NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

DUAL APPROACH TO SUPERQUANTILE ESTIMATION AND APPLICATIONS TO DENSITY FITTING

by

John J. Sabol III

June 2016

Thesis Advisor:	Johannes O. Royset
Second Reader:	Samuel E. Buttrey

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704–0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202–4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.

1. AGENCY USE ONLY *(Leave Blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	June 2016	Master's Thesis 07-07-2014 to 06-17-2016

4. TITLE AND SUBTITLE

DUAL APPROACH TO SUPERQUANTILE ESTIMATION AND APPLICATIONS TO DENSITY FITTING

5. FUNDING NUMBERS

6. AUTHOR(S)

John J. Sabol III

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Postgraduate School
Monterey, CA 93943

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

N/A

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

The views expressed in this document are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol Number: N/A.

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(maximum 200 words)*

Analysts often concern themselves with the tail regions of distributions, sometimes called "extreme events," in order to measure or predict risk. One risk metric, the superquantile, possesses several properties that make it particularly well-suited for risk quantification. Observable data, however, often lack information on extreme events due to various resource constraints, resulting in sample superquantile estimates that often undervalue the true level of risk. By leveraging the dual relationship between superquantiles and superexpectations, we apply constrained optimization on second-order epi-splines to arrive at incrementally better approximations of superquantile values. With these improved estimates, we incorporate additional constraints to improve the fidelity of density estimates in tail regions. We limit our investigation to data with heavy tails, where risk quantification is typically the most difficult. Demonstrations are provided in the form of a known distributional benchmark, historical financial data, and a fluid dynamics model used in the development of a high-speed naval vessel. Results show that accurate quantile and superquantile constraint implementation, in conjunction with empirical statistics and distributional knowledge, can improve tail density estimates by up to 15% for small samples of various heavy-tailed distributions.

14. SUBJECT TERMS

probability density estimation, epi-splines, optimization, risk quantification, superquantiles, non-parametric statistics

15. NUMBER OF PAGES 85

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2–89)
Prescribed by ANSI Std. 239–18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

DUAL APPROACH TO SUPERQUANTILE ESTIMATION AND APPLICATIONS TO DENSITY FITTING

John J. Sabol III
Captain, United States Marine Corps
B.S., University of Notre Dame, 2010

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
June 2016

Approved by: Johannes O. Royset, Ph.D.
Thesis Advisor

Samuel E. Buttrey, Ph.D.
Second Reader

Patricia A. Jacobs, Ph.D.
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Analysts often concern themselves with the tail regions of distributions, sometimes called "extreme events," in order to measure or predict risk. One risk metric, the superquantile, possesses several properties that make it particularly well-suited for risk quantification. Observable data, however, often lack information on extreme events due to various resource constraints, resulting in sample superquantile estimates that often undervalue the true level of risk. By leveraging the dual relationship between superquantiles and superexpectations, we apply constrained optimization on second-order epi-splines to arrive at incrementally better approximations of superquantile values. With these improved estimates, we incorporate additional constraints to improve the fidelity of density estimates in tail regions. We limit our investigation to data with heavy tails, where risk quantification is typically the most difficult. Demonstrations are provided in the form of a known distributional benchmark, historical financial data, and a fluid dynamics model used in the development of a high-speed naval vessel. Results show that accurate quantile and superquantile constraint implementation, in conjunction with empirical statistics and distributional knowledge, can improve tail density estimates by up to 15% for small samples of various heavy-tailed distributions.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

List of Figures

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

AAD	Average Absolute Deviation
CDF	Cumulative Distribution Function
DSE	Dual of Superexpectation
GAMS	General Algebraic Modeling System
IID	Independent and Identically Distributed
LCB	Lower Confidence Bound
MAD	Median Absolute Deviation
MEP	Maximum Entropy Problem
MLE	Maximum Likelihood Estimation
MLP	Maximum Likelihood Problem
NPS	Naval Postgraduate School
PDF	Probability Density Function
SSE	Sum of Squared Errors
SSTE	Sum of Squared Tail Errors
SVM	Support Vector Machines
UCB	Upper Confidence Bound

THIS PAGE INTENTIONALLY LEFT BLANK

Executive Summary

Experimental data collection is often limited in the number of observations due to budgetary constraints, computational complexity, or time. In these cases, samples often lack the "extreme events" that constitute the tail regions of the underlying probability density function. In order to assess and quantify risk, analysts must often extrapolate probability estimates into regions beyond the range of observable data. This risk assessment is particularly important for distributions that exhibit heavy-tailed behavior, where extreme events occur more frequently.

Our goal is to provide a set of accurate risk quantification constraints for limited data that enhance the fidelity of probability density estimates in outlying tail regions. Due to its desirable qualities as a risk metric, we use the superquantile at a given percentile to help inform tail density estimates. By incorporating an accurate superquantile estimate (or set of estimates) into a constrained optimization problem using first- and second-order epi-splines, we allow for the incorporation of a flexible combination of empirical knowledge and distributional soft information that can be uniquely tailored to the application in question.

This thesis proposes a two-step process that first approximates superquantile values before using them within an overall density estimation framework. Superquantiles are estimated using constrained second-order epi-spline optimization that leverages the dual relationship between superquantiles and superexpectations as well as additional soft information to help inform shape constraints. We investigate the impact of distributional knowledge on both quantile and superquantile predictions, demonstrating how enhanced constraint formulations arrive at incrementally better estimates. Once obtained, these quantile and superquantile estimates are incorporated into a second constrained optimization problem that estimates the underlying density function using first-order epi-splines. Within this second phase, we explore optimizations under both maximum likelihood and maximum entropy formulations.

The method posited here is evaluated across three cases. In the first case we assess the method against two well-defined parametric benchmarks, demonstrating and measuring the impact of additional information. Next, we apply the method to a heavy-tailed set of financial data, comparing our results with those attained via a widely used existing method.

Finally, we investigate a multi-fidelity data set from a fluid dynamics model to assess the feasibility of surrogacy for quantile estimates and the benefits of mixture modeling in risk quantification.

Results show that accurate quantile and superquantile predictions are key to successful constraint implementation for overall density estimation. Given accurate estimates for outlying quantiles and superquantiles, overall density estimation improves by up to 15%, with tail density estimates improving by up to 80% in the case of maximum entropy formulations. The impact here is improved accuracy in tail density estimation, providing better risk mitigation planning for "extreme events," and without the need for large sample sizes.

CHAPTER 1:
Background

The world is filled with uncertainty — a lack of total knowledge that forces us to make decisions with access to only a fraction of the complete picture. How much money should a city allocate in the budget to cover the recovery costs of the next category-five hurricane? How confident are engineers that the tensile strength of the steel cables they tested will live up to the design specifications for a new bridge? In attempting to answer these questions, risk analysts rely heavily on data to provide accurate estimates or bounds on the likelihood of an event, and their findings represent a significant and growing area of research in numerous fields including insurance, financial management, and reliability engineering. The question, however, of how to estimate the probability of events that exist beyond the range of observed data requires extrapolation that often leads to inaccurate estimates for the probability of extreme events. Within this thesis, we look at a specific risk measure (*superquantiles*) to help inform probability density estimates for rare or extreme events coming from an unknown distribution with access to limited samples of data.

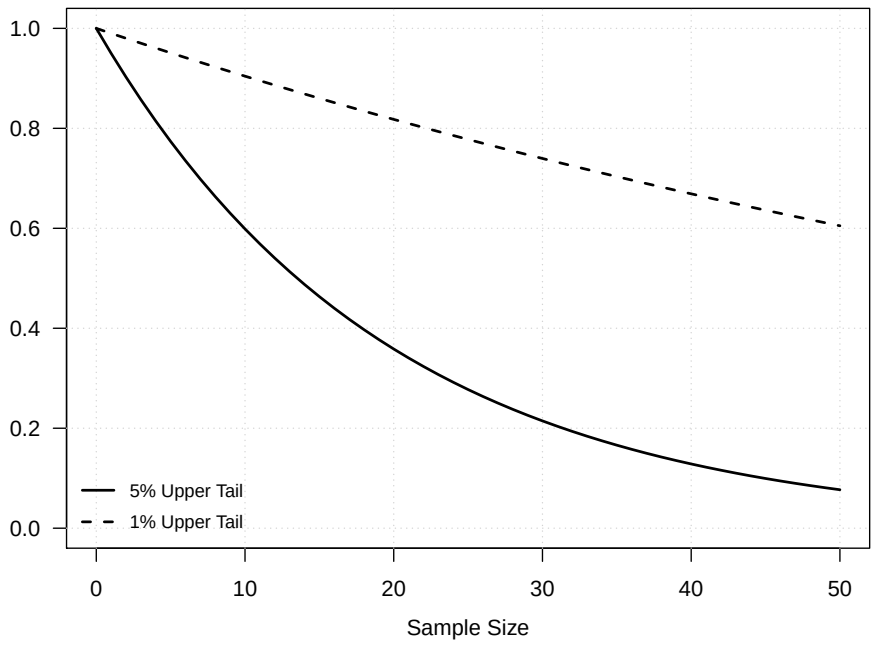
1.1 Probability Density Estimation

Probability density estimation is the construction of probability density functions (PDFs) from an observed set (or sets) of data for which the true underlying distribution is unknown. By estimating the underlying density of a random variable, we hope to predict future observations according to some probabilistic model, where predictions become more accurate as the fidelity of the model improves. Methods for density estimation include both parametric methods such as the maximum likelihood estimation (MLE) formulations proposed by Fisher between 1912 and 1922 [1], as well as non-parametric techniques such as kernel smoothing [2], support vector machines (SVM) [3], and quasi-MLE for unspecified models [4]. While parametric techniques can be quite powerful even with limited sample sizes, they are highly dependent on the assumption of the density's underlying parametric structure. If one assumes an invalid distributional family, or if the data itself is non-parametric in nature, these methods can result in grossly inaccurate predictions.

Nonparametric techniques provide more flexibility in modeling data, but can often

rely heavily on sample size to arrive at accurate density estimates. In theory, if one were afforded millions of observations, a nonparametric density estimate would perform quite well, even for wildly nonparametric distributions. When provided only limited samples, however, nonparametric techniques can perform poorly in estimating densities for outlying tail regions where observations are likely to be absent. In fact, in a sample of 30 independent and identically distributed (IID) observations, there is a roughly 20% chance the data will lack an observation coming from the 5% upper tail region, as shown in Figure 1.1.

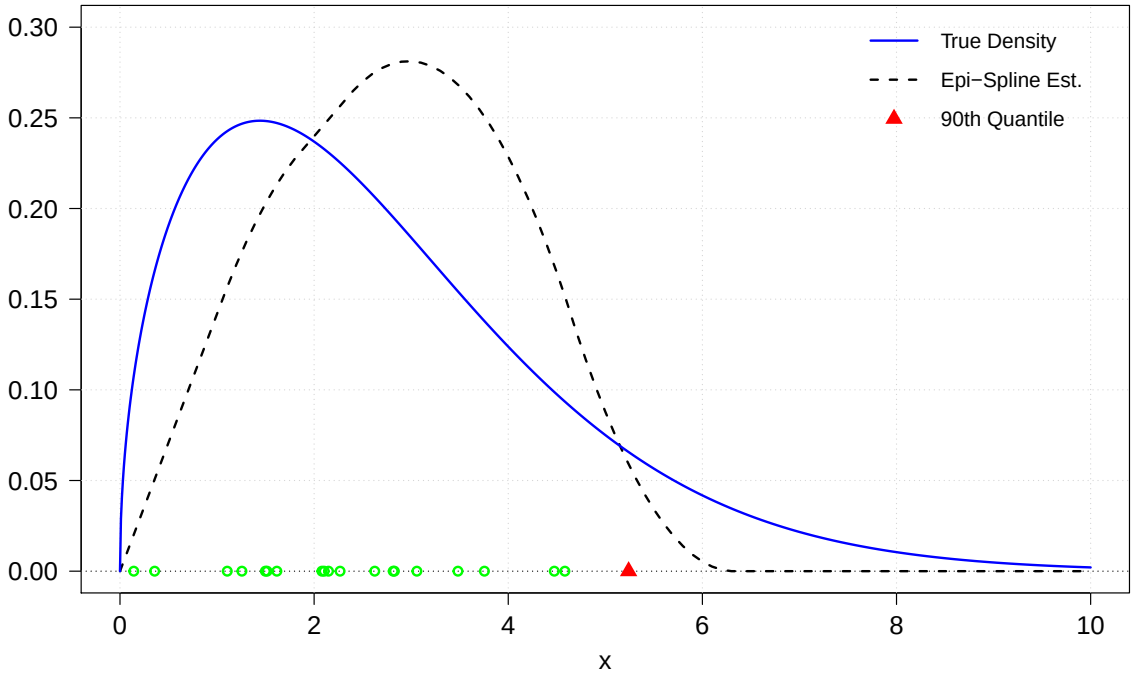
Figure 1.1: Upper Tail Data Inclusion by Sample Size



Probability of not seeing any values within the 5% (solid) and 1% (dotted) upper tail regions for IID observations. Clearly, sample size is of great consequence in estimating tail density.

Inaccurate estimates on tail densities become especially problematic when one attempts to answer questions regarding the uncertainty or risk of extreme events, such as the probability of falling above or below some threshold value. Given realistic constraints on resources, analysts are often forced to confront limited data sets possessing few or no observations from the extreme tail regions. This shortfall is especially prominent in densities where extreme events are more likely to occur, a characteristic referred to as *heavy-tailed*. An example of this is offered in Figure 1.2, which displays a prototypical underestimation of tail density from a well-known heavy-tailed distribution.

Figure 1.2: Example of Tail Density Underestimation



PDF estimate for 20 observations coming from a Weibull distribution with shape parameter $k = 1.5$ and scale parameter $\lambda = 3$. Optimization of first-order epi-splines was performed on the basis of maximizing log-likelihood subject to various constraints such as monotonicity and tail convexity. Notice that no data exists beyond the 90th quantile. As a result, the estimate fails to provide sufficient density in the tail region.

1.2 Tail Density Characterizations

The risk implications for tail estimation can depend greatly on the distribution's tail behavior. Take for instance a normal random variable, $X \sim N(\mu, \sigma)$. We know by virtue of its light tails that 99.7% of the data is bounded within three standard deviations of the mean. Now compare this to a Cauchy random variable (known to have a heavy tail) with an undefined variance, and the prospect of bounding the uncertainty becomes much more difficult. In fact, even something as simple as estimating the mean for a heavy-tailed distribution can be quite difficult, as explained in [5]. In cases where the underlying distribution exhibits behavior that can be characterized as *heavy-tailed*, the consequences of inaccurate tail estimation become of greater consequence. Given our specific interest in such distributions, we limit the remainder of our analysis to those distributions and data sets for which we know, or suspect, heavy-tailed behavior.

1.2.1 Heavy Tails

Heavy-tailed distributions have tails which are not exponentially bounded, with common examples including the Pareto, Burr, and Cauchy distributions. Since the tail of an exponential distribution is (by definition) exponentially bounded, these can serve as illustrative comparisons to demonstrate characteristics of heavy-tails. Defined more explicitly, a distribution is characterized as heavy-tailed if

$$\lim_{x\to\infty} e^{\lambda x} Pr[X > x] = \infty \quad \forall \lambda > 0. \tag{1.1}$$

Heavy-tails can be further defined as one of three general types: fat-tailed, long-tailed, or sub-exponential. Specialized algorithms for simulating heavy-tailed distributions can be found in [6], with more detailed methodologies proposed for handling specific cases. Here, we explore the particular characteristics of fat-tailed distributions.

1.2.2 Fat Tails

Fat-tailed distributions are a sub-class of heavy-tailed distributions for which the probability density function goes to zero as a power of a for large x. As such, they are always bounded below by the probability density of an exponential distribution. Mathematically, a distribution is said to have a fat tail if

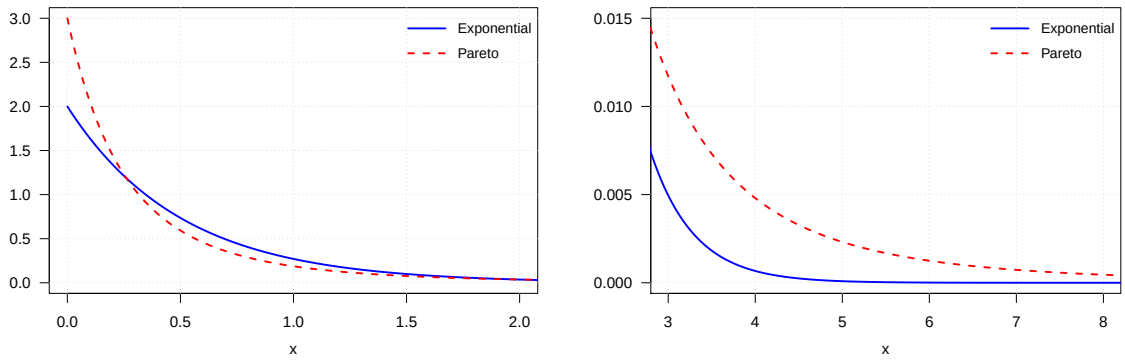
$$\Pr[X > x] \sim x^{-a} \quad \text{as } x \to \infty, \quad a > 0, \tag{1.2}$$

where $\sim$ refers to the asymptotic equivalence of functions. Fat tails are frequently used to model financial and economic applications where extreme events must be accounted for, such as the insurance industry and commodity markets [7]–[9]. Another application is in modeling natural disasters [10], where the question posed earlier regarding the category-five hurricane serves as an illustrative example. A visual comparison between a fat tail and an exponential case is provided in Figure 1.3.

1.3 Superquantiles as a Risk Measure

The methods by which we interpret, measure, and communicate uncertainty forms the basis for risk quantification. The probability of exceeding a threshold value (as mentioned

Figure 1.3: Comparison of Heavy vs. Exponentially Bounded Tails



(a) Density Functions (b) Tail Region

PDF plots with an expanded view of the tail region (left) for an exponential (solid blue) and Pareto (dotted red) distribution possessing equal expected values. Notice the lower bounding performed by the exponential case in comparison to the fat-tailed Pareto distribution in the tail region.

earlier) is one such risk measure, which in this case uses a particular quantile value as the risk measure. Quantiles, however, are just one of many such risk measures that can be used. In fact, a quantitative evaluation of risk relies on a measure that properly relates to the specific application in question, where certain desirable qualities can make a measure more or less useful and appropriate for a given circumstance [11]. Of particular note are *superquantiles*, which possess several features that make them particularly well-suited as risk measures. Known more commonly as "conditional value-at-risk," "average value-at-risk," "tail value-at-risk," and "expected shortfall" from their applications in financial analysis, many analysts have come to recognize superquantiles for their desirable properties of coherency and regularity in assessing risk under incomplete or inaccurate probabilistic models [11], [12]. The analysis of superquantiles originated in financial engineering where risk lies predominantly in the lower tail of the underlying distribution. We modify this convention for our purposes here to assess the upper tails instead, realizing that the methods are equally valid with the addition of a simple sign change on the data of interest.

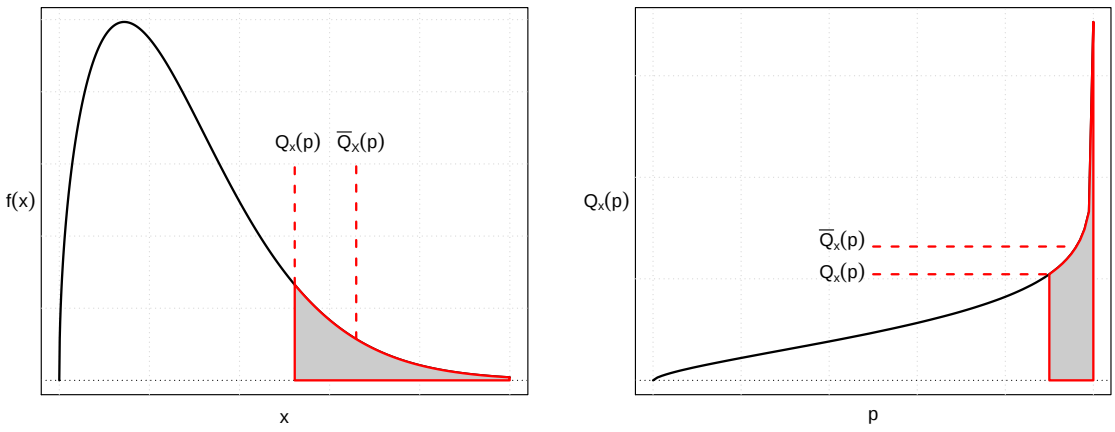
Given a random variable X, we can calculate the quantile of X at a certain percentile p as the threshold value below which proportion p of the underlying density function exists. If X is continuous, then the quantile function $Q(p)$ is simply the inverse of the cumulative distribution function (CDF). A *super*-quantile is (generally) an average of quantiles for

percentiles above a given threshold. When the CDF of X is continuous, the superquantile becomes the conditional expectation of X above the given quantile threshold value [11]. Put another way, a superquantile is the expected outcome provided you are known to be operating in the top p-tail of the underlying probability distribution. Visual depictions of these relations between superquantiles and the density and quantile functions are provided in Figure 1.4. Mathematically, we define the superquantile $\bar{Q}_X$ of the distribution as a function of a specific percentile p, where $\bar{Q}_X(p)$ equals the expectation in the upper p-tail distribution of X, for $p \in (0, 1)$. As such, the superquantile function becomes

$$\bar{Q}_X(p) = \frac{1}{1-p} \int_p^1 Q_X(p')dp' \quad \text{for} \quad p \in (0, 1). \tag{1.3}$$

Provided the knowledge of a distribution's superquantile at a given percentile, we seek to leverage this information in order to make better estimates of that distribution's tail density. By arriving at better estimates of superquantile values for p in ranges closer to one, we hope to more accurately assess the probability of exceeding a particular risk threshold.

Figure 1.4: Quantile and Superquantile Relationship



(a) Density Function (b) Quantile Function

PDF and quantile plots for a Weibull distribution. Quantile and superquantile values are identified by the dotted red lines, with the upper tail region shaded. With continuity in the density, the superquantile value is the expected value of this upper tail region.

1.4 Superexpectations and their Dual

Related to the superquantile function, the superexpectation function E_X of a random variable X is the expected value of the maximum of X and some threshold value x, which we term the level. Thus, the value $E_X(x)$ is the superexpectation of X at level x defined by

$$E_X(x) = E[\max\{x, X\}] = \int_0^1 \max\{x, Q_X(p)\}dp. \tag{1.4}$$

If we require $E|X| < \infty$, then the superexpectation function has the properties of being closed, proper, and convex. Using the Legendre-Fenchel transform, we are able to arrive at a conjugate ("dual") function of the superexpectation which is also closed, proper, and convex [11]. We refer to this as the dual of superexpectations (DSE) $E_X^*(p)$, here denoted with a star, and defined as a function of p such that

$$E_X^*(p) = \begin{cases} -(1-p)\bar{Q}_X(p) & \text{for } p \in (0,1), \\ -E[X] & \text{for } p = 0, \\ 0 & \text{for } p = 1, \\ \infty & \text{for } p \notin [0,1]. \end{cases} \tag{1.5}$$

1.4.1 Relation of Dual of Superexpectation to Superquantile

Of importance is the relationship of the DSE to superquantiles, where one can simply be written as a function of the other on the interval $p \in (0, 1)$, as in Equation 1.5. Furthermore, we note the known endpoints for cases where $p = 0$ and $p = 1$, as well as the fact that the left-derivative of E_X^* at a percentile p is in fact the underlying density function's quantile value at that same p, as can be seen in Figure 1.5 [11]. That is,

$$\frac{d}{dp}E_X^*(p) = Q_X(p). \tag{1.6}$$

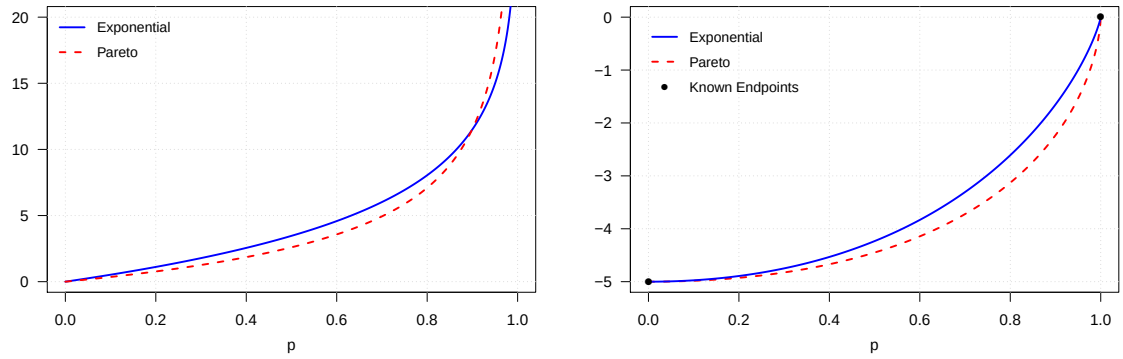
Additionally, if X possesses a finite second moment, there exist upper and lower bounds on superquantile estimates for values between the endpoints [11]. In this way, for $p \in [0, 1)$,

superquantiles are bounded according to

$$E[X] \quad \leq \quad \bar{Q}_X(p) \quad \leq \quad E[X] + \frac{\sigma(X)}{\sqrt{1-p}}, \tag{1.7}$$

provided that $E|X^2| < \infty$, and where $\sigma(X)$ is the standard deviation of X. This requirement is not extraneous, as several known heavy-tailed distributions, such as the Cauchy distribution mentioned earlier, do not in fact meet this criterion. Nonetheless, these characteristics provide the foundation for estimating the superquantile for a set of data, which underpins much of the methodology pursued here and outlined in Chapter 2.

Figure 1.5: Heavy-Tailed Quantile Characteristics



(a) Quantiles (b) DSE

A comparison of the quantile (left) and DSE (right) functions for the exponential distribution (blue solid) and Pareto (dotted red). Notice that the quantile function values correspond exactly to the slope of the DSE function. Furthermore, note the bounding provided by the exponential case in (b). This exponential upper-bounding is related to the lower-bounding it performs on the tail density seen before in Figure 1.4b.

1.5 First- and Second-Order Epi-Splines

Epi-splines are a flexible set of piecewise polynomial functions that can be used to describe practically any function one could reasonably expect to encounter. By modifying a set of constraints that define the problem either through empirical information from the data itself or from external *soft information* one can develop a framework that identifies the coefficient values that define the piecewise epi-spline function. Epi-splines are well suited to handling multiple shape-constraints and approximating density functions through the maximization

of either a likelihood function or entropy function, even for small sets of data, as shown in [13].

1.5.1 Epi-Splines Model Formulations

Given a closed interval $[l, u]$ constituting the bounds of estimation within $\mathbb{R}$, we segment the interval into evenly spaced mesh segments. Here, we will distinguish between a second-order epi-spline mesh k used for DSE estimation from a first-order mesh m used for density estimation. As such, we have K evenly spaced second-order mesh segments defining p according to $p = \{p^k \mid k = 1, 2, ..., K\}$, and M evenly spaced first-order mesh segments defining x according to $x = \{x^m \mid m = 1, 2, ..., M\}$. Mesh segments are right-continuous, with endpoints defined as either left $(p_L,\ x_L)$ or right $(p_R,\ x_R)$ respectively, as seen in Figure 1.6. Our epi-spline segments are thus defined either as first- or second-order polynomials according to their mesh assignment.

$$\text{Second Order:} \quad \hat{E}^*_X(p) = a_0^k + a_1^k p + a_2^k p^2 \qquad \text{for } p \in [p_L^k,\ p_R^k) \tag{1.8}$$

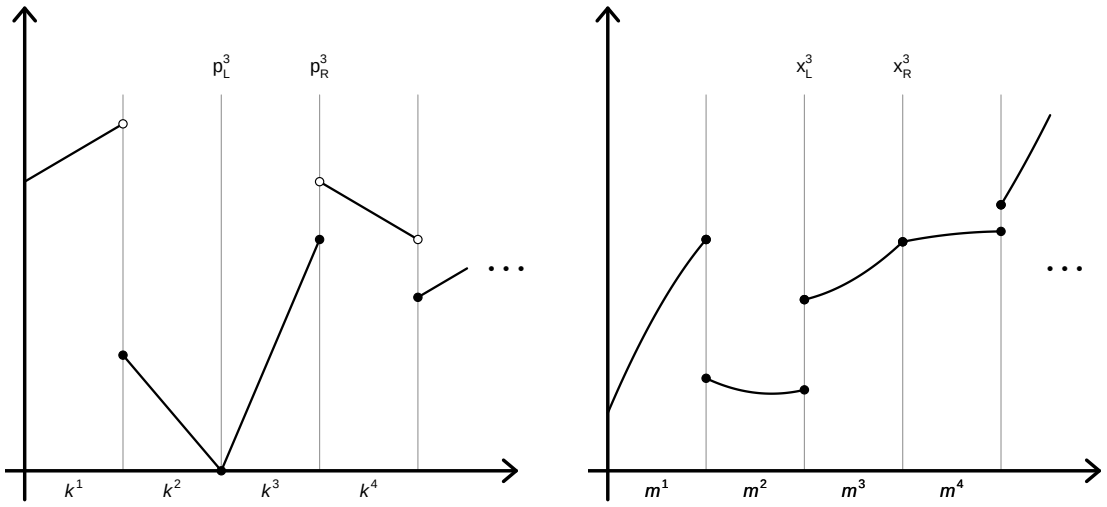
$$\text{First Order:} \quad \hat{f}(x) = b_0^m + b_1^m x \qquad \text{for } x \in [x_L^m,\ x_R^m) \tag{1.9}$$

To differentiate notation between formulations, we use a-coefficients for the second-order optimization defined via mesh index k, and b-coefficients for the first-order optimization defined via mesh index m. The details of epi-spline implementations are covered in Chapter 2.

1.6 Soft Information Incorporation

Shortfalls in sample size can often be partially mitigated through the incorporation of soft information that provides qualitative information on the underlying density function in addition to the empirical information provided by the data itself. Examples of soft information include density characteristics such as monotonicity, tail convexity, uni-modality, and many others. Combinations of soft information through constrained optimization leads to better density estimates for limited data sets when implemented in an intelligent framework, as shown in [14], [15].

Figure 1.6: First- and Second-Order Epi-Spline Meshes



(a) First-Order Epi-Splines

(b) Second-Order Epi-Splines

First- and second-order epi-splines with associated mesh. First-order epi-splines are on the left, with second-order epi-splines displayed on the right. Notice the lower semi-continuous (lcs) property enforced at mesh intersections. This will require continuity to be enforced throughout. By increasing the mesh resolution, we can arrive at arbitrarily close approximations to virtually any curve using these low-order splines.

1.7 Overview

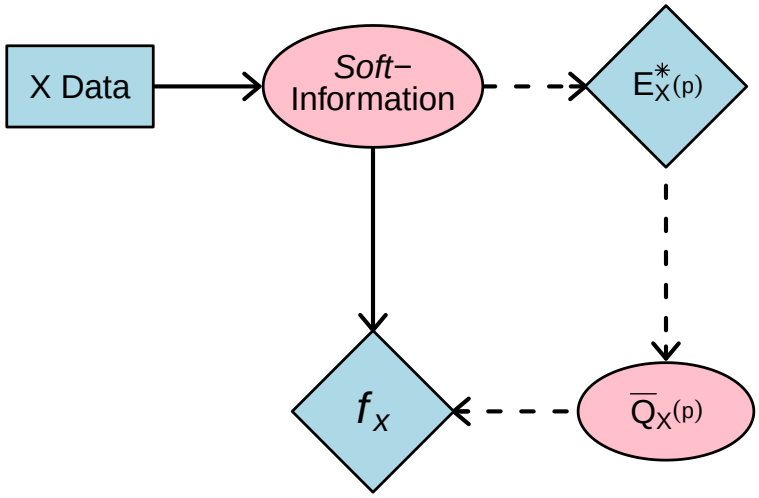
Expanding on earlier work done using epi-splines for density estimation, we plan to implement an additional set of constraints within the constrained optimization formulations to incorporate information on tail characteristics and superquantile bounds. In Chapter 2, we provide detailed steps for constrained optimization formulation for both superquantile estimation and density estimation, identifying objective functions and a set of possible constraints used. In Chapter 3, we estimate the superquantile values and density functions for two parametric distributions, exploring the impact of the superquantile constraint on overall tail density estimation by comparison to known benchmarks. Following this, Chapter 4 applies our method to samples from a financial data set known to possess heavy-tailed characteristics with comparisons made another commonly used density estimation method. Chapter 5 provides an application of the technique on a mixed data set of both high- and low-fidelity observations to show the method's use in a hierarchical model framework. Finally, Chapter 6 provides conclusions from our findings.

CHAPTER 2:
Methodology

2.1 Analytic Framework

We ultimately seek to estimate the underlying distribution of a data set for which limited observations exist. Epi-splines, shaped through multiple design constraints informed by both the data and additional soft information, and formulated to optimize either log-likelihood or entropy, provide the foundation of our approach. In general, the formulation follows the process outlined in the flow diagram in Figure 2.1.

Figure 2.1: Density Estimation Methodology



The generalized methodology for estimating probability density from a sample of data. Note the inclusion of both statistical information derived from the sample itself as well as soft information that can describe known tail behavior or other distributional characteristics. We hope that the additional steps shown via the dotted lines will help inform tail estimation.

We begin with an IID set of univariate sample data $x_1, ...x_n$ for which we wish to estimate the underlying density. From the data, we calculate sample statistics (such as mean, variance, median, and quartile values), as well as associated confidence intervals as per the methods outlined in section A.3 and section A.4. Second-order epi-spline segments that approximate the DSE of the underlying distribution are obtained using constrained

optimization subject to various constraints as outlined in section 2.3. The choice of second-order epi-splines allows us to leverage the rigidly defined shape constraints imposed on DSE functions without requiring an ultra-fine mesh. Additionally, as we will later see, the objective function for DSE approximation will require epi-splines of at least the second-order. From these epi-spline estimates, we acquire quantile and superquantile estimates for right-tailed p-values which we will subsequently incorporate into a second constrained optimization problem. This second optimization uses first-order epi-splines to approximate the underlying density function, using soft information to inform shape characteristics and optimized in the form of either a maximum log-likelihood problem (MLP) or a maximum entropy problem (MEP) as detailed in section 2.4. We elect first-order splines for this second optimization so as to enable a much finer mesh resolution subject to simpler constraint formulations. This choice helps to balance computational run time and formulation complexity, though higher-order epi-splines could certainly be utilized if desired.

2.2 Modeling Assumptions

We assume all data are IID from an unknown, but well-defined, density function with both finite mean and variance structure. The distribution need not be parametric, as epi-splines are well suited to nonparametric estimation, however, without a finite second moment on the underlying distribution, some methods posited here will have an improper application due to the attempts to estimate both mean and variance for use in select constraints within the optimizations. Furthermore, the data is assumed to possess no time dependence, or other ordering structure.

We further assume the data are free from measurement error or noise, making no attempt at deconvolution. If data is known or suspected of possessing Gaussian noise, application of a deconvolution constraint within the constrained density optimization can be applied as in [14].

2.3 Superquantile Estimation Formulation

We intend to show how the estimation of a desired superquantile can be arrived at from a constrained optimization of second order epi-splines. Using known characteristics of the DSE and its relation to superquantiles, we conduct our constrained optimization to arrive

at estimates of $E_X^*(p)$. We then convert these to estimates of $\bar{Q}_X(p)$ via the relationship outlined in Section 1.4.

Given that we know $E_X^*(p)$ exists only for $p \in [0, 1]$, we are able to freely define the length of our second-order epi-spline segments according to a desired mesh resolution. We refer back to subsection 1.5.1 in defining our epi-spline segments.

2.3.1 Dual of Superexpectation Objective

We desire conservative approximations for right-tailed densities such that our estimates for quantiles and superquantiles near $p = 1$ are *at least* as high as the actual values. To achieve this, we select an objective function that maximizes the curvature of the individual epi-spline segments, which we will term $\kappa(k)$, in order to achieve a shape similar to that seen in Figure 1.5b. In order to promote curvature over the entire mesh (rather that at a few individual segments) we apply an additional smoothing term that penalizes changes to a_2 between consecutive epi-splines. The degree of penalization is governed by a smoothing parameter ρ which we initialize at zero and increase as necessary to achieve smooth DSE estimates based on visual observation. Thus, given the second order of our epi-splines and our objective of maximizing curvature across the entire mesh, we arrive at an objective formulation defined by

$$\max_{a_0,\, a_1,\, a_2} \left\{ (1-\rho) \sum_k \kappa(k) - \rho \int \left(\frac{d^2}{dp^2} E_X^*(p) \right)^2 dp \right\}, \qquad \begin{array}{c} \text{Objective Function:} \\ \text{Dual of Superexpectations} \end{array}$$

$$\kappa(k) = \frac{2a_2^k}{\left(1 + \sqrt{a_1^k + 2a_2^k}\right)^{1.5}}, \qquad \text{Curvature Function}$$

$$\rho \int \left(\frac{d^2}{dp^2} E_X^*(p) \right)^2 dp = \rho \left(a_2^k - a_2^{k+1} \right)^2. \qquad \text{Smoother Penalty}$$

We note here that maximization of the sum of curvatures displayed a reasonable shape response for estimating $E_x^*(p)$, though other objective functions could reasonably be utilized. Values for the penalty parameter $\rho \approx 0.01 - 0.1$ were shown to work reasonably well when smoothing was required.

2.3.2 Constraints on $E_X^*(p)$

We impose constraints according to known criteria as well as external information as available. The following examines the significant constraints in formulations used here, though this does not constitute an all-inclusive list.

Convexity, Continuity, and Differentiability

From Theorem 1 of the characterization of superexpectations [11], we know E_X^* to be convex. We can therefore impose convexity on each epi-spline segment by requiring

$$a_2^k \geq 0 \qquad \forall k \in K. \tag{2.1}$$

This requirement further ensures that all curvature values will be positive, thus avoiding any potential cancellation across the sum of segments. If we further require continuity across the segments by tying mesh endpoints together and impose equal slopes at these intersections, we enforce convexity across the entire epi-spline function. This also has the added benefit of providing a smoother curve. Again, for p_R and p_L corresponding to the right and left ends of the mesh segment accordingly, we arrive at

$$a_0^k + a_1^k p_R^k + a_2^k (p_R^k)^2 = a_0^{k+1} + a_1^{k+1} p_L^{k+1} + a_2^{k+1} (p_L^{k+1})^2, \tag{2.2}$$

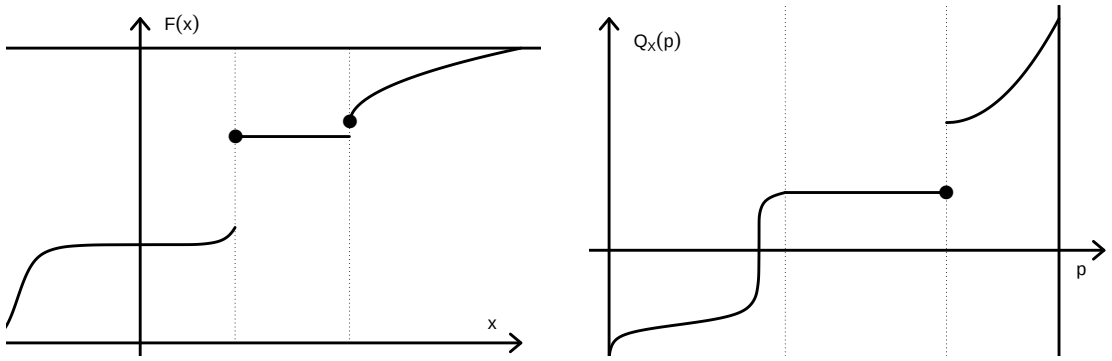
$$a_1^k + 2a_2^k p_R^k = a_1^{k+1} + 2a_2^{k+1} p_L^{k+1}. \tag{2.3}$$

Density Continuity

If we assume the underlying distribution to have a continuous CDF, we can require a_2 to be strictly positive across the mesh. This, again, relies on the DSE-quantile relationship of Equation 1.6 so that if the quantile is constant across an interval p, there is a vertical jump at that quantile value on the CDF, as demonstrated in Figure 2.2. As such, we can tighten our convexity constraint to a strict inequality so that

$$a_2^k > 0 \qquad \forall k \in K. \tag{2.4}$$

Figure 2.2: Relationship of CDF to Quantile Function



(a) CDF (b) Quantile Plot

CDF and quantile plots for an arbitrary, but non-continuous distribution. Notice that in areas where $Q_X(p)$ is constant there is a corresponding jump in the CDF that prohibits continuity.

Endpoints

From Equation 1.5 we know that $E^*_X(0) = -E[X]$ and $E^*_X(1) = 0$. Provided we may not know $E[X]$ exactly, we can approximate it from our sample data as either a point estimate or as a confidence interval on the sample mean $\bar{X}$. The Lower Confidence Bound (LCB) and Upper Confidence Bound (UCB) for $\bar{X}$ can be determined via a number of statistical techniques such as the student's T-distribution or bootstrap sampling depending on a desired confidence level. As such, we can impose a starting point constraint at $p = 0$ depending on our knowledge of X and the degree of flexibility we wish to provide our formulation as one of either

$$a_0^k = E[X], \tag{2.5}$$

$$\bar{X}_{LCB} \leq a_0^k \leq \bar{X}_{UCB}, \qquad \text{for } k = 1. \tag{2.6}$$

Additionally, the end point is always fixed, thus providing a constraint for the right-most epi-spline segment as

$$a_0^k + a_1^k + a_2^k = 0 \qquad \text{for } k = K. \tag{2.7}$$

Lower Bounds

As per Equation 1.7, we have a lower bound on $E_X^*(p)$ according to whether or not we have a known mean, known variance ($\sigma(X)$), or some combination thereof. We therefore impose a lower-bounding constraint as

$$a_0^k + a_2^k p + a_2^k p^2 \leq (p-1)\left(E[X] + \frac{\sigma_x}{\sqrt{1-p}}\right), \tag{2.8}$$

$$a_0^k + a_2^k p + a_2^k p^2 \leq (p-1)\left(\bar{X} + \frac{s}{\sqrt{1-p}}\right), \tag{2.9}$$

where s refers to the sample standard deviation. If $E[X]$ or $\sigma(X)$ are unknown, we can again estimate their values within some range of a confidence interval. Although there exists the corresponding upper bound for $E_x^*(p)$ as well, due to the requirement for convexity, it becomes a redundant constraint, and is therefore omitted.

Quantile Constraints

We recall from earlier that the slope of the DSE function is in fact the distribution's quantile function as per Equation 1.6. As such, if we know quantile values for X at specific p, we can enforce

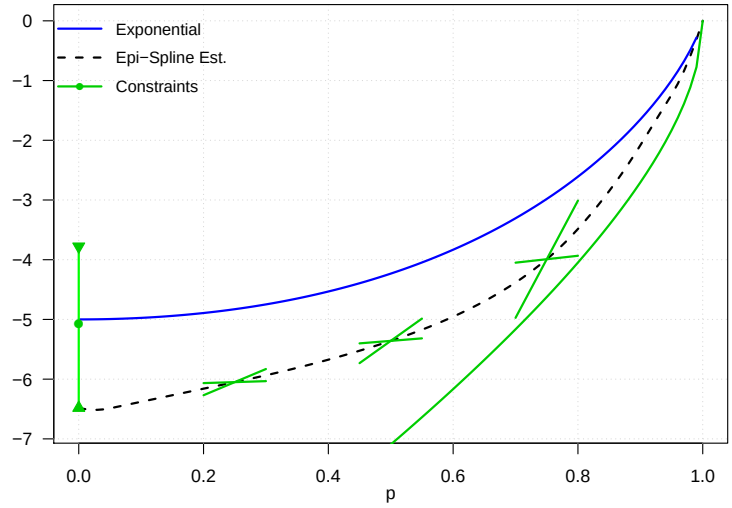
$$\frac{d}{dp}\hat{E}_X^*(p) = a_1^k + 2a_2^k p = Q_X(p) \qquad \text{for } p \in [p_L^k,\ p_R^k], \tag{2.10}$$

within each epi-spline segment containing such p. Furthermore, although mean and standard deviation are notoriously difficult to estimate for small, asymmetric distributions, quantiles are robust to outliers and can be estimated quite easily via binomial confidence intervals [16]. Even in relatively small samples, reasonable intervals for the median, 25th, and 75th quantiles ($p = 0.50$, 0.25, and 0.75 respectively) can be arrived at for modest confidence levels. Utilizing this non-parametric technique, we modify our constraint to include quantile lower and upper bound estimates $\hat{Q}(p)$ for any desired p as

$$\widehat{Q}_{LCB}(p) \leq a_1^k + 2a_2^k p \leq \widehat{Q}_{UCB}(p) \qquad \text{for } p \in [p_L^k,\ p_R^k]. \tag{2.11}$$

For the remainder of this thesis, we use the 25th, 50th, and 75th quantiles to bound epi-spline slopes for DSE estimation. An illustration of this constraint, as well as those for the end points and lower bounding, can be seen in Figure 2.3.

Figure 2.3: Visualization of Constraints for DSE Optimization



A graphical depiction of several constraints (green) imposed on the DSE optimization for an exponential random sample. The leftmost vertical confidence interval implies the approximation of expected value, while the three slope constraints seen at $p = 0.25,\ 0.50$, and 0.75 imply unknown but estimated quantile intervals. The lower bound constraint is informed through an estimation of both the sample mean and variance as per Equation 2.9.

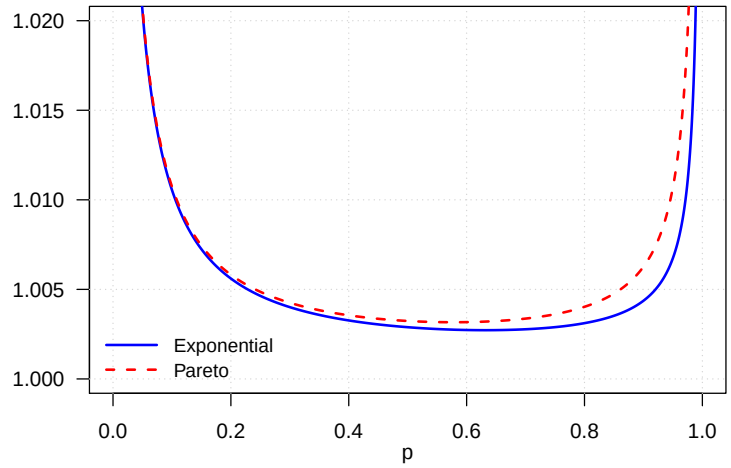
Tail Weight

If the distribution is known to be heavy-tailed, meaning the tail decays sub-exponentially, we can apply an additional constraint by leveraging knowledge of the exponential quantile function. As identified in subsection 1.4.1, we know that the left derivative of $E_X^*(p)$ is in fact the quantile value for the distribution. For the exponential case, we take the ratio of consecutive quantile values to obtain a relative rate of slope change for the epi-spline estimate. Heavy-tailed distributions will have a quantile function derivative that is bounded above by this change rate within the tail region, as exemplified in Figure 2.4. Thus, if we know or suspect heavy-tailed behavior, we enforce

$$\frac{Q(p_L^{k+1})}{Q(p_L^k)} = \frac{\log(1-p_R^k)}{\log(1-p_L^k)} \quad \leq \quad \frac{a_1^k + 2a_2^k p_R^k}{a_1^k + 2a_2^k p_L^k} \qquad \forall k \in \text{Tail Region},$$

$$\left(a_1^k + 2a_2^k p_L^k\right)\log(1-p_R^k) \quad \leq \quad \left(a_1^k + 2a_2^k p_R^k\right)\log(1-p_L^k). \tag{2.12}$$

Figure 2.4: Relative Gradient Change of Quantile Comparison



Ratio of quantile values for the exponential (solid blue) and Pareto (dotted red) densities possessing equal means. Note the lower bounding performed by the exponential on the fat-tailed Pareto, particularly in the tail region right of $p \approx 0.8$.

Minimal and Maximum Values

Since we know that the distribution's 0th quantile must be at most as small as the minimum observation, and that its 100th quantile must be at least as large as the maximum observation in the sample, we can develop constraints on the starting and ending slopes according to Equation 1.5, and implemented as

$$a_1^k + 2a_2^k \geq \max\{x_i\} \qquad \text{for } k = K, \tag{2.13}$$

$$a_1^k \leq \min\{x_i\} \qquad \text{for } k = 1. \tag{2.14}$$

Similarly, if we know the distribution's minimum or maximum value (or perhaps some reasonable bound on it), we can apply these equations as equalities or inequalities on the known value in the form of

$$a_1^k + 2a_2^k = \max\{f(x)\} \qquad \text{for } k = K, \tag{2.15}$$

$$a_1^k = \min\{f(x)\} \qquad \text{for } k = 1, \tag{2.16}$$

$$a_1^k + 2a_2^k \leq f(x)_{UCB} \qquad \text{for } k = K, \tag{2.17}$$

$$a_1^k \geq f(x)_{LCB} \qquad \text{for } k = 1. \tag{2.18}$$

Monotonicity

We leverage the DSE-quantile relation as per Equation 1.6 in order to enforce a characteristic of quantile gradients in regions of monotonicity. Within the underlying density, any region of monotonic increase or decrease will have a corresponding quantile gradient that is negative or positive respectively. Secondly, given the second-order of our epi-splines, we know that

$$\frac{d^2}{dp^2}E_X^* = Q_X'(p) = 2a_2^k \qquad \text{for } p \in [p_L^k,\ p_R^k]. \tag{2.19}$$

Because we segment our mesh into evenly spaced segments, we combine these two relations to form a DSE constraint on a_2 enforced over regions of known or suspected monotonicity. This relationship is depicted in Figure 2.5 and is applied as

$$a_2^k \geq a_2^{k+1} \qquad \forall k \in \text{Monotonically Increasing Region}, \tag{2.20}$$

$$a_2^k \geq a_2^{k-1} \qquad \forall k \in \text{Monotonically Decreasing Region}. \tag{2.21}$$

2.4 Density Estimation Formulation

In contrast to the situation with DSE estimation where endpoints for the mesh p were predefined on the interval $[0,\ 1]$, for density estimation we typically will not know the endpoints of the underlying distribution beforehand (if they are even finite). As such, our mesh x is defined by both a resolution M as well as endpoints l and u, as formulated in subsection 1.5.1.

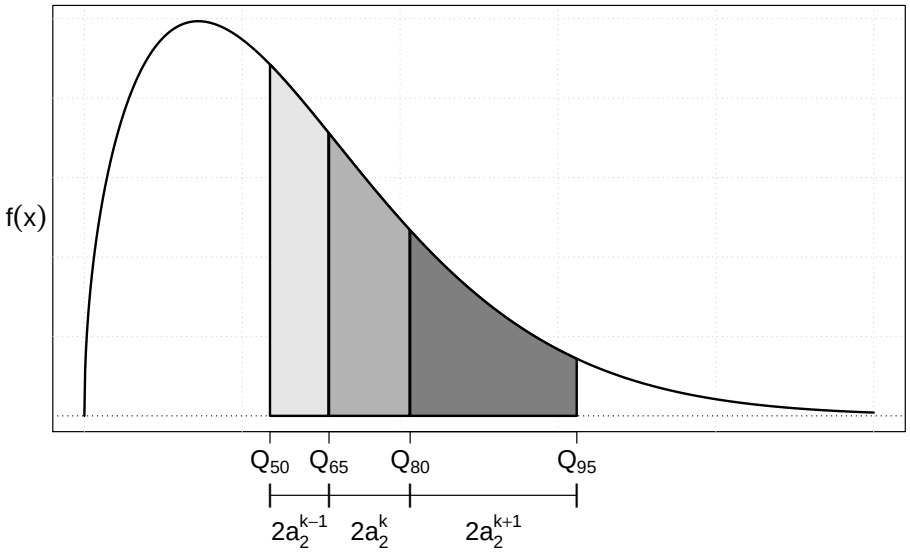
2.4.1 Density Objectives

Given our choice of objective function as outlined earlier, the general formulations for density become

$$\text{MLP:} \qquad \max_{\hat{f}} \sum_{i=1}^{n} \log\left(\hat{f}(x_i)\right) \qquad s.t. \quad \hat{f} \in F, \tag{2.22}$$

$$\text{MEP:} \qquad \max_{\hat{f}} \int -\hat{f}(x)\log\left(\hat{f}(x)\right) dx \qquad s.t. \quad \hat{f} \in F, \tag{2.23}$$

Figure 2.5: Quantile Gradient Relationship to Monotonicity



The Weibull distribution from before with several quantile values plotted. Since quantiles are shown for evenly spaced p segments (0.50, 0.65,...) the corresponding areas under the curve between them are equal (in this case 0.15). Since the PDF is monotonically decreasing in this region, basic geometry tells us that the distances between our quantiles must increase over the same region. This increase directly corresponds to our change in a_2^k across our epi-spline mesh.

where the set of constraints F are determined by properties of density functions (non-negative, integrate to one), as well as soft information such as continuity and shape constraints. With first-order epi-splines, we optimize our segment coefficients according to one of either

$$\text{MLP:} \quad = \max_{b_0,\, b_1} \sum_{i=1}^{n} \log(b_0^m + b_1^m x_i) \quad \forall i, m \mid x_i \in m, \tag{2.24}$$

$$\begin{aligned}\text{MEP:} \quad &= \max_{b_0,\, b_1} \int -(b_0^m + b_1^m x)\log(b_0^m + b_1^m x)dx \\ &= \max_{b_0,\, b_1} \sum_{m} \left[z_R^m \log(z_R^m) + z_L^m \log(z_L^m) + 2(z_R^m + z_L^m)\log\left(\frac{z_R^m + z_L^m}{2}\right)\right]. \end{aligned} \tag{2.25}$$

Here we use $z_R^m = b_0^m + b_1^m x_R^m$ and $z_L^m = b_0^m + b_1^m x_L^m$ as substitution variables, and make use of Simpson's Rule for approximating the integration contained in the MEP. Given that Simpson's Rule provides exact results for polynomials of degree three or lower, we implement it in later quantile and superquantile constraints to simplify terms.

2.4.2 Constraints on Density

We provide the minimal set of constraints that define a valid PDF, as well as those additional constraints dealing with quantile and superquantile estimates. Many other shape constraints, or soft information, such as monotonicity, unimodality, and convexity can be incorporated depending on one's knowledge of the underlying distribution. These methods are utilized throughout the remainder of this study, though the exact constraint derivations are omitted for the sake of brevity. Many of these derivations can be found in [14].

Unity

By definition, the density function must integrate to one. We can formulate this integrality constraint in linear form, where Δ_x is the mesh resolution $(u - l)/M$, shown here as

$$\sum_m \Delta_x \left[b_0^m + b_1^m \left(\frac{x_L^m + x_R^m}{2} \right) \right] = 1. \tag{2.26}$$

Continuous

We further assume densities explored here come from the family of continuous functions so that

$$b_0^m + b_1^m x_R^m = b_0^{m+1} + b_1^{m+1} x_L^{m+1} \qquad \forall m \mid m < M. \tag{2.27}$$

Non-Negative

Finally by definition of a density function, our estimates must contain all non-negative values such that

$$b_0^m + b_1^m x_L^m \geq 0 \qquad \forall m \in M, \tag{2.28}$$

$$b_0^m + b_1^m x_R^m \geq 0 \qquad \forall m \in M. \tag{2.29}$$

Quantile Constraints

If we know (or believe we have accurately estimated) the value of a particular quantile, we can, in much the same way that we formulate the unity constraint (Equation 2.26), require

that some percentage of density occupy the region left or right of the value itself. This works for not just quantiles estimated from the data itself, which we used to inform Equation 2.11, but also for quantile values closer to $p = 0$ or $p = 1$ by estimating the slope of the DSE function. In either case, quantile constraints for a specific p can be implemented as

$$\sum_{m|x_L^m \geq Q(p)} \Delta_x \left[b_0^m + b_1^m \left(\frac{x_R^m + x_L^m}{2} \right) \right] = 1 - p. \tag{2.30}$$

Superquantile Constraints

We wish to incorporate a constraint utilizing our estimate of the superquantile(s) derived from the previous optimization. Using the definition of superquantile outlined in Equation 1.3, and leveraging Simpson's Rule to simplify the expression, we arrive at a superquantile constraint for a particular quantile p as

$$\frac{1}{1-p} \sum_{m|x_L^m \geq Q(p)} \frac{\Delta_x}{6} h(x) = \bar{Q}(p) \tag{2.31}$$

$$h(x) = 3b_0^m (x_L^m + x_R^m) + b_1^m \left((x_R^m)^2 + (x_L^m)^2 \right) + b_1^m (x_R^m + x_L^m)^2 \tag{2.32}$$

which remains convex in b_0 and b_1. The corresponding quantile value $Q(p)$ can be estimated if unknown. One may also recall that since the superquantile at $p = 0$ is in fact the expected value, we see that the above formulation doubles as an expected value constraint for m summed across the entire mesh.

2.4.3 Quantile/Superquantile Constraint Implementation

In subsection 2.4.2, we formulated our quantile and superquantile constraints as equalities. Given the approximate nature of our quantile/superquantile values from DSE estimation, in practice we implement our constraints as either lower or upper bounds, the choice of which depends on the objective function for density (MLP vs. MEP) as well as any assumption made regarding a potential bias in estimates.

Typically, MLP formulations utilize lower bounding so as to "push" density into the tails, while MEP formulations use upper bounding to "rein in" density from the tails.

Conversely, if one believes the values attained though DSE approximation to overestimate the true superquantile values then perhaps the implementation of superquantile constraint as an upper bound may be more appropriate. When not otherwise stated, we assume that all MLP formulations incorporate expected value, quantile, and superquantile constraints via a lower-bounding of right-tailed densities, while MEP formulations use an upper-bounding of right tailed densities. With this in mind, we present the following quantile and superquantile constraints, where both $\widehat{Q}(p)$ and $\widehat{\bar{Q}}(p)$ are the results of DSE estimation in the previous step for quantiles and superquantiles respectively.

$$\sum_{m|x_L^m \geq \widehat{Q}(p)} \Delta_x \left[b_0^m + b_1^m \left(\frac{x_R^m + x_L^m}{2} \right) \right] \leq 1 - p \qquad \text{(Quantile Upper Bound)} \tag{2.33}$$

$$\sum_{m|x_L^m \geq \widehat{Q}(p)} \Delta_x \left[b_0^m + b_1^m \left(\frac{x_R^m + x_L^m}{2} \right) \right] \geq 1 - p \qquad \text{(Quantile Lower Bound)} \tag{2.34}$$

$$\frac{1}{1-p} \sum_{m|x_L^m \geq \widehat{Q}(p)} \frac{\Delta_x}{6} h(x) \leq \widehat{\bar{Q}}(p) \qquad \text{(Superquantile Upper Bound)} \tag{2.35}$$

$$\frac{1}{1-p} \sum_{m|x_L^m \geq \widehat{Q}(p)} \frac{\Delta_x}{6} h(x) \geq \widehat{\bar{Q}}(p) \qquad \text{(Superquantile Lower Bound)} \tag{2.36}$$

Given the possibility of inaccurate and perhaps even infeasible quantile and superquantile estimates, we further introduce a highly-penalized set of slack variables by which we elasticize the constraint bounds to ensure feasibility in the solution space. In nearly all cases, these slack variables remain zero.

2.5 Approximation Metrics

In order to assess the validity of our method, we must quantify the error of our estimates numerically. For evaluation of the quantile and superquantile estimates, we use the average absolute deviation (AAD) and the median absolute deviation (MAD) between the estimated and known values at various p across all optimization iterations ($j \in J$). Here we evaluate our DSE estimation results at $p = 0.80$, 0.90, 0.95, and 0.99. Since we are dealing with

heavy-tailed densities, we prefer two centrality measures so as to assess the impact of potential outlier skewing. Thus, we formulate our DSE error metrics as

$$\text{AAD}(p) = \frac{1}{J}\sum_{j}\left|\hat{\bar{Q}}(p) - \bar{Q}(p)\right|,$$
$$\text{MAD}(p) = \underset{j}{\text{median}}\left\{\left|\hat{\bar{Q}}(p) - \bar{Q}(p)\right|\right\},$$

with quantile errors calculated in the same manner.

For density estimation, we also propose two metrics. The first is a measure of overall fit, where we sum the squared errors (SSE) at epi-spline segment endpoints across the entire mesh. The second metric uses the same measure, but only across endpoints within the tail region. As such, we term this the sum of squared tail errors, or SSTE, and use the 80th, 90th, and 95th quantiles to define the start of right tail regions according to

$$SSE = \sum_{m}\left(\hat{f}(x) - f(x)\right)^2,$$
$$SSTE = \sum_{m|x_L^m \geq Q(p)}\left(\hat{f}(x) - f(x)\right)^2.$$

CHAPTER 3:
Distributional Benchmarks

In order to evaluate and quantify the impact of accurate superquantile estimation within the density estimation framework, we evaluate the methodology posited in Chapter 2 by considering samples from known distributional benchmarks. We explore two distributions, the exponential and the Pareto, making use of their well-defined density and quantile functions in order to measure estimation errors in quantiles, superquantiles, overall density, and tail density.

3.1 Exponential Superquantile Estimation

We begin with the exponential case, where a comparison is relevant because an exponential decay in the tail density serves as the boundary case for classification of heavy tails. Thus, this exploration serves as a validation for such data that might just barely be considered heavy-tailed. We begin by taking a sample of 30 IID observations, randomly generated from an exponential distribution with a rate of one-fifth.

$$x_1, x_2, ..., x_{30} \sim f_x = \lambda e^{-\lambda x}, \qquad \text{for } \lambda = \frac{1}{5}$$

Given knowledge of the underlying distribution, we can compare the results attained via epi-spline estimates with the true values for quantiles and superquantiles at specific p-values. From the PDF, we derive the following quantile function $Q_X(p)$, superquantile function $\bar{Q}_X(p)$, and DSE function $E^*_X(p)$ for the exponential case. Probability density, quantile, and DSE plots for the exponential case can be seen in Figure 1.3 and Figure 1.5.

$$Q_X(p) = -\frac{1}{\lambda}\ln(1-p) \tag{3.1}$$

$$\bar{Q}_X(p) = \frac{1}{\lambda}\left[1 - \ln(1-p)\right] \tag{3.2}$$

$$E^*_X(p) = (p-1)\bar{Q}_X(p) = \frac{p-1}{\lambda}\left[1 - \ln(1-p)\right] \tag{3.3}$$

With our random sample, we apply different combinations of constraints that constitute varying degrees of distributional knowledge and soft information by way of four progressive scenarios. As the scenarios progress, we gradually increase distributional knowledge in a manner that seems reasonable. Scenario 1 serves as a base case, with no prior knowledge outside of what we can deduce from the sample itself. Using bootstrapping, we obtain 95% confidence intervals for the mean and standard deviation (as per section A.3), as well as 95% confidence intervals for the median and quartiles using binomial approximations (as per section A.4). In Scenario 2, we acquire knowledge of the distribution's median as well as the "heavy-tailed" characteristic. In Scenario 3 we learn its median and quartile values, as well as the fact that X cannot be negative (i.e., a minimum value threshold). Finally, given the difficulty in their estimation, we wait until Scenario 4 to add perfect knowledge of the distribution's mean and variance, which are used to inform both the starting conditions and the lower bounding constraint. We also add the knowledge that the functions is monotonically decreasing. These constraint formulations are summarized with their associated equations in Table 3.1.

Scenario	Constraints	Equations
DSE 1	Continuous	Equation 2.2
	Differentiable	Equation 2.3
	Convex	Equation 2.1
	Endpoint	Equation 2.7
	Startpoint	Equation 2.6
	Lower Bound	Equation 2.9
	Quartiles	Equation 2.11
	Min. Value	Equation 2.14
	Max. Value	Equation 2.13
DSE 2	+ Median Known	Equation 2.10 ($p = 0.5$)
	+ Heavy-Tailed	Equation 2.12
DSE 3	+ Quartiles Known	Equation 2.10 ($p = 0.25,\ 0.75$)
	+ Minimum Value	Equation 2.18 ($x \geq 0$)
DSE 4	+ Mean Known	Equation 2.5
	+ Variance Known	Equation 2.8
	+ Monotonic Decrease	Equation 2.21

Table 3.1: Constraint Configuration Scenarios for Estimating DSE

For each scenario, 100 iterations of the optimization are performed, with each iteration using a different randomly generated sample of 30 observations and a mesh of

$K = 20$ segments. The results of these scenario replications are summarized in Table 3.2. For the objective function, we begin with $\rho = 0$, incrementally increasing the penalty term if results are determined to require additional smoothing. For the exponential case, no smoothing was required.

Quantiles	$Q_X(0.80)$	$Q_X(0.90)$	$Q_X(0.95)$	$Q_X(0.99)$
Scenario 1	1.129	2.397	4.637	9.426
Scenario 2	2.659	4.000	2.604	5.425
Scenario 3	0.943	1.688	2.334	7.038
Scenario 4	0.313	0.318	0.271	1.771
Superquantiles	$\bar{Q}_X(0.80)$	$\bar{Q}_X(0.90)$	$\bar{Q}_X(0.95)$	$\bar{Q}_X(0.99)$
Scenario 1	3.496	5.586	7.623	7.190
Scenario 2	3.529	3.791	4.261	4.152
Scenario 3	2.524	3.718	5.287	4.919
Scenario 4	0.096	0.473	1.037	2.514

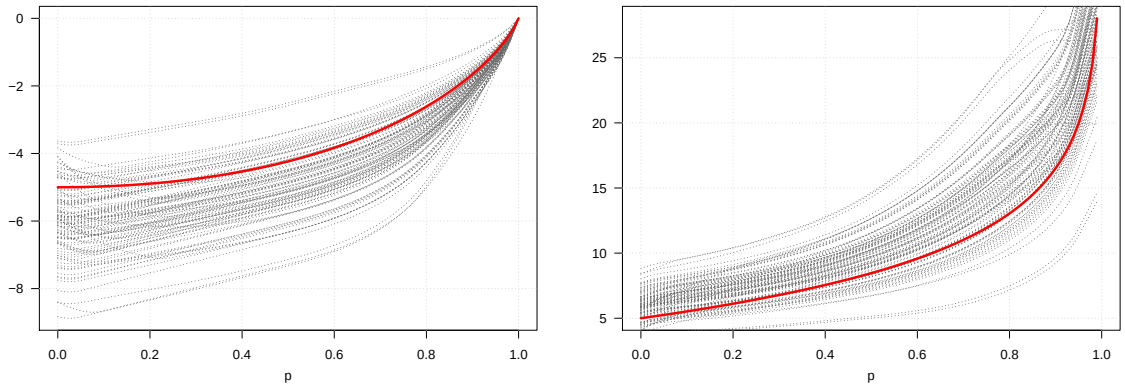
Table 3.2: AAD Estimation Errors for 100 Iterations on Exponential Benchmark

We see from Table 3.2 that as the degree of distributional knowledge increases, the accuracy of both quantile and superquantile estimates generally improves. Somewhat intuitively, estimates further in the tail (closer to $p = 1$) are less accurate than those nearer the center. Additionally, we find a significant jump in accuracy for Scenario 4, suggesting that knowledge of the mean and variance is valuable in both quantile and superquantile estimation. A plot of the 100 DSE epi-spline estimates and their associated superquantile curves for Scenario 2 is provided in Figure 3.1. Additional scenario graphics are included in Section A.6.

3.2 Exponential Density Estimation

With the estimates of Q_X and $\bar{Q}_X$ for $p = 0.80$, 0.90, 0.95, and 0.99, we now perform a second constrained optimization to estimate the probability density of the underlying distribution. As before, we constrain the optimizations to correspond to the same levels of distributional knowledge as in the DSE scenarios. For example, Scenario 3 here would assume perfect knowledge of median and quartiles in the same manner as Scenario 3 during DSE estimation. In this way, the scenarios for density estimation represent a logical progression within the overall methodological framework.

Figure 3.1: DSE and Superquantile Estimates for Exponential Scenario 2



(a) Scenario 2 DSE Estimates

(b) Scenario 2 Superquantile Estimates

Dual of superexpectations (left) and superquantile (right) second-order epi-spline estimates for the exponential case using constraints via Scenario 2 and optimized for curvature. The true values are shown in solid red. Notice the generally conservative estimates made, regularly overestimating the superquantiles in the range near $p = 1$.

For each scenario we optimize on the basis of both MLP and MEP. In addition, we run each optimization twice. In the first iteration, we ignore the quantile and superquantile estimates derived in the previous section. In the second iteration we apply these estimates for $p = 0.80,\ 0.90,\ 0.95$, and 0.99 so as to assess the impact of the previous step. This second iteration is denoted by a star ("*") in the scenario name, indicating that it uses $E^*_X(p)$ optimization to obtain additional constraints in density estimation. In total, this requires 16 formulations, each with 100 iterations using different sample data, for 1,600 total optimizations. We recall that constraints for mean, quantile, and superquantiles are bounded as per subsection 2.4.3 with MLP utilizing lower bounding and MEP utilizing upper bounding.

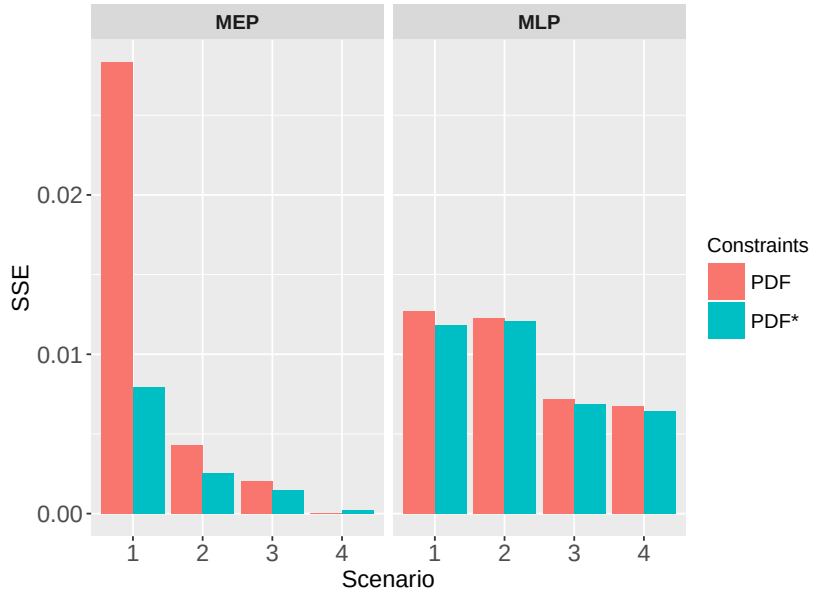
Due to potential inaccuracies in sample estimates and quantile/superquantile estimates from DSE optimization, for any constraint reliant upon "less than perfect" knowledge we incorporate a highly penalized slack variable so as to ensure feasibility throughout the 100 iterations. Finally, soft information is applied consistently throughout scenarios as per Table 3.3.

The results of our 16 scenarios are depicted in Figure 3.2, which shows the SSE across the entire mesh for both MLP and MEP formulations and averaged across the 100

Soft Information for Exponential PDF Estimation
Unity, Continuous, Non-Negative
Lower/Upper Limits: $[0,\ 60]$
Monotonic Decrease: $x \geq 0$
Convex Right Tail: $x \geq 0$
Max. Gradient Change: $\Delta b_1 \leq 0.01$

Table 3.3: Soft Information for Exponential Density Estimation

Figure 3.2: SSE Results for Exponential Benchmark

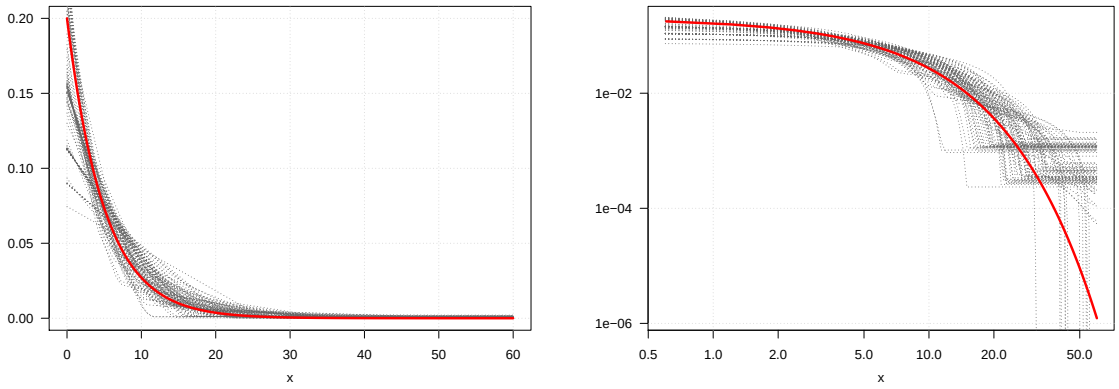


Sum of Squared Errors (SSE) for first-order epi-spline estimates evaluated across all mesh segments for both MLP and MEP formulations and averaged across 100 iterations. Results show an improvement as distributional knowledge increases, and when quantile/superquantile estimates are incorporated.

iterations. We note two general trends. First, as scenarios increase in their level of distributional knowledge there is a reduction in overall density error, the degree to which seems to depend on objective formulation. Second, the inclusion of estimated quantile and superquantile values causes a noticeable decrease in overall error, especially in early MEP scenarios. A more detailed numerical summary of these results is included in Table A.1.

The results of density estimation across the 100 iterations for Scenario 1a* is provided in Figure 3.3. We include here a second plot on a log-scale to better visualize the tail region estimates. The plots of additional scenarios are provided in section A.6.

Figure 3.3: PDF Estimates for Exponential Scenario 1a*



(a) Scenario 1a* PDF

(b) Scenario 1a* PDF Log-Scale

Density estimates (left), also shown on a log scale (right), for second-order epi-spline estimates for Scenario 1a* of the exponential benchmark. The true values are shown in solid red.

3.3 Pareto Superquantile Estimation

For our second benchmark, we explore a distribution known to possess a heavy tail, the Pareto. Using the same methodology as before, we begin with a sample of 30 observations taken from a Pareto distribution with shape parameter (α) of three and a scale parameter (x_m) of ten.

$$x_1, x_2, ..., x_{30} \sim f_x = \frac{\alpha x_m^\alpha}{(x + x_m)^\alpha}, \qquad \text{for } \alpha = 3,\ x_m = 10$$

From the probability density function, we attain the following quantile function $Q_X(p)$, superquantile function $\bar{Q}_X(p)$, and DSE function $E_X^*(p)$ for the Pareto case. We again reference Figure 1.3 and Figure 1.5 of Chapter 1 for plots of probability density, quantile, and DSE functions.

$$Q_X(p) = \frac{x_m}{(1-p)^{1/\alpha}} - x_m \tag{3.4}$$

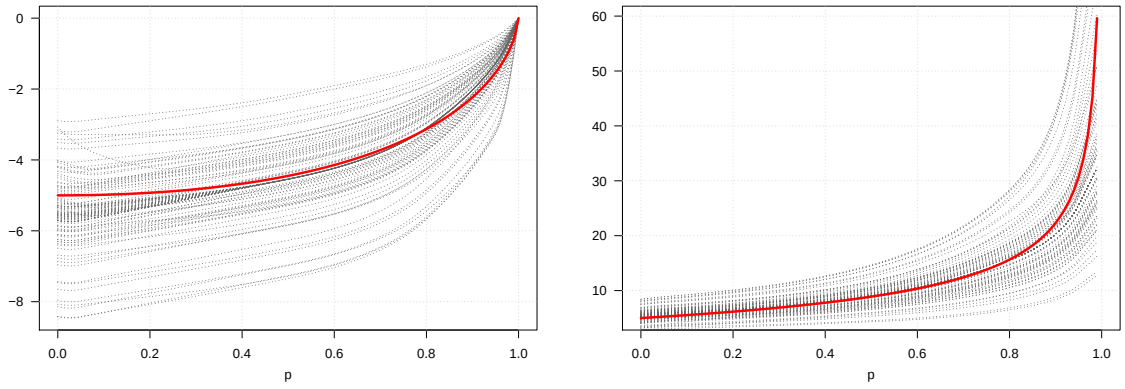
$$\bar{Q}_X(p) = \frac{\alpha x_m}{(\alpha-1)(1-p)^{1/\alpha}} - x_m \tag{3.5}$$

$$E_X^*(p) = \frac{-\alpha x_m}{(\alpha-1)}(1-p)^{\frac{\alpha-1}{\alpha}} + x_m(1-p) \tag{3.6}$$

3.3.1 Superquantile Estimation

Using the same constraint scenarios as outlined for the exponential case, we arrive at the following results summarized in Table 3.4. We first notice the considerable increase in quantile and superquantile estimation errors, particularly for higher p-values. Still though, we continue to see the general trend of decreasing errors as the scenarios progress. Scenario 2 epi-spline estimates are provided in Figure 3.4, where we see similar results to those of the exponential case.

Figure 3.4: DSE and Superquantile Estimates for Pareto Scenario 2



(a) Scenario 2 DSE Estimates

(b) Scenario 2 Superquantile Estimates

Dual of superexpectations (left) and superquantile (right) second-order epi-spline estimates for Scenarios 2 for the Pareto case. The true values are shown in solid red.

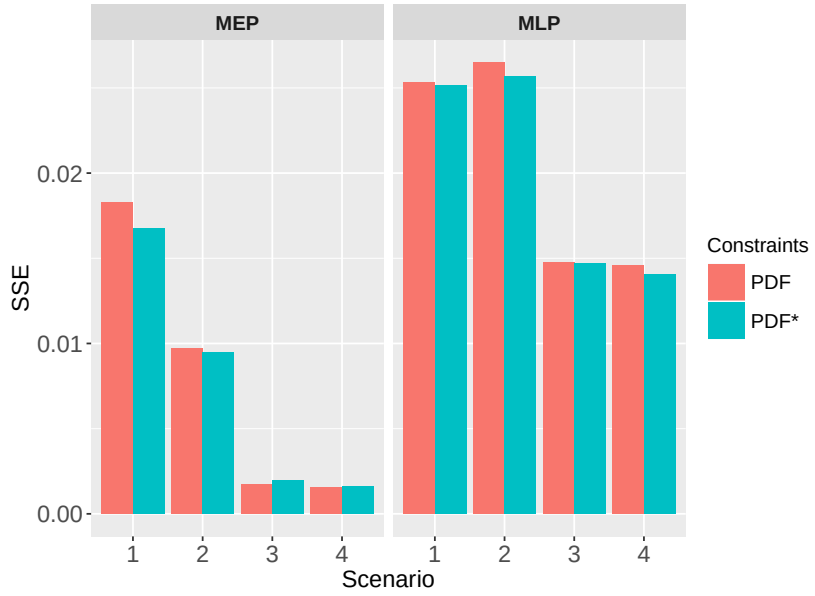
Quantiles	$Q_X(0.80)$	$Q_X(0.90)$	$Q_X(0.95)$	$Q_X(0.99)$
Scenario 1	2.107	4.952	9.347	13.474
Scenario 2	1.968	2.184	2.181	11.176
Scenario 3	0.882	0.801	2.405	13.032
Scenario 4	1.162	1.064	2.091	5.929
Superquantiles	$\bar{Q}_X(0.80)$	$\bar{Q}_X(0.90)$	$\bar{Q}_X(0.95)$	$\bar{Q}_X(0.99)$
Scenario 1	6.229	9.222	11.575	22.412
Scenario 2	2.768	4.873	8.577	27.326
Scenario 3	2.841	5.436	9.743	28.872
Scenario 4	0.378	1.866	4.531	22.092

Table 3.4: AAD Estimation Errors for 100 Iterations on Pareto Benchmark

3.4 Pareto Density Estimation

We again run our optimizations across the 16 formulations, plotting density estimates and recording errors for each of the 100 iterations. We modify our mesh, expanding it from $[0,\ 60]$ (as in the exponential case) to $[0,\ 110]$ so as to encompass all X observations. All other soft information remains as before. The numerical results of these error calculations is provided in Section A.6 and plotted in Figure 3.5. Scenario 1a* again serves as an illustrative case for optimization on the basis of log-likelihood with minimal information, where we again include a log-scale plot to highlight estimation in the tail region (Figure 3.6).

Figure 3.5: SSE Results for Pareto Benchmark

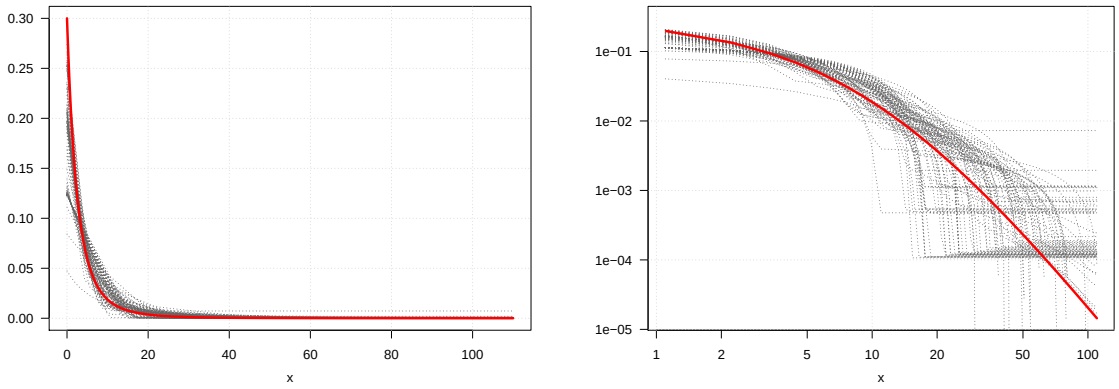


Sum of Squared Errors (SSE) for first-order epi-spline estimates evaluated across all mesh segments for both MLP and MEP formulations and averaged across 100 iterations. In contrast to the exponential case, the addition of additional quantile/superquantile constraints doesn't appear to help as much in overall density estimation, at least from a quantitative perspective.

3.5 Benchmark Summary

The exponential and Pareto cases provided somewhat mixed results. On one hand, the exponential benchmark seems to suggest that the addition of quantile and superquantile constraints informed though DSE approximation can be quite beneficial, particularly for MEP formulations that have access to very little distributional knowledge (as in Scenarios 1 or 2). Contrast this to the Pareto cases, where the addition of these same constraints

Figure 3.6: PDF Estimates for Pareto Scenario 1a*



(a) Scenario 1a* PDF

(b) Scenario 1a* PDF Log-Scale

Density estimates (left), also shown on a log scale (right), for second-order epi-spline estimates for Scenario 1a of the Pareto benchmark across 100 iterations. The true values are shown in solid red.

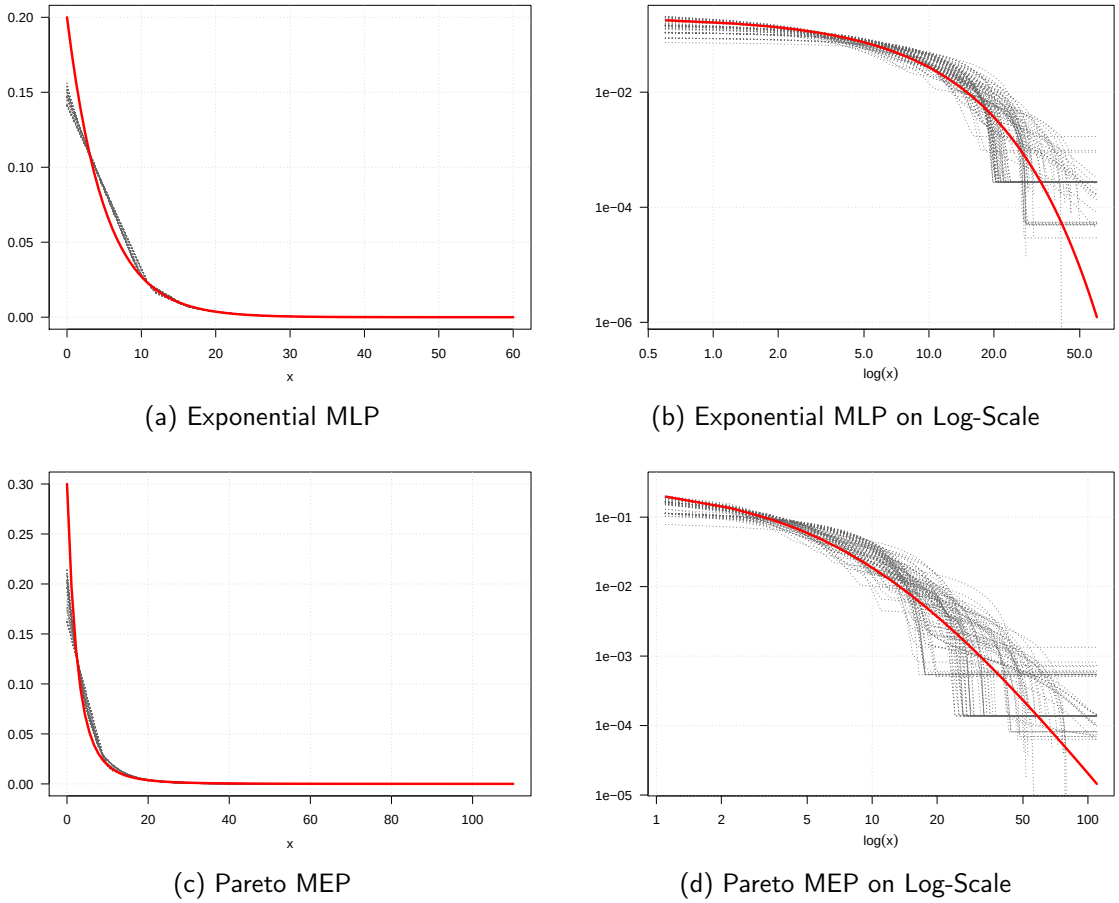
seemed to have almost no impact. This, however, might be slightly misleading. Recall that the error of quantile/superquantile estimates for the Pareto case was higher than that of the exponential. If predicted values possess significant error, then it follows that their inclusion in an density estimation framework could actually be detrimental rather than helpful. For this reason, we do not rely purely on a numerical evaluation of the technique's efficacy, but also include plots displaying the fidelity of the estimates themselves for visual comparison. Again, these can be found in Section A.6.

3.5.1 Hypothetical Constraint Implementation

One trend that remains consistent, however, is the improvement in accuracy as we move from Scenario 1 to 4. The addition of new information into the density estimation framework improves overall estimation considerably, with perhaps the most noticeable improvement occurring between Scenarios 2 and 3, as seen in Figure 3.2 and Figure 3.5. In light of this observation, we feel confident that quantile and superquantile constraint incorporation can improve density estimates provided the constraints are near the true quantile/superquantile values. To assess this, we provide a fifth scenario, showing MLP and MEP formulations utilizing a perfect knowledge of the quantile/superquantile constraints used in Scenario 1 and without any other additional information. We refer to these scenarios as run "*P*," with

results depicted in Figure 3.7 and summarized in Table 3.5.

Figure 3.7: PDF Estimates with Perfect Quantile Knowledge



(a) Exponential MLP

(b) Exponential MLP on Log-Scale

(c) Pareto MEP

(d) Pareto MEP on Log-Scale

Scenario 1 runs for both the Exponential case (MLP) and Pareto case (MEP) given a hypothetical perfect knowledge of quantile/superquantile values for $p = 0.80,\ 0.90,\ 0.95$ and 0.99. Compare these to the results shown earlier and it becomes apparent that accurate quantile/superquantile estimates can greatly enhance density optimization.

Although improvement is not made in every case, we generally find that MEP formulations are greatly improved by adding accurate quantile/superquantile constraints. Aggregating the errors across both cases and formulations we calculate an overall improvement in SSE of roughly 25% between Scenarios 1 and P, and an overall improvement in SSTE (averaged across $p = 0.80,\ 0.90,\ 0.95$) of 15% for MLP and greater than 80% for MEP.

Scenario	SSE	SSTE (0.80)	SSTE (0.90)	SSTE (0.95)
Exp 1a	0.0127	2.11e-03	1.11e-03	5.43e-04
Exp 1a*	0.0118	1.52e-03	8.36e-04	4.59e-04
Exp Pa	0.0119	1.70e-03	9.52e-04	4.82e-04
Exp 1b	0.0283	2.94e-03	2.66e-03	2.06e-03
Exp 1b*	0.0079	0.55e-03	2.46e-04	1.71e-04
Exp Pb	0.0064	0.28e-03	8.64e-06	1.96e-06
Par 1a	0.0253	1.85e-03	6.45e-04	1.85e-04
Par 1a*	0.0251	1.39e-03	5.20e-04	2.16e-04
Par Pa	0.0243	1.23e-03	5.28e-04	2.47e-04
Par 1b	0.0183	0.84e-03	3.38e-04	8.03e-05
Par 1b*	0.0168	0.75e-03	2.92e-04	7.80e-05
Par Pb	0.0164	0.21e-03	3.54e-05	1.11e-05

Table 3.5: Impact of Perfect Quantile Knowledge on Error

3.5.2 Observations

The use of second-order epi-splines provides a relatively simple way to leverage a curvature objective function while requiring relatively few mesh segments (only 20 here). While this eases computation, it also limits the degrees of freedom afforded the optimization solver and as such, limits the flexibility of the epi-spline estimates. This issue came to light in trying to impose multiple simultaneous constraints that exhausted the degrees of freedom provided by second-order splines, resulting in infeasibility issues.

One could of course simply increase the complexity of the epi-spline model through enhancing the degree of the polynomial. This would enable a more flexible fit that could accommodate more elaborate constraint combinations at the expense of increased computational cost. Alternatively, one could also simply add slack variables to those constraints deemed desirable, though not necessarily required. This approach would have the added benefit of providing the marginal costs of taught constraints, which provides the intuitive appeal of relating to a confidence in imposing the constraint. As such, that was the method pursued here.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 4:
Non-Parametric Financial Data

Having observed the impact of DSE estimation and the relative merit of the method on known distributional benchmarks, we now turn our focus to the financial sector, where data sets are known to commonly exhibit heavy-tailed characteristics. We obtain this data from Dr. Uryasev, a professor and Director of the Risk Management and Financial Engineering Lab at the University of Florida. The data includes 1,000 observations of both positive and negative values of a particular financial measure, which we will term "cost."

4.1 Data Evaluation

We begin with a summary of the raw data Y, shown in Table 4.1, as well as a PDF estimate derived using the **density** function from the R base-package **stats**, which we will term $d(Y)$, across all 1,000 observations. Here, the density is estimated using kernel smoothing as described in [17]. Many other density estimation packages are available, with a comparison of efficiency and accuracy explored in [18].

Statistic	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Y	−603.40	−127.90	−10.33	7.44	118.80	1095.00
$d(Y)$	−728.30	−129.89	−11.71	7.44	125.51	1219.40

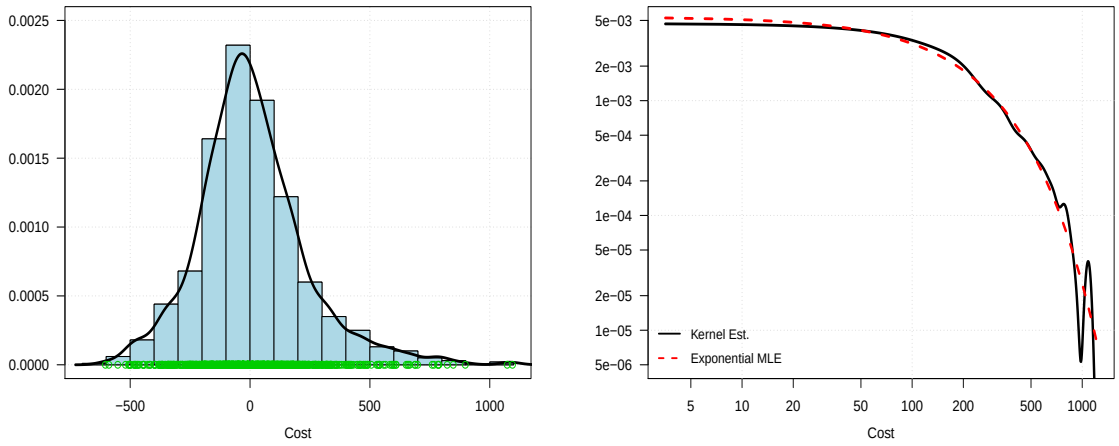
Table 4.1: Summary Statistics of Financial Data

We assume the data to have come from some continuous, but unknown distribution. With 1,000 observations, the PDF estimate arrived at by $d(Y)$ provides us with a qualitatively acceptable approximation to the true distribution within the range of observed values. We realize, however, that $d(Y)$ is not "the true" distribution, and that its approximations (particularly in the extreme tail regions) will invariably be flawed. Without the underlying ground-truth, a true numerical evaluation of results (as in Chapter 3) remains ill-posed. As such, we focus the majority of this chapter on a qualitative evaluation of density, avoiding quantitative comparisons of tail regions.

Inspecting the data, we suspect (as per Figure 4.1a) that the right tail of the data may be heavy-tailed. In fact, if we look only at the density to the right of the mode (which in this

case is roughly zero) and rescale accordingly, we see that by comparison with an exponential (Figure 4.1b), the data does appear to possess a heavy tail. Though in reality an investigation such as this would not be possible (as we are using all 1,000 observations), we include it here to show that our data can reasonably be considered heavy-tailed, a specification we earlier cited in the outline for this approach.

Figure 4.1: Financial Data Density Analysis



(a) Finance Data PDF

(b) Finance Data Tail Region

The PDF (left) and tail density (right) for the financial data. The tail density here is compared with an exponential distribution using MLE to estimate the rate parameter and plotted on a log-scale. We see evidence here that the data can be considered heavy-tailed.

4.2 Quantile and Superquantile Estimations

We intend to evaluate the efficacy of our approach on small data sets for which limited knowledge is available. As such, we depart from the scenarios outlined in Chapter 3 to rely solely on information derived from the sample data itself.

We begin by taking 30 randomly sampled observations (without replacement) from the original data and calculating summary statistics in the same manner as Chapter 3. Doing this over 100 replications, we apply constraint equations that closely relate to Scenario 1 from earlier, and outlined in Table 4.2. We again optimize for curvature, in this case utilizing a smoothing parameter of $\rho = 0.01$. The results of the 100 replications for quantile/superquantile predictions are summarized in Table 4.3, along with their associated standard errors (SE). We also include the quantile and superquantile values derived from

both Y and $d(Y)$ as a means of comparison. From these estimates, and from visual inspection of Figure 4.2, we find our method routinely overestimating both quantile and superquantile values. Recall, however, the curvature objective function was implemented with the intent of achieving conservative predictions.

Constraints	Equations
Continuous, Differentiable, Convex	Equation 2.2, Equation 2.3, Equation 2.1
Startpoint, Endpoint, Quantiles	Equation 2.6, Equation 2.7, Equation 2.11
Min. Value, Max. Value	Equation 2.14, Equation 2.13
Lower Bound, Heavy-Tailed	Equation 2.9, Equation 2.12

Table 4.2: Constraint Configuration for Financial DSE Estimation

Quantiles	$Q_x(0.80)$	$Q_x(0.90)$	$Q_x(0.95)$	$Q_x(0.99)$
Y	158.33	278.49	419.09	699.71
$d(Y)$	163.63	289.41	430.44	750.61
Epi-Spline Est.	288.89	414.43	537.25	944.34
Epi-Spline SE.	69.10	94.82	124.42	321.01
Superquantiles	$\bar{Q}_x(0.80)$	$\bar{Q}_x(0.90)$	$\bar{Q}_x(0.95)$	$\bar{Q}_x(0.99)$
Y	340.44	472.81	604.03	861.88
$d(Y)$	343.17	471.17	598.49	781.69
Epi-Spline Est.	490.27	633.76	791.68	995.23
Epi-Spline SE.	110.91	145.55	187.39	246.03

Table 4.3: Average Epi-Spline Quantile/Superquanile Estimates

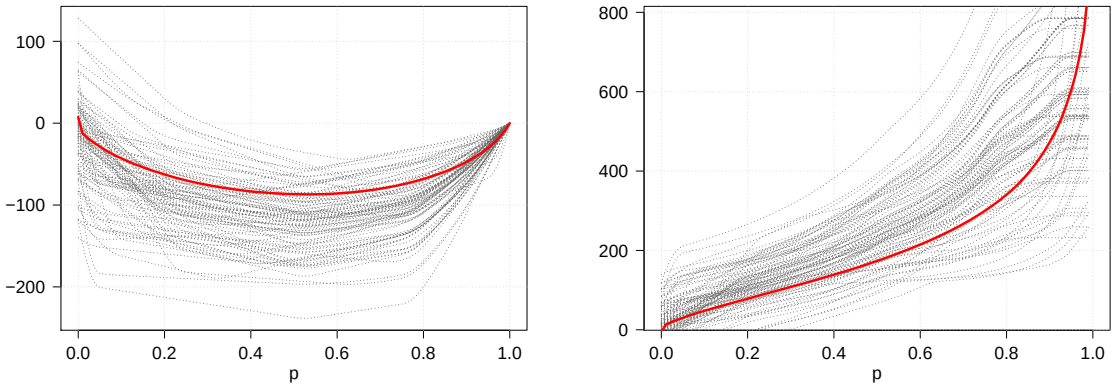
4.3 Density Estimates

Using the quantile and superquantile estimates previously derived, we again estimate the financial PDF under constrained optimization of both MLP and MEP as in Chapter 3. Additionally, we apply soft information for all formulations given on Table 4.4.

4.4 Comparison of Methods

We compare the results achieved with epi-splines to densities obtained using kernel smoothing on 100 replications, each with a different sample of 30 observations. Sample kernel estimates are again obtained using the **density** function from the **stats** package in R. For kernel smoothing, we enforce mesh endpoints and resolution to match that of the epi-spline

Figure 4.2: DSE and Superquantile Estimates for Financial Data



(a) Dual of Superexpectations

(b) Superquantiles

DSE and superquantile estimates for the financial data samples of size 30. Constrained optimization is performed as per Table 4.2. The red line corresponds to the $d(Y)$ kernel estimate across the entire data set. We can see the general overestimation performed by the objective curvature function on the superquantile plot, leading to conservative predictions.

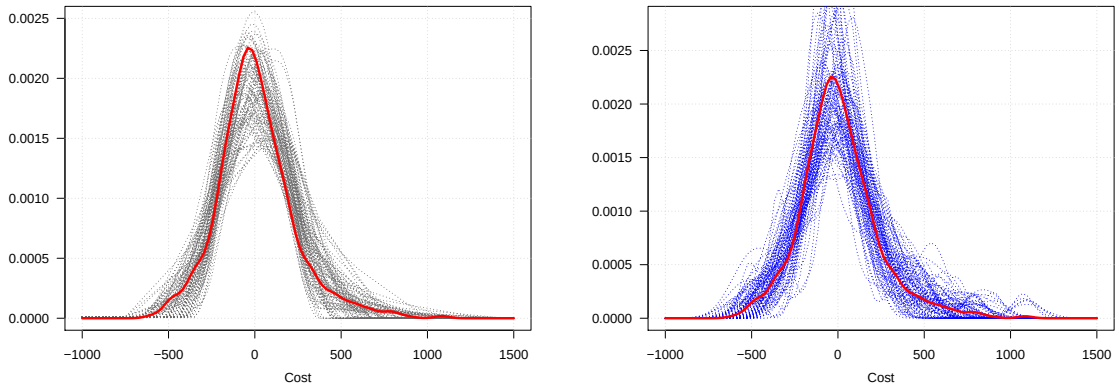
Financial Soft Information
Unity, Continuous, Non-Negative
Lower/Upper Limits: $[-1,000,\ 1,500]$
Unimodal Inflection Points: $x = -250,\ 250$
Convex Right Tail: $x \geq 250$
Convex Left Tail: $x \leq -250$
Mode: $\max\{f(x)\} \in [-50, 50]$
Minimize Gradient Change: $\Delta b_1 \leq 2e-06$

Table 4.4: Constraint Formulations by Scenario for Density Estimation

estimates. This allows for direct comparison of density estimates at each mesh intersection. The results of these kernel smoothing estimates can be seen in Figure 4.3b, with a comparison of estimation errors summarized in Table 4.5.

With few assumptions made on the data, we can arrive at decent approximations that are comparable to those estimates made via naive kernel smoothing. If more soft information becomes available, epi-spline estimates can leverage the additional knowledge for incrementally better approximations. As an illustrative example of this, we run both methods over a single sample using an epi-spline MEP enhanced with an accurate knowledge of select superquantile values. In this case, our "accurate" superquantile values are those found

Figure 4.3: Epi-Spline vs. Kernel Smoothing Density Estimation



(a) MLP* Epi-Spline Estimates

(b) Kernel Smoothing Estimates

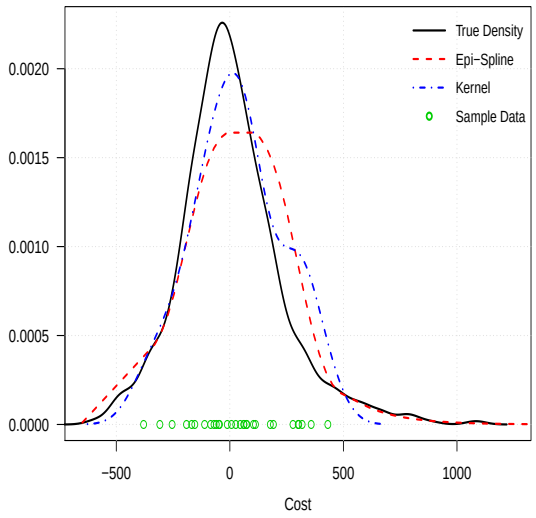
Density estimates for second-order epi-splines under MLP* (left) and kernel smoothing (right). The $d(Y)$ estimate is shown in solid red.

Estimation Method	SSE	SE
Epi-Spline MLP*	3.1720e-06	2.3709e-06
Epi-Spline MEP*	7.5303e-06	3.0263e-06
Kernel Smoothing	3.9536e-06	3.0835e-06

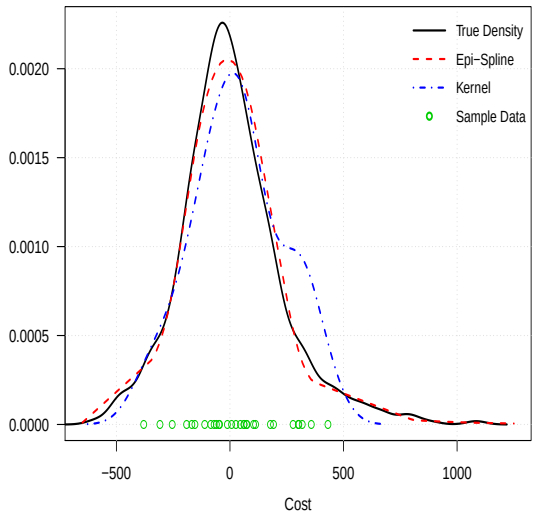
Table 4.5: PDF Estimation Error for Financial Data

in Table 4.1 for Y. The results are shown in Figure 4.4 and demonstrate the improvements possible with accurate quantile/superquantile estimates, particularly in the tail region. We recall that the density approximation of $d(Y)$ in solid black is not the "true" density, and that in all likelihood, the much smoother epi-spline estimate seen in the log-scale plots are probably better reflections of the true underlying tail density.

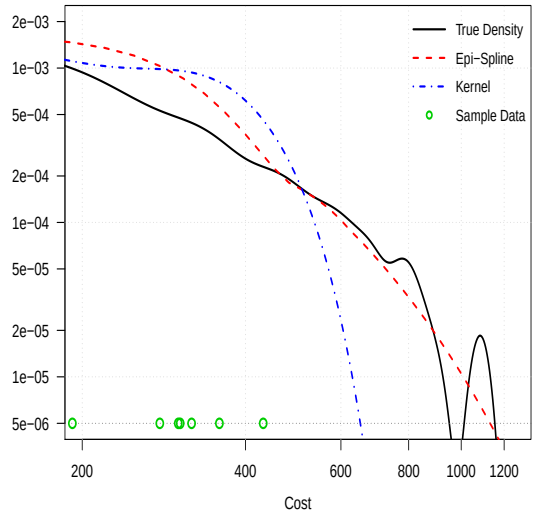
Figure 4.4: Comparison of PDF Estimation Methods



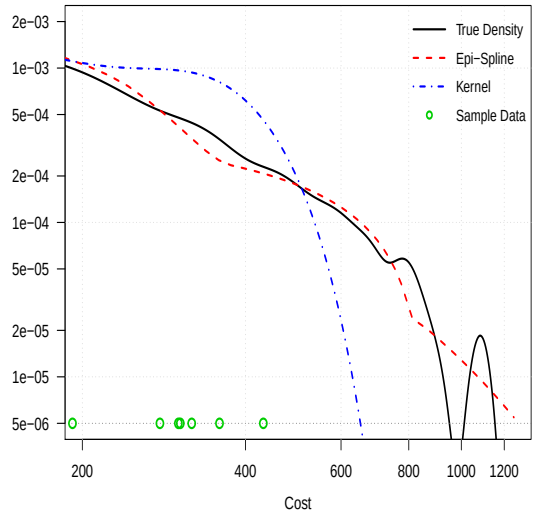
(a) Superquantiles Estimated



(b) Superquantiles Known



(c) $\bar{Q}_X$ Estimated (Log-Scale)



(d) $\bar{Q}_X$ Known (Log-Scale)

A comparison of density estimation using MEP optimized epi-splines informed through quantile and superquantile approximations (red), and kernel smoothing (blue). Both methods are used on a sample of 30 observations. With minimal knowledge, quantiles and superquantiles can be approximated and incorporated with additional soft information. As the fidelity and scope of knowledge grows, estimates improve. The left figures use quantile and superquantile estimates corresponding to Scenario 1 in DSE estimation. If these estimates were 100% accurate, the improved estimates would correspond to the figures on the right. Bottom figures are plotted on a log-scale and enhanced to show the right tail region.

CHAPTER 5:
Multi-Fidelity Hydrofoil Data

We now investigate a multi-fidelity data set that comes to us from Dr. S. Brizzolara of the MIT SeaGrant program. The data contains the output of 878 runs of both high-fidelity and low-fidelity fluid dynamic models that simulate a particular hydrofoil concept's drag-to-lift ratio. Since computational cost can vary significantly between low-fidelity vs. high-fidelity simulation runs, we wish to explore the accuracy of density estimates that rely on only a few high-fidelity observations supplemented with many low-fidelity runs. A more detailed exploration of multi-fidelity modeling and its applications to fluid dynamics simulations can be found in [19].

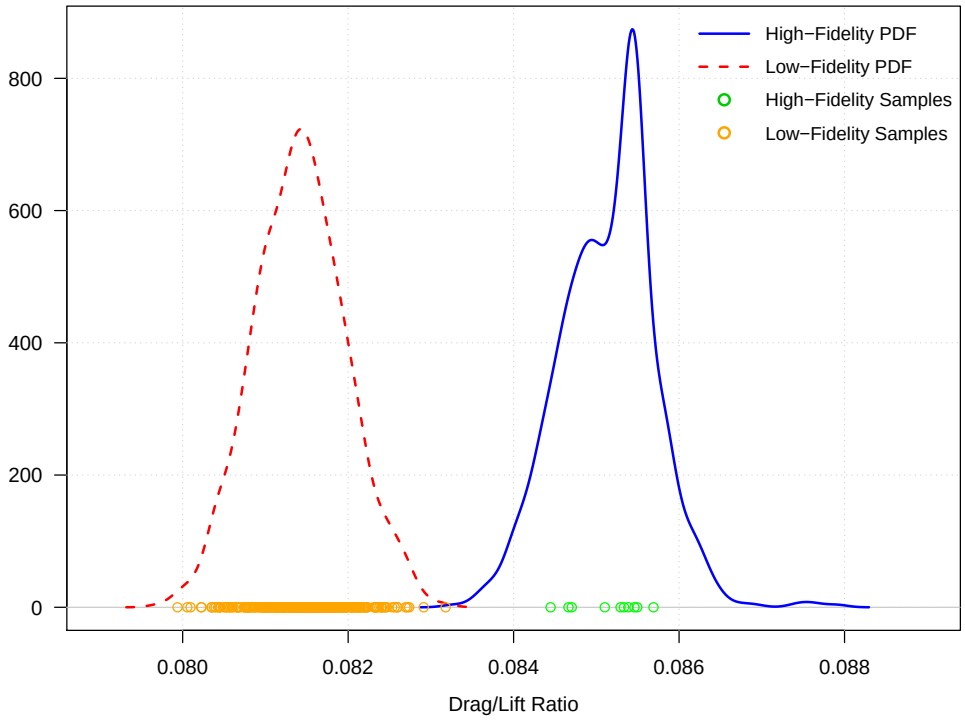
5.1 Hierarchical Model Blending

In this endeavor, we explore three options for informing the quantile and superquantile estimates that ultimately help shape our PDF approximations. In the first, we use only a small sample of high-fidelity observations, running a two-stage optimization in much the same way as Chapters 3 and 4. Next, we look at the correlated relation between the low-fidelity and high-fidelity observations, forming a linear model which we then use to predict high-fidelity quantiles and superquantiles from low-fidelity observations. Finally, we look at a blended model which uses a combination of the previous two methods. As our basis for comparison, we fit a kernel density across 875 of the 878 entries (we omit three as outliers) for both the high and low fidelity outputs. We will henceforth loosely refer to these as the kernel estimates for high- or low-fidelity data respectively. These distributions and their associated sample data are displayed in Figure 5.1.

5.2 High-Fidelity Modeling

With the 10 observation sample of high-fidelity simulation runs, we calculate the sample statistics and bootstrapped confidence intervals as in Chapter 3. Performing the first optimization using the same constraint formulations provided in Table 4.2, we obtain the results shown in Figure 5.2, where we see fairly severe overestimation of superquantiles.

Figure 5.1: Hydrofoil Data Inspection



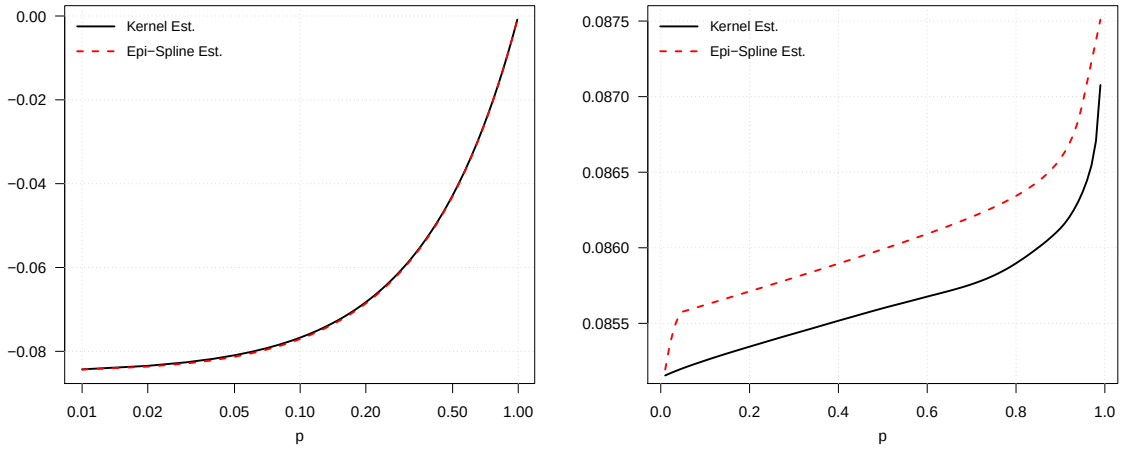
Densities of the two hydrofoil data sets formulated via kernel smoothing across 875 observations. We use 10 high-fidelity (green) and 300 low-fidelity (orange) samples respectively. Note the unimodality of the low-fidelity kernel estimate vice the dual-modality of the high-fidelity estimate.

Number of High-Fidelity Samples: $n_h = 10$
Number of Low-Fidelity Samples: $n_l = 300$

In addition to quantile predictions for upper tail regions, DSE functions can (as per Equation 1.6) also be used for quantile predictions in the lower tail regions, at least in theory. Although the fidelity of this approach remains unclear, we will attempt it here so as to obtain quantile estimates for regions where they cannot be reasonably attained from the sample itself. In this way, we estimate additional quantile values for $p = 0.05,\ 0.10$, and 0.20, including them as bounded quantile constraints within the density estimation formulation.

Using the estimates attained, we approximate density functions using the following combination of sample statistics, predicted quantiles/superquantiles, and additional soft information shown in Table 5.1. We elect here to optimize on the basis of log-likelihood rather than entropy due to the small sample size so as to provide further emphasis for

Figure 5.2: High-Fidelity Quantile Estimation



(a) Dual of Superexpectations

(b) Superquantiles

Results of quantile estimation on the high-fidelity sample data. Notice the log-scale used to display DSE (left). Although the DSE appears to be well-estimated, we see from the superquantiles (right) that we are overestimating virtually all superquantiles beyond the mean.

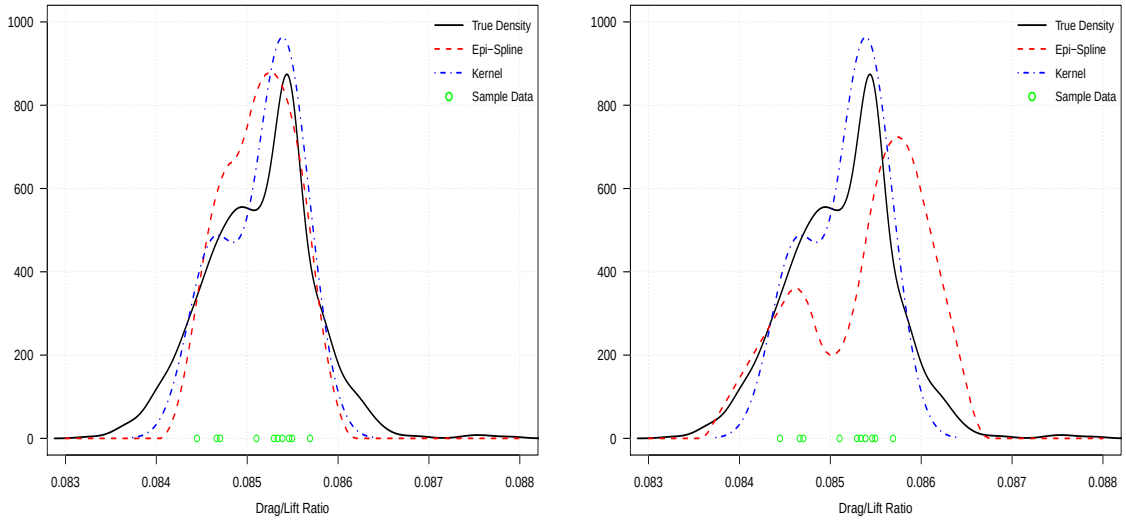
density around the high-fidelity observations. Results are shown with and without the implementation of the quantile and superquantile constraints in Figure 5.3.

HSV Soft Information
Unity, Continuous, Non-Negative
Lower/Upper Limits: $[0.083,\ 0.088]$
Unimodal Inflection Points: $x = 0.0825,\ 0.0845$
Convex Right Tail: $x \geq 0.0857$
Minimize Gradient Change: $\Delta b_1 \leq 10e6$

Table 5.1: Constraint Formulations by Scenario for Density Estimation

We see from Figure 5.3 that although constrained optimization of log-likelihood can help identify the non-parametric dual-modality of the high-fidelity density, poor quantile and superquantile estimates result in tail weights that actually prove detrimental to overall approximations. Since quantile estimates for low/high p-values were underestimated/overestimated due to the nature of our DSE optimization objective, the resulting constraints for the MLP formulation ended up pushing too much density into the tail regions.

Figure 5.3: PDF Estimates Using High-Fidelity Sample Only



(a) Without Additional Constraints

(b) With Additional Constraints

Results of PDF estimation on the high-fidelity sample data. Using only soft information and log-likelihood (left) we see a failure to properly estimate the tail regions for which no observations are present. Adding additional constraints, however, causes overestimation of the quantiles and superquantiles leads to an "over-filling" of the tail regions.

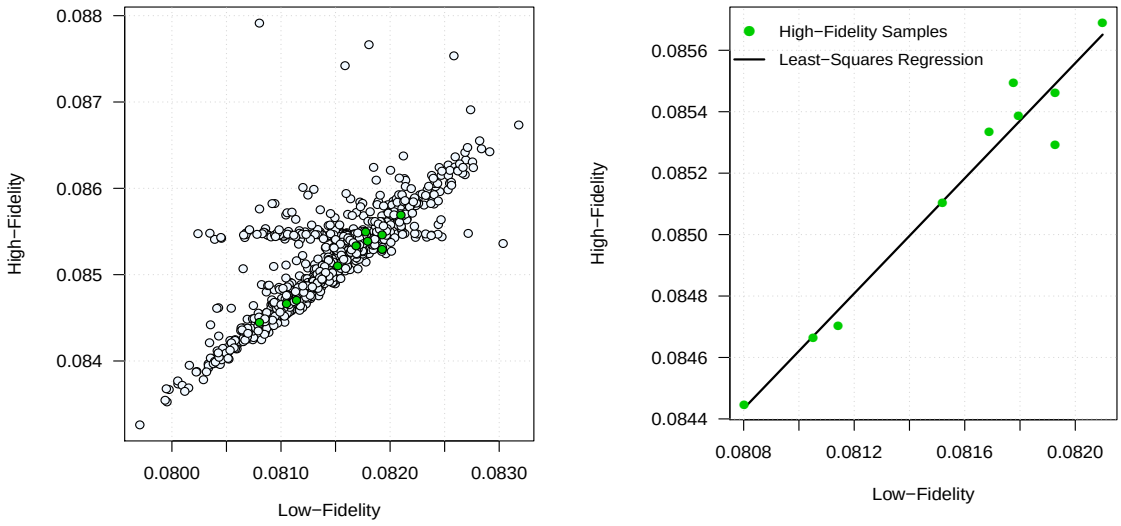
5.3 Low-Fidelity Surrogate Modeling

Estimates relying on only 10 observations can lead to predictions that prove either uninformative or perhaps even counterproductive. As such, we attempt to leverage the plethora of low-fidelity observations by relying on the underlying correlation between data sets. We begin by fitting a linear model and assessing the quality of fit for the 10 observations from the high-fidelity sample and their associated low-fidelity values. A scatter plot showing the general correlations, as well as the 10 points used for the regression model are provided in Figure 5.4. As such, the resulting least squares linear model becomes

$$\begin{aligned} X_h &= \beta_0 + \beta_1 X_l + \epsilon, \\ \widehat{\beta}_0 &= 8.702 \times 10^{-3}, \\ \widehat{\beta}_1 &= 0.937. \end{aligned} \tag{5.1}$$

Provided a reasonable fit of the linear model (here we have $R^2 \approx 0.95$) and the fact

Figure 5.4: PDF Estimates Using Low-Fidelity Linear Approximations



(a) High/Low-Fidelity Relation

(b) Multi-Fidelity Linear Model

Scatter plot (left) of high-to-low fidelity observations for all 875 observations and the linear model for predicting high-fidelity output from low-fidelity input values (right). The high-fidelity sample data (green) is common to both plots.

that linear transformations preserve shape and distribution, we can translate mean, variance, and quantile values for the low-fidelity sample data into intervals for the high-fidelity set by

$$\bar{X}_h \approx \widehat{\beta}_0 + \widehat{\beta}_1 \bar{X}_l, \tag{5.2}$$

$$\mathrm{Var}(X_h) \approx \widehat{\beta}_1^2 \mathrm{Var}(X_l), \tag{5.3}$$

$$Q_{X_h} \approx \widehat{\beta}_0 + \widehat{\beta}_1 Q_{X_l}. \tag{5.4}$$

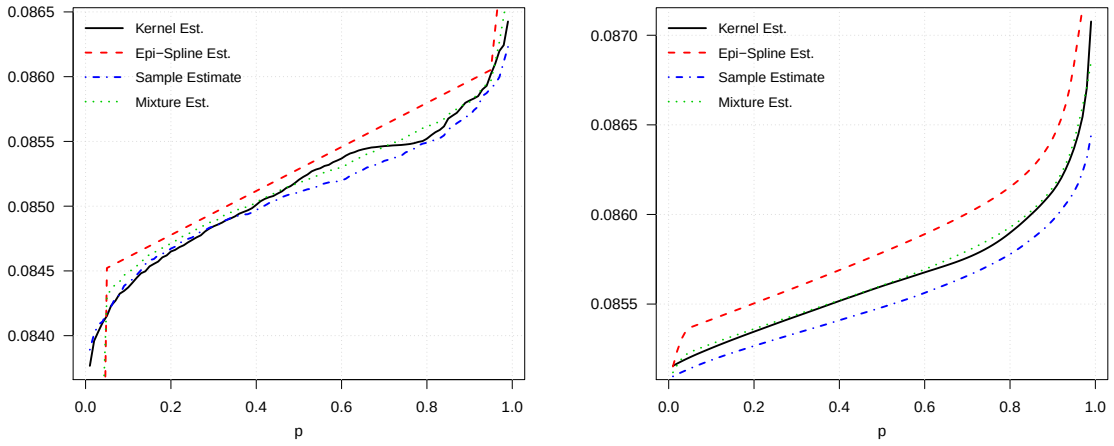
Additionally, we can estimate superquantile values from the transformed low-fidelity data itself, and by DSE approximation. For superquantiles estimated from the transformed low-fidelity data, we can further generate confidence intervals using bootstrapping if desired, though we will not attempt this here. Due to the larger sample size of 300, we can more confidently predict quantile/superquantile values for p closer to 0 and 1. In fact, because sample estimates tend to underestimate both quantile and superquantile values (due to lack of tail representation), and the tendency for these same estimates to be overestimated in DSE approximation (as seen earlier), we propose a weighted averaging of these predictions

to arrive at more accurate values. In this way, our estimates for quantiles and superquantiles are calculated according to

$$\widehat{Q}_{avg}(p) = \theta\widehat{Q}_s(p) + (1-\theta)\widehat{Q}_e(p), \tag{5.5}$$

where $\widehat{Q}_{avg}$ represents the weighted averaged of the sample quantile value (Q_s) and the epi-spline estimated quantile value (Q_e). θ provides the ratio of sample to epi-spline estimate weighting. Here, we will use $\theta = 0.6$. The results of this process are seen in Figure 5.5, where we note the greatly improved estimates obtained using a 60/40 weighted average.

Figure 5.5: Quantile & Superquantile Estimation Comparisons



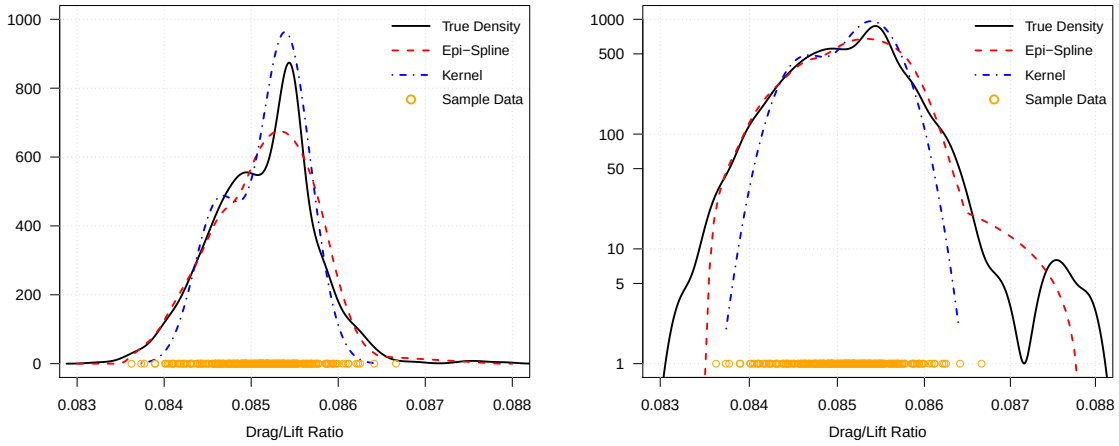
(a) Quantile Estimation Accuracy

(b) Superquantile Estimation Accuracy

Comparison of quantile and superquantile estimation techniques. Recall that the constraints informed via DSE estimation (red) were consistent over-estimators, while those informed only by sample statistics (blue) were consistent under-estimators. Here, the averaged values (green) provide the most accurate quantile and superquantile estimates.

Thus, we use the 300 low-fidelity observations to inform distributional characteristics of the high-fidelity data, such as mean, variance, quantiles, and superquantiles. Then, using these *surrogate* statistics, we impose the same formulation as before, but now with log-likelihood on the basis of the transformed low-fidelity observations. Results are seen in Figure 5.6. We note that by arriving at more accurate quantile and superquantile estimates, our tail densities have improved considerably. Despite this, because we rely solely on low-fidelity observations, we do not achieve the dual-modality we saw earlier with the high-fidelity formulations.

Figure 5.6: Density Using Low-Fidelity Surrogate Constraints



(a) PDF

(b) PDF on Log-Scale

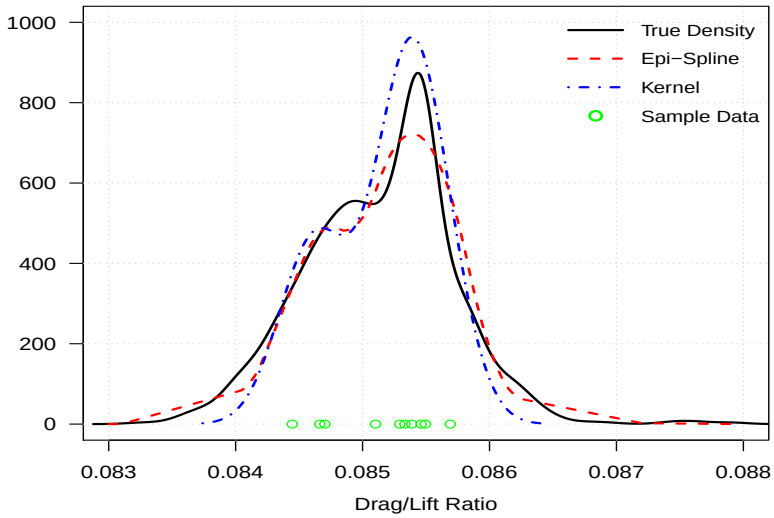
Density estimation results when using the low-fidelity observations. Estimates are transformed via a linear relationship determined by the correlations of the 10 observations from the high-fidelity sample. Quantile and superquantile estimates are further estimated by a weighted average of sample statistics and DSE approximations. Optimization is done via a MLP for the transformed low-fidelity samples. Results show greatly improved tail densities when compared to using high-fidelity observations alone.

5.4 High and Low Mixture Modeling

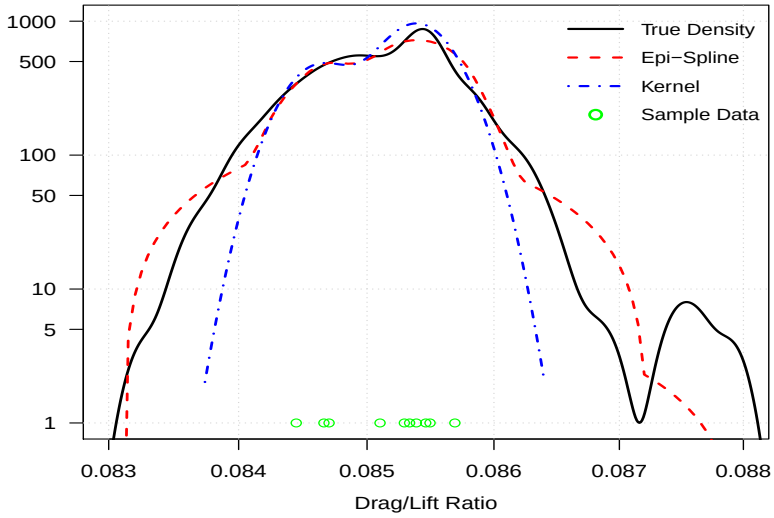
As a third and final method, we intend to combine the strengths of the previous two techniques so as to achieve both the non-parametric dual-modality of the high-fidelity model, as well as the improved tail densities observed in the low-fidelity surrogate model. To do this, we retain our quantile and superquantile averaged estimates from the low-fidelity sample, and implement them in a MLP formulation on the high-fidelity observations. All other soft information remains consistent with previous iterations. The results of this blended model are displayed in Figure 5.7.

The success of this method remains highly dependent on the accuracy of the linear model developed in section 5.3. In this case, the 10 high-fidelity data points used in model creation accurately reflected the preponderance of the high-to-low fidelity diagonal relationship seen in Figure 5.4. With only 10 observations, this will often not be the case, and in such circumstances, this method may actually be counterproductive. Thus, we point out not the validity of this method in all circumstances, but rather its potential uses provided a relatively high confidence in the high-to-low fidelity correlations.

Figure 5.7: PDF Estimates Using High/Low-Fidelity Mixture



(a) PDF Estimate



(b) PDF Estimate Log-Scale

The results of constrained optimization of log-likelihood on the high-fidelity observations using a quantile and superquantile constraints derived from averaged low-fidelity sample and DSE estimates. Notice that we have achieved the dual-modality desired, as well as much more accurate tail density estimates. When plotted on a log-scale (bottom), we can see the points corresponding to where our quantile/superquantile constraints take effect.

CHAPTER 6:
Conclusions

We have introduced a new method for improving density estimates for data coming from known or suspected heavy tailed distributions. By leveraging the dual relationship between superquantiles and superexpectations, we incorporate an additional set of constraints within a density optimization framework in a manner that, to the best of this author's knowledge, had not yet been attempted. By incorporating all potential sources of available information, we are able to provide tailored formulations for density approximations that enhance the fidelity of estimates for even small sample sizes.

The work presented here represents an initial excursion into epi-spline density approximations utilizing quantile and superquantile estimates. We demonstrate and quantify the value of accurate quantile and superquantile estimates in two benchmark examples that cover exponential tails and a heavy-tailed Pareto distribution. The financial scenario applies the method to a non-parametric data set where we compare our method against an existing density estimation technique and show that even gross-level quantile approximations can improve tail density estimates when properly bounded. Finally, the hydrofoil example demonstrates the feasibility of informing quantile and superquantile constraints for densities of limited observations through a linear approximation of more abundant data on the basis of reasonable correlation.

Although initial results show promise, they represent only a limited application of quantile and superquantile constraints within constrained epi-spline optimizations. Much work is left to be done, particularly in the realm of superquantile estimation itself. Second-order epi-splines, though flexible and relatively easy to formulate, offer limited degrees of freedom that can limit constraint combinations, a shortcoming that might be overcome through higher order epi-splines, such as those of the third or fourth order. Additionally, the choice of objective function in estimating the DSE arose through empirical observation on limited distributional examples. Better objective functions likely exist that can encourage the smooth and convex shapes seen in the DSE plots of Chapters 3-5.

Right tail density estimation remained the focus of this work, though we recognize

there are numerous applications in which both the left and right tails are of significance. Although left unexplored, we expect that the addition of a second iteration of superquantile estimation prior to density approximation could quite reasonably achieve this goal by simply applying the optimizations of section 2.3 on the negatives of observed values.

Quantification of density estimation accuracy remains an issue throughout this thesis. In cases where the underlying density is unknown (such as Chapters 4 and 5), we rely on other density approximations to reflect the "true" distribution. In some cases, these approximations inaccurately portray what we intuitively believe to be the true tail densities, as in the example of section 4.3. Additionally, because density values can range so dramatically between data sets, numerical errors should be scaled to properly reflect the relative, rather than absolute, impact of constraint implementation. Take, for instance, the extreme difference in density values between the data sets of Chapters 4 and 5. In light of these factors, we remain quite reliant on a qualitative evaluation of density approximation through the use of visual comparisons.

The choice between an MLP and MEP formulation for density estimation remains an ambiguous issue with no clear answer. Given the objective of each, one can reasonably conclude that MLP formulations will likely underestimate tail density due to a potential lack of observations from those regions, thus recommending quantile and superquantile constraint formulations that utilize lower bounds on tail densities in order to "push" weight into the tail regions. Likewise, one can envision using these same constraints as an upper bound within an MEP framework in order to "rein in" the tails. Although this objective/constraint relationship generally performs as intended, it is important to recognize an additional aspect distinguishing the two objective functions. For small samples with little additional soft information, we often cannot foresee issues of modality, non-symmetry, and other non-parametric characteristics. In these cases, because MEP formulations rely solely on constraints rather than the sample data itself, these distributional shape attributes can be overlooked. As in the hydrofoil example of Chapter 5, an MEP formulation would have failed to identify the dual-modality of the underlying distribution barring the explicit incorporation of a constraint requiring it.

With the ability to incorporate a flexible and adaptive method of constraint formulation, epi-spline optimizations enable analysts to uniquely modify the process proposed

here to their particular application of interest, and based on their total available knowledge. Given the limited data sets explored, we find that incorporation of accurate quantile and superquantile estimates can significantly enhance density estimation, particularly for those distributions with heavy tails. The key in this implementation remains in the accuracy of the quantile/superquantile estimates themselves, which, through more evolved constrained optimization can be enhanced without the need for additional samples. We expect further exploration of the methods and processes posed here to engender epi-spline optimization for more extensive usage.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX: Computations

A.1 Software Interface Algorithm

We utilize the **GDXRRW** utility suite [20] to provide an interface between R and the General Algebraic Modeling System (GAMS). We begin by initializing simulation parameters and indexes such as sample size, iteration number, and parameter values (i.e., λ, α). Passing these inputs from GAMS to R, we generate the random samples, compute sample statistics, and pass results back to GAMS for DSE and density optimizations. The results of these optimizations are then returned to R for plotting and error calculation. The general algorithm is as follows:

Algorithm 1: DSE and PDF Optimization

Results: $Q_X(p)$, $\bar{Q}_X(p)$, and f_x Estimation Errors
Initialize *Sets, Parameters*
foreach *Trial* **do**
  Generate X sample data for specified distribution;
  Calculate sample statistics;
  Input *soft information* for f_x per DSE scenario;
  Solve for DSE using *Maximum Curvature*
  Record a-coefficients and $Q_X(p)$, $\bar{Q}_X(p)$ predictions;
  Plot and record AAD, MAD errors;
  Apply quantile, superquantile constraints;
  Input *soft information* for f_x;
  if *Maximum Log-Likelihood* **then**
    Solve for $\hat{f}_x$ using MLP;
  else
    Solve for $\hat{f}_x$ using MEP;
  end
  Record b-coefficients;
  Plot and record SSE, SSTE errors;
end

A.2 Computational Time

We performed all statistical and optimization calculations on a 2010 PC laptop with a 2.53 GHZ Intel Core i5 processor and 4GB RAM. Statistical computations and error calculations were performed using R, while optimization were performed using GAMS. DSE optimization was done using the CONOPT solver, with 100 iterations taking less than 4 minutes to perform 100 iterations in all scenarios. PDF optimization was performed using either the CONOPT or COINIPOPT solver depending on the complexity of constraint configurations. Numerical experimentation using 100 iterations as in Chapter 3 took less than 10 minutes in all cases for optimizations, with less than 5 minutes required in most cases.

A.3 Bootstrapped Confidence Intervals

From our original sample of X of size n, we perform $B = 1,000$ bootstrap iterations in which we sample from X (with replacement) so that each bootstrapped sample is also of size n. Calculating statistics of interest on these bootstrapped sample provides 1,000 values per statistic of interest. Sorting these statistics allows us to identify the corresponding 25th and 975th largest statistics, $s(x^{*b})$, of the distribution of B. These values then correspond to an approximate 95% confidence interval for the statistic s of the original sample X.

Algorithm 2: Bootstrapping Sample Statistics

Result: 95% CI for $s(x)$
Generate X sample data for specified distribution;
foreach *Iteration* in $1 : B = 1,000$ **do**
  Generate bootstrap sample: x^{*b} = Sample(X, n, replace = T);
  Calculate statistic: $s(x^{*b})$
Sort bootstrapped statistics
return 250th and 975th largest $s(x^{*b})$

This relatively straightforward method of bootstrapping is called *quantile-based intervals*, and though simple, is often affected by bias and tends to provide less than the nominal coverage. For a more in-depth explanation of bootstrapping, see [16].

A.4 Non-Parametric Binomial Confidence Intervals

The use of binomial confidence intervals for quantile estimation is well-known. For a more in-depth explanation on the validity of this technique, see [21]. Here, we use the binomial approximation for estimating the median and quartile values of small samples, using a 95% confidence interval, $[X_r, X_s]$ where integer values r and s are the indexes of the sorted X determined as

$$r = \text{largest } r \quad s.t. \quad \sum_{i=1}^{r} \binom{n}{i} p^i (1-p)^{n-i} \leq 0.025,$$

$$s = \text{smallest } s \quad s.t. \quad \sum_{i=1}^{s} \binom{n}{i} p^i (1-p)^{n-1} \geq 0.975.$$

A.5 Quantile and Superquantile Constraint Calculation

Provided their central importance in constraint formulations, details regarding both quantile and superquantile constraint derivations are provided.

$$\text{Quantile: } \int_{Q(p)}^{\infty} f(x)dx = 1 - p$$

$$\text{Superquantile: } \frac{1}{1-p} \int_{Q(p)}^{\infty} x f(x)dx = \bar{Q}_X(p)$$

Using Simpson's Rule, which provides exact approximations for third order and below polynomials, we can simplify the integrals as follows.

$$\begin{aligned} Q_X(p) : \int_{Q(p)}^{\infty} f(x)dx &= \sum_{a,\, b} \int_{a}^{b} f(x)dx \approx \sum_{a,\, b} \frac{b-a}{6} \left[f(a) + f(b) + 4f\left(\frac{a+b}{2}\right) \right] \\ &= \sum_{m} \frac{\Delta x}{6} \left[(b_0^m + b_1^m x_L^m) + (b_0^m + b_1^m x_R^m) + 4\left(b_0^m + b_1^m \left(\frac{x_L^m + x_R^m}{2} \right) \right) \right] \\ &= \sum_{m} \Delta x \left[b_0^m + b_1^m \left(\frac{x_L^m + x_R^m}{2} \right) \right] \end{aligned}$$

$$\bar{Q}_X(p) : \frac{1}{1-p}\int_{Q(p)}^{\infty} xf(x)dx = \sum_{a,\ b}\int_{a}^{b} f(x)dx \approx \sum_{a,\ b}\frac{b-a}{6}\left[f(a)+f(b)+4f\left(\frac{a+b}{2}\right)\right]$$

$$= \frac{1}{1-p}\sum_{m}\frac{\Delta x}{6}\Bigg[\left(b_0 x_L^m + b_1^m (x_L^m)^2\right) + \left(b_0^m x_R^m + b_1^m (x_R^m)^2\right)$$

$$+ 4\left(b_0^m\left(\frac{x_L^m + x_R^m}{2}\right) + b_1^m\left(\frac{x_L^m + x_R^m}{2}\right)^2\right)\Bigg]$$

$$= \frac{1}{1-p}\sum_{m}\frac{\Delta x}{6}\left[3b_0^m(x_L^m + x_R^m) + b_1^m\left((x_R^m)^2 + (x_L^m)^2\right) + b_1^m(x_R^m + x_L^m)^2\right]$$

For the quantile constraints we sum over all m such that x_L^m is greater than $Q(p)$. This allows us to achieve an arbitrarily close approximation to the true integrals as our mesh resolution approaches zero.

A.6 Chapters 3 Additional Results

Table A.1: SSE and SSTE for Exponential Benchmark

Scenario	SSE	SSTE (0.80)	SSTE (0.90)	SSTE (0.95)
PDF 1a	1.272e-02	2.117e-03	1.114e-03	5.434e-04
PDF 1a*	1.180e-02	1.524e-03	8.362e-04	4.587e-04
PDF 2a	1.226e-02	2.037e-03	1.098e-03	5.480e-04
PDF 2a*	1.207e-02	1.662e-03	1.156e-03	8.138e-04
PDF 3a	7.179e-03	1.129e-03	5.155e-04	3.032e-04
PDF 3a*	6.849e-03	8.079e-04	4.785e-04	3.169e-04
PDF 4a	6.751e-03	1.050e-03	4.839e-04	2.854e-04
PDF 4a*	6.403e-03	8.063e-04	4.692e-04	2.703e-04
PDF 1b	2.832e-02	2.943e-03	2.661e-03	2.057e-03
PDF 1b*	7.924e-03	5.520e-04	2.460e-04	1.714e-04
PDF 2b	4.318e-03	9.375e-04	7.304e-04	6.809e-04
PDF 2b*	2.540e-03	5.610e-04	4.077e-04	2.388e-04
PDF 3b	2.056e-03	1.100e-03	2.623e-04	2.250e-04
PDF 3b*	1.480e-03	4.209e-04	1.312e-04	9.259e-05
PDF 4b	2.578e-05	< 1.0e-08	< 1.0e-10	< 1.0e-10
PDF 4b*	2.087e-05	2.070e-05	3.697e-06	4.462e-07

The sum of squared errors (SSE) and tail errors (SSTE) of first-order epi-spline estimates against the true exponential distribution evaluated at each mesh intersection. Notice the impact of both increased distributional knowledge and the effect of accurate quantile/superquantile estimates. Here the tails correspond to prediction values to the right of particular exponential quantiles according to

$$Q(0.80) = 8.05$$
$$Q(0.90) = 11.51$$
$$Q(0.95) = 14.98$$

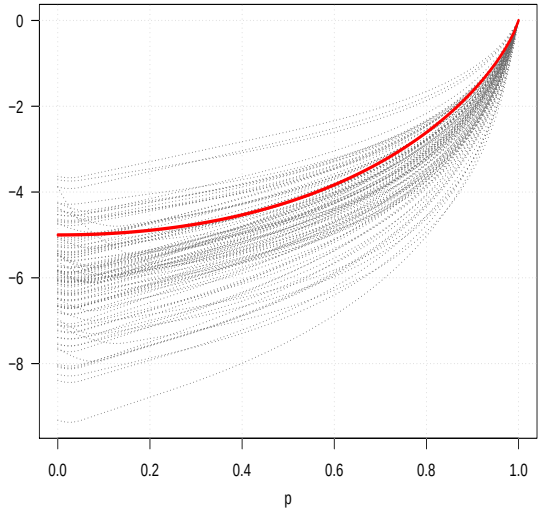
Table A.2: SSE and SSTE for Pareto Benchmark

Scenario	SSE	SSTE (0.80)	SSTE (0.90)	SSTE (0.95)
PDF 1a	2.532e-02	1.858e-03	6.454e-04	1.850e-04
PDF 1a*	2.514e-02	1.391e-03	5.196e-04	2.164e-04
PDF 2a	2.648e-02	1.914e-03	6.390e-04	1.878e-04
PDF 2a*	2.570e-02	1.599e-03	6.070e-04	2.041e-04
PDF 3a	1.474e-02	9.646e-04	3.452e-04	1.241e-04
PDF 3a*	1.469e-02	8.039e-04	3.601e-04	1.305e-04
PDF 4a	1.460e-02	9.121e-04	3.274e-04	1.163e-04
PDF 4a*	1.406e-02	6.856e-04	3.486e-04	1.284e-04
PDF 1b	1.830e-02	8.479e-04	3.381e-04	8.028e-05
PDF 1b*	1.675e-02	7.580e-04	2.917e-04	7.802e-05
PDF 2b	9.742e-03	4.146e-04	2.123e-04	8.107e-05
PDF 2b*	9.455e-03	4.896e-04	2.398e-04	6.041e-05
PDF 3b	1.739e-03	3.094e-04	9.100e-05	4.425e-05
PDF 3b*	1.991e-03	2.629e-04	9.746e-05	1.719e-05
PDF 4b	1.529e-03	1.352e-04	6.548e-05	1.328e-05
PDF 4b*	1.642e-03	1.566e-04	6.632e-05	7.796e-06

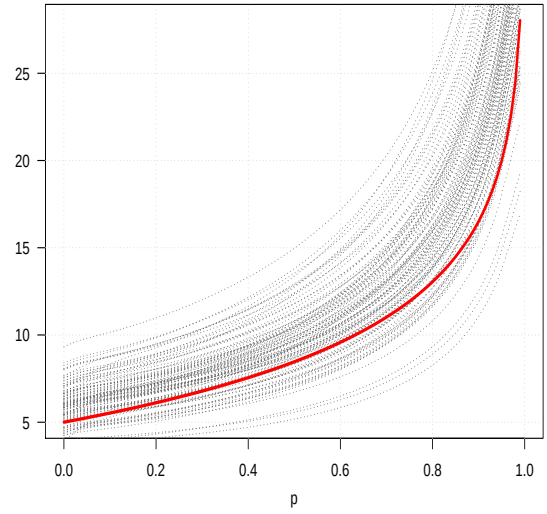
The sum of squared errors (SSE) and tail errors (SSTE) of first-order epi-spline estimates against the true Pareto distribution evaluated at each mesh intersection. Here the tails correspond to prediction values to the right of particular Pareto quantiles according to

$$Q(0.80) = 7.10$$
$$Q(0.90) = 11.51$$
$$Q(0.95) = 17.14$$

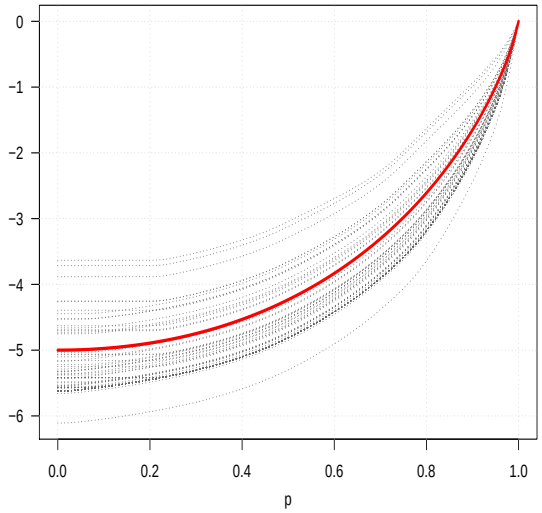
Figure A.1: DSE and Superquantile Estimates for Exponential Scenarios 1 and 3



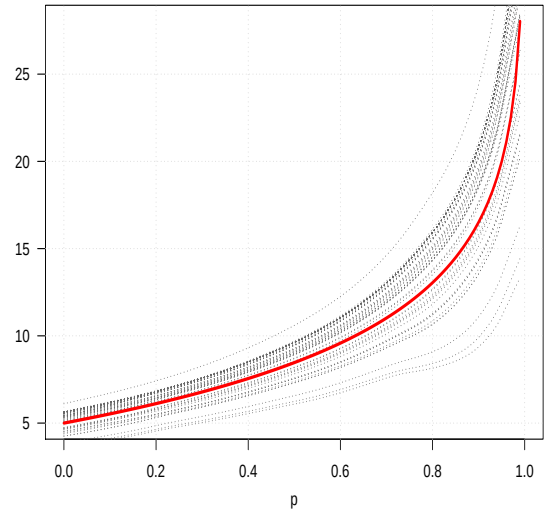
(a) Scenario 1 DSE Estimates



(b) Scenario 1 Superquantile Estimates



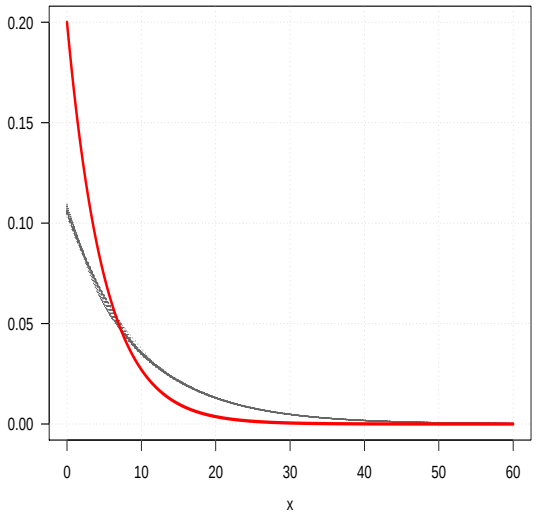
(c) Scenario 3 DSE Estimates



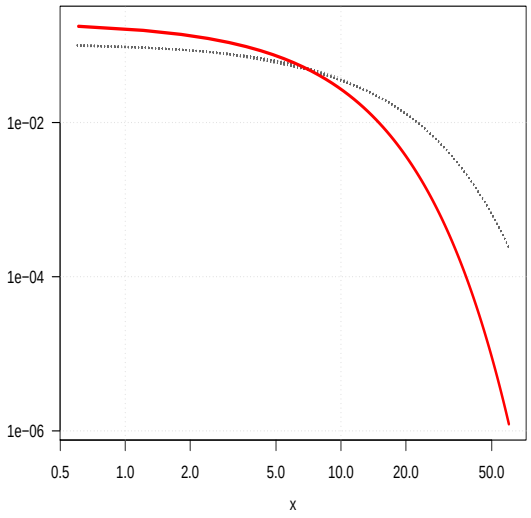
(d) Scenario 3 Superquantile Estimates

Dual of superexpectations (left) and superquantile (right) second-order epi-spline estimates for the exponential case using constraints via Scenario 1 (top) and Scenario 3 (bottom) and optimized for curvature. The true values are shown in solid red. Notice the generally conservative estimates made, regularly overestimating the superexpectations.

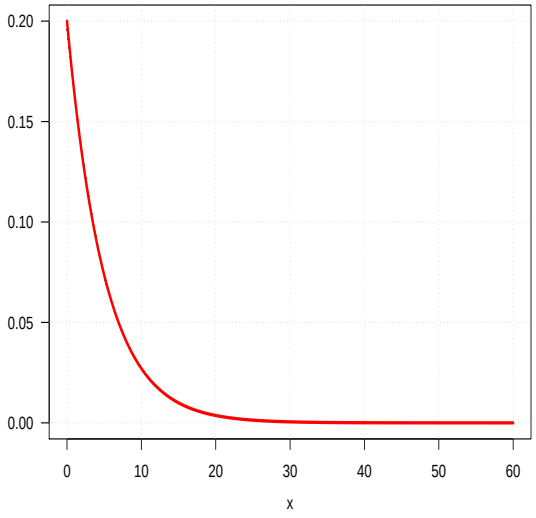
Figure A.2: PDF Estimates for Exponential Scenarios 1b and 4b



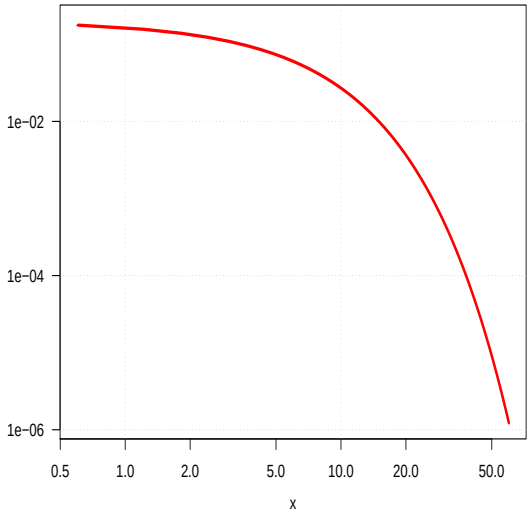
(a) Scenario 1b PDF



(b) Scenario 1a PDF Log-Scale



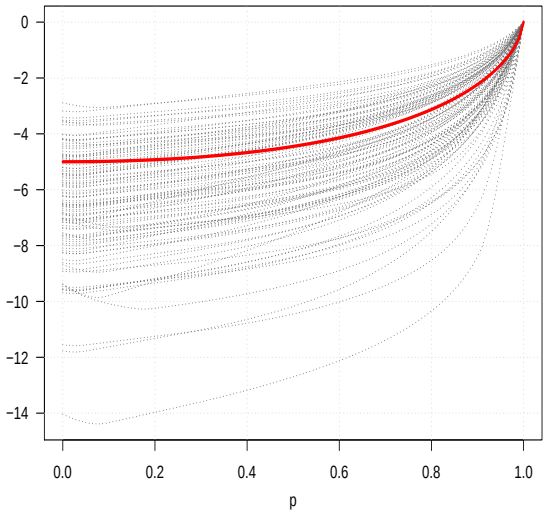
(c) Scenario 4b PDF



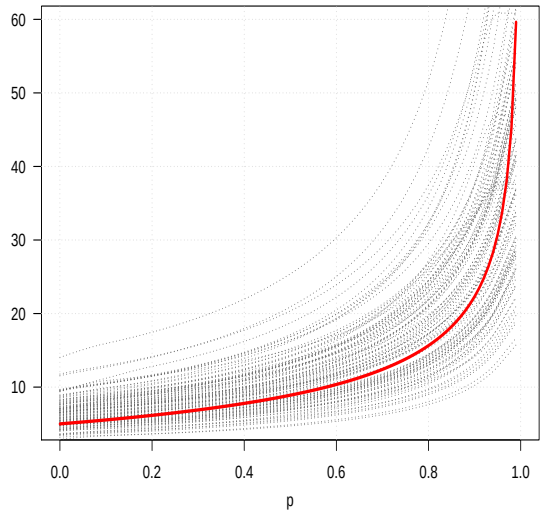
(d) Scenario 4b PDF Log-Scale

Density estimates (left), also shown on a log-scale (right), for first-order epi-spline estimates for Scenarios 1b (top) and 4b (bottom) of the Exponential benchmark. The true values are shown in solid red. Increased knowledge between scenarios results in more accurate density estimates. For Scenario 4b, results are so accurate that epi-spline estimates are no longer visible.

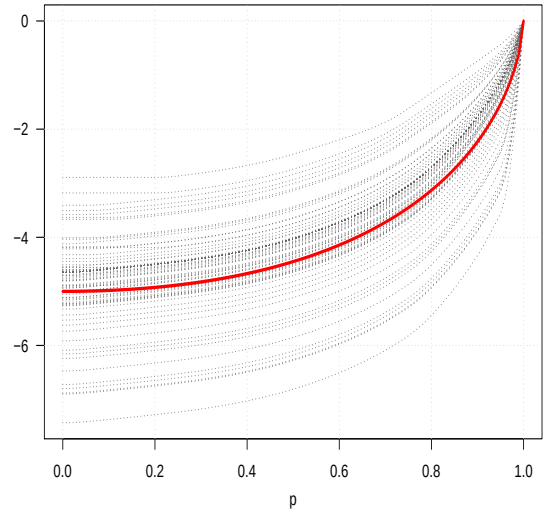
Figure A.3: DSE and Superquantile Estimates for Pareto Scenarios 1 and 3



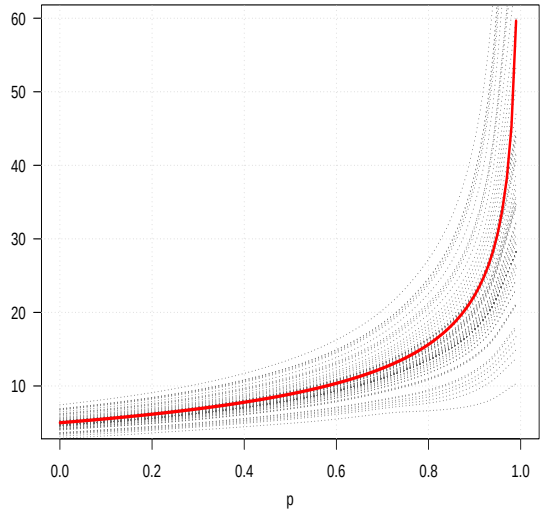
(a) Scenario 1 DSE Estimates



(b) Scenario 1 Superquantile Estimates



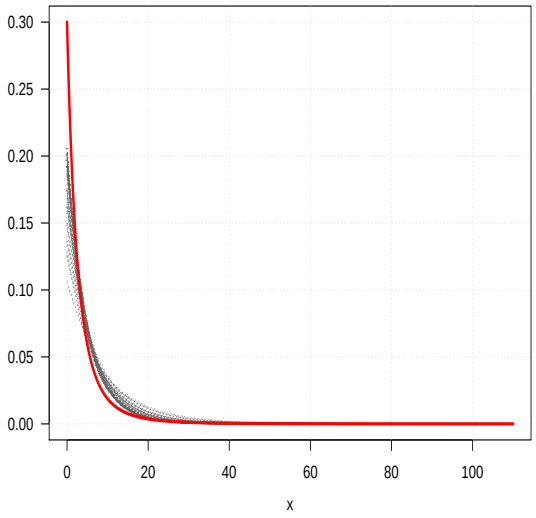
(c) Scenario 3 DSE Estimates



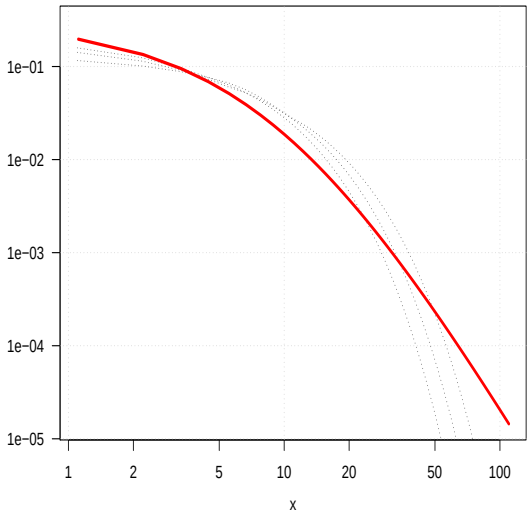
(d) Scenario 3 Superquantile Estimates

Dual of superexpectations (left) and superquantile (right) second-order epi-spline estimates for the Pareto case using constraints via Scenario 1 (top) and Scenario 3 (bottom) and optimized for curvature. The true values are shown in solid red.

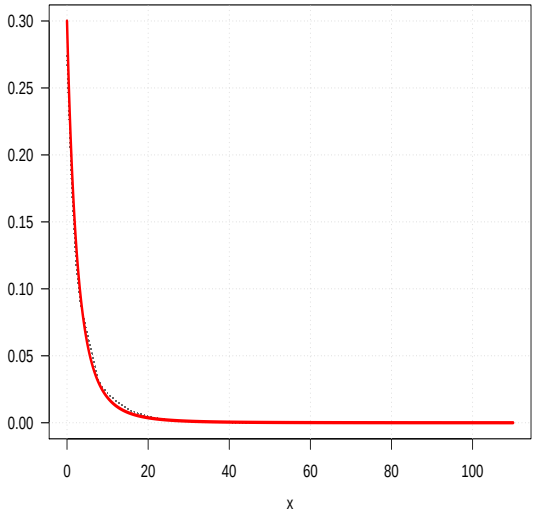
Figure A.4: PDF Estimates for Pareto Scenarios 1b and 4b



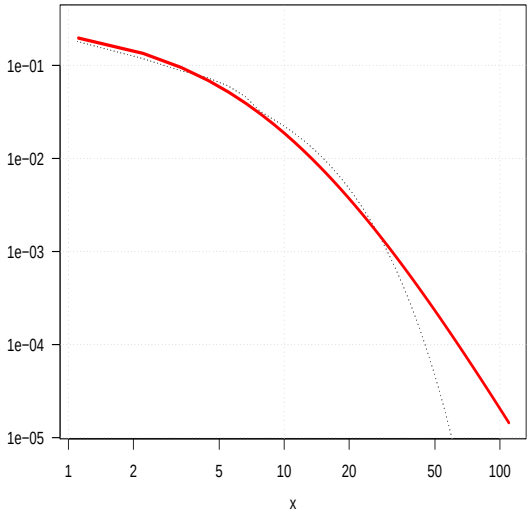
(a) Scenario 1b PDF



(b) Scenario 1a PDF Log-Scale



(c) Scenario 4b PDF



(d) Scenario 4b PDF Log-Scale

Density estimates (left), also shown on a log-scale (right), for first-order epi-spline estimates for Scenarios 1b (top) and 4b (bottom) of the Pareto benchmark. The true values are shown in solid red.

List of References

[1] I. J. Myung, "Tutorial on maximum likelihood estimation," *Journal of mathematical Psychology*, vol. 47, no. 1, pp. 90–100, 2003.

[2] N. S. Altman, "Kernel smoothing of data with correlated errors," *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 749–759, 1990. Available: http://www.tandfonline.com/doi/abs/10.1080/01621459.1990.10474936

[3] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004. Available: http://dx.doi.org/10.1023/B:STCO.0000035301.49549.88

[4] H. White, "Maximum likelihood estimation of misspecified models," *Econometrica*, vol. 50, no. 1, pp. 1–25, 1982. Available: http://www.jstor.org/stable/1912526

[5] L. Peng, "Estimating the mean of a heavy tailed distribution," *Statistics & Probability Letters*, vol. 52, no. 3, pp. 255–264, APR 15 2001.

[6] D. P. K. Søren Asmussen, "Improved algorithms for rare event simulation with heavy tails," *Advances in Applied Probability*, vol. 38, no. 2, pp. 545–558, 2006. Available: http://www.jstor.org/stable/20443455

[7] R. Feldman and M. Taqqu, *A practical guide to heavy tails: statistical techniques and applications*. Springer Science & Business Media, 1998.

[8] K. Dowd, *Measuring market risk*. John Wiley & Sons, 2007.

[9] N. Champagnat, M. Deaconu, A. Lejay, N. Navet, and S. Boukherouaa, "An empirical analysis of heavy-tails behavior of financial data: The case for power laws," July 2013, working paper or preprint. Available: https://hal.inria.fr/hal-00851429

[10] V. Pisarenko and M. Rodkin, *Heavy-tailed distributions in disaster analysis*. Springer Science & Business Media, 2010, vol. 30.

[11] R. Rockafellar and J. Royset, "Superquantiles and their applications to risk, random variables, and regression," in *Tutorials in Operations Research* (INFORMS). INFORMS, 2013, pp. 151–167. Available: http://dx.doi.org/10.1287/educ.2013.0111

[12] R. T. Rockafellar and J. O. Royset, "Random variables, monotone relations, and convex analysis," *Mathematical Programming*, vol. 148, no. 1, pp. 297–331, 2014. Available: http://dx.doi.org/10.1007/s10107-014-0801-1

[13] J. O. Royset and R. J.-B. Wets, *From Data to Assessments and Decisions: Epi-Spline Technology*. INFORMS, 2014, ch. 3, pp. 27–53. Available: http://pubsonline.informs.org/doi/abs/10.1287/educ.2014.0126

[14] J. C. Carbaugh, "Density deconvolution with epi-splines," Master's thesis, Dept. of Operations Research, Naval Postgraduate School, Monterey, California, 2015.

[15] J. O. Royset and R. J. Wets, "Fusion of hard and soft information in nonparametric density estimation," *European Journal of Operational Research*, vol. 247, no. 2, pp. 532–547, 2015.

[16] P. Sprent and N. C. Smeeton, *Applied nonparametric statistical methods*. CRC Press, 2007.

[17] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. Available: https://www.R-project.org/

[18] H. Deng, "Density estimation packages in r," in *The R User Conference, useR! 2011 August 16-18 2011 University of Warwick, Coventry, UK*. Citeseer, 2011, p. 120.

[19] J. R. P. Perdikaris, D. Venturi and G. Karniadakis, "Multi-fidelity modeling via recursive co-kriging and gaussian markov random fields," in *Proceedings IEEE of the Royal Society (MEMS'97)*, Nagoya, Japan, Jan. 1997, pp. 290–294.

[20] O. coding by Rishabh Jain. Adopted, packaged, and extended by Steve Dirkse., *gdxrrw: An interface between GAMS and R*, 2014, r package version 0.4.0.

[21] W. J. Conover and W. Conover, *Practical nonparametric statistics*. Wiley New York, 1980.

Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California