Lim E, Glynn PW (2012) Consistency of multidimensional convex regression. *Oper. Res.* 60(1):196–208.
McNeil AJ (1997) Estimating the tails of loss severity distributions using extreme value theory. *ASTIN Bulletin: J. Internat. Actuarial Assoc.* 27(1):117–137.
Nicola VF, Nakayama MK, Heidelberger P, Goyal A (1993) Fast simulation of highly dependable systems with general failure and repair processes. *IEEE Trans. Comput.* 42(12):1440–1452.
Petersen IR, James MR, Dupuis P (2000) Minimax optimal control of stochastic uncertain systems with relative entropy constraints. *IEEE Trans. Automatic Control* 45(3):398–412.
Pickands J III (1975) Statistical inference using extreme order statistics. *Ann. Statist.* 3(1):119–131.
Popescu I (2005) A semidefinite programming approach to optimal-moment bounds for convex classes of distributions. *Math. Oper. Res.* 30(3):632–657.
Rockafellar RT (1997) *Convex Analysis* (Princeton University Press, Princeton, NJ).
Seijo E, Sen B, et al. (2011) Nonparametric least squares estimation of a multivariate convex regression function. *Ann. Statist.* 39(3):1633–1657.
Seregin A, Wellner JA (2010) Nonparametric estimation of multivariate convex-transformed densities. *Ann. Statist.* 38(6):3751–3781.
Smith JE (1995) Generalized Chebychev inequalities: Theory and applications in decision analysis. *Oper. Res.* 43(5):807–825.
Smith RL (1985) Maximum likelihood estimation in a class of nonregular cases. *Biometrika* 72(1):67–90.
Smith RL (1987) Estimating tails of probability distributions. *Ann. Statist.* 1174–1207.
Talluri KT, Van Ryzin GJ (2006) *The Theory and Practice of Revenue Management*, Vol. 68 (Springer, New York).
Winkler G (1988) Extreme points of moment sets. *Math. Oper. Res.* 13(4):581–587.

Henry Lam is an associate professor in the Department of Industrial Engineering and Operations Research at Columbia University. His research focuses on Monte Carlo simulation, risk and uncertainty quantification, and stochastic optimization. His work has been recognized by the National Science Foundation CAREER Award, the INFORMS Junior Faculty Interest Group Paper Competition Second Prize, the INFORMS George Nicholson Paper Competition Honorable Mention Prize, and the Adobe Faculty Research Award.

Clementine Mottet is a PhD student in the Department of Mathematics and Statistics at Boston University. She joined their PhD program in 2013 after graduating from the National Institute of Applied Sciences of Toulouse with a Master's Degree in Applied Mathematics. She is currently conducting research under the supervision of Professor Lam. Their collaborative research has been focusing on robust estimation of tail quantities under limited information, which is relevant in fields related to risk management.