ABSTRACT

Title of dissertation: SEMIPARAMETRIC METHODS IN THE ESTIMATION OF TAIL PROBABILITIES AND EXTREME QUANTILES

Lemeng Pan, Doctor of Philosophy, 2016

Dissertation directed by: Professor Benjamin Kedem
Department of Mathematics

In quantitative risk analysis, the problem of estimating small threshold exceedance probabilities and extreme quantiles arise ubiquitously in bio-surveillance, economics, natural disaster insurance actuary, quality control schemes, etc. A useful way to make an assessment of extreme events is to estimate the probabilities of exceeding large threshold values and extreme quantiles judged by interested authorities. Such information regarding extremes serves as essential guidance to interested authorities in decision making processes. However, in such a context, data are usually skewed in nature, and the rarity of exceedance of large threshold implies large fluctuations in the distribution's upper tail, precisely where the accuracy is desired mostly. Extreme Value Theory (EVT) is a branch of statistics that characterizes the behavior of upper or lower tails of probability distributions. However, existing methods in EVT for the estimation of small threshold exceedance probabilities and extreme quantiles often lead to poor predictive performance in cases where the underlying sample is not large enough or does not contain values in the distribution's

tail. In this dissertation, we shall be concerned with an *out of sample* semiparametric (SP) method for the estimation of small threshold probabilities and extreme quantiles. The proposed SP method for interval estimation calls for the fusion or integration of a given data sample with external computer generated independent samples. Since more data are used, real as well as artificial, under certain conditions the method produces relatively short yet reliable confidence intervals for small exceedance probabilities and extreme quantiles.

This dissertation is organized as follows: In Chapter One, an overview of Extreme Value Theory will be given, and the existing methods for exceedance probability and extreme quantile estimation in EVT will be presented in some detail. Chapter Two introduces some necessary background about the Density Ratio Model. In Chapter Three, the idea of out of sample fusion (OSF) and repeated out of sample fusion (ROSF) are reviewed. We will show how to estimate tail probabilities and construct confidence intervals through OSF and ROSF. Results from extensive simulation studies are presented to demonstrate the performance of the proposed model when the underlying sample is from a highly skewed distribution. The results are compared with those obtained by EVT, and other well known methods. In Chapter Four, how extreme quantiles are estimated based on ROSF is presented with results from simulation studies. In Chapter Five, applications of the proposed method to real data problems in food safety and a clinical trial will be given. Finally, asymptotic theorems and results for quantiles under the density ratio model appear in the appendix.

SEMIPARAMETRIC METHODS IN THE ESTIMATION OF TAIL PROBABILITIES AND EXTREME QUANTILES

by

Lemeng Pan

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2016

Advisory Committee:
Professor Benjamin Kedem, Chair/Advisor
Professor Tingni Sun
Professor Myron Katzoff
Professor Frank Alt
Professor Jing Zhang

Dedication

To all my family

Acknowledgments

I owe my gratitude to all the people who have made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever.

First and foremost, I would like to extend my sincere gratitude to my adviser, Professor Benjamin Kedem for giving me an invaluable opportunity to work on challenging and extremely interesting projects over the past few years. I am also deeply grateful for his instructive advice and useful suggestions on my thesis. Professor Kedem is not only my teacher of statistics but also a mentor for my life. It has been a great pleasure to work with and learn from such a wise individual.

I am also deeply indebted to Professor Tingni Sun, Professor Frank Alt, Dr. Myron Katzoff and Professor Jing Zhang for agreeing to serve on my thesis committee and for sparing their invaluable time reviewing the manuscript. Their expertise and insight are very valuable and greatly enriched my work.

I would also like to thank all professors that have helped me in my graduate studies. Dr. Paul Smith taught me applied statistics and statistical modeling. Dr. Erid Slud guided me on the ICES imputation contest. Dr. Kagan has taught me mathematical statistics and multivariate analysis which helped me in understanding many fundamental questions in statistics.

I would also like to acknowledge Ms. Celeste Regaldo, and Mr. William Schildknecht for their help and support on administrative issues.

I would like to acknowledge the financial support received from the USDA, and the Department of Mathematics for all the projects discussed herein.

Lastly, I would like to thank all my family and friends. I owe my deepest thanks to my family for their love consideration and great confidence in me through all these years. Words cannot express the gratitude I owe them. My friends and fellow classmates at the Mathematics Department have enriched my graduate life in many ways and deserve a special mention. I would like to thank Xuan Yao, Zi Ding, Bin Han, Cheng Jie and Jinghang Xue for all their help during my course of study at Maryland.

It is impossible to remember all, and I apologize to those I've inadvertently left out.

Table of Contents

List of Figures

List of Abbreviations

AC	Agresti-Coull Method
BM	Block Maxima
DRM	Density Ratio Model
EP	Empirical Method
EVD	Extreme Value Distribution
EVT	Extreme Value Theory
GEV	Generalized Extreme Value Distribution
GHD	Generalized Hyperbolic Distribution
GPD	Generalized Pareto Distribution
IG	Inverse Gaussian
MAE	Mean Absolute Error
OSF	Out of Sample Fusion
POT	Peaks Over Threshold
ROSF	Repeated Out of Sample Fusion
SP	Semiparametric Method
VaR	Value at Risk

Chapter 1: Extreme Value Theory

1.1 Introduction

The estimation of the probability of rare and hazardous events is of interest in many disciplines, including environmental studies, finance modeling, engineering and earth sciences etc. Extreme Value Theory (EVT) is a branch of statistics that characterizes the behavior of upper or lower tails of probability distributions. Given a sample from a distribution, EVT seeks to model large deviations far away from the median.

This chapter briefly reviews the theoretical underpinnings of EVT. Three classical methods in modeling extreme values will be covered: the block maxima approach, the peaks over threshold approach, and Poisson processes. This introduction is by no means exhaustive, and its purpose is just to review traditional methods in estimation of tail probabilities and extreme quantiles which will later serve as benchmarks to assess the performance of the semiparametric methods. For more rigorous and thorough treatment of EVT, the reader is referred to Beirlant *et al.* (2004 [3], Coles (2001) [10], Haan and Ferreira (2006) [24], Leadbetter *et al.* (1983) [33], and Resnick (1987) [42].

Section 1.2 provides model formulation and inference. The Block Maxima approach is introduced in Section 1.3 and threshold models are given in Section 1.4. The notations are adopted from Coles (2001) [10].

1.2 Model Formulation

Consider a sequence of independent and identically distribution (i.i.d.) random variables $X_1, \dots, X_n$, the extreme value model focuses on the statistical behavior of

$$M_n = \max\{X_1, \dots, X_n\},$$

which is the maximum of the sequence of random variables. Determining which distribution M_n follows is the essential problem in EVT. In real applications, the sequence X_i's could represent either independent measurements of certain quantities on different individuals in the sample or values of a process measured over a time span. For example, in our food safety application, X_i's represent lead intake levels of 3000 Americans resulting from consuming seafood, while in a financial data application the X_i's may represent daily log-returns.

Theoretically, the distribution of M_n could be derived exactly, given that the distribution function F of X_i is known:

$$\mathrm{P}(M_n \leq z) = \mathrm{P}(X_1 \leq z, \dots X_n \leq z) = \mathrm{P}(X_1 \leq z) \times \cdots \times \mathrm{P}(X_n \leq z) = (F(z))^n. \tag{1.1}$$

In practice, however this approach is not feasible for the following reasons. First, the distribution function F is unknown in general. The problem might be overcame by using standard statistical techniques. One possibility would be estimating F by a kernel density estimate, the other would be assuming that X_i's are coming from a particular distribution. Then the estimated F needs to be raised to the power of n to obtain the distribution function of M_n. Small discrepancies in the estimates of F may lead to substantial discrepancies in F^n. Alternatively, a family of distribtuons F^n that approximate any unknown F may be found. In other words, the characteristics and asymptotic properties of F^n are needed. However, one difficulty could arise. Suppose z_+ is the smallest value of z such that $F(z) = 1$, or put differently, let z_+ be the upper end point of F. Then for any $z < z_+$, $F^n(z) \to 0$ as $n \to \infty$. In this case, the distribution of M_n degenerates to a point mass on z_+. To circumvent this problem, a linear transformation of the variable M_n is introduced:

$$M_n^* = \frac{M_n - b_n}{a_n}$$

where $a_n > 0$ and b_n are sequences of constants. Appropriate choices of a_n and b_n would prevent a probability mass collapse over a single point and stabilize the location and scale of M_n^* as n increases.

Theorem 1.1 (Fisher-Tippett-Gnedenko)**.** *Let X_n be a sequence of i.i.d. random variables. If there exist constants $a_n > 0, b_n \in \Re$ and some non-degenerate distribution function G such that*

$$\frac{M_n - b_n}{a_n} \xrightarrow{d} G, \tag{1.2}$$

then G belongs to one of the three standard extreme value distributions:

$$\begin{array}{llll}
\text{I} & \textit{F\'echet:} & \Phi(z) & = \exp\Big\{-\exp\Big[-\Big(\tfrac{z-b}{a}\Big)\Big]\Big\}, \quad -\infty < z < \infty; \\
\text{II} & \textit{Weibull:} & \Psi(z) & = \begin{cases} \exp\Big\{-\Big(\tfrac{z-b}{a}\Big)^{-\alpha}\Big\}, & z > b \\ 0, & z \leq b \end{cases} \\
\text{III} & \textit{Gumbel:} & \Lambda(z) & = \begin{cases} \exp\Big\{-\Big[-\Big(\tfrac{z-b}{a}\Big)^{\alpha}\Big]\Big\}, & z < b \\ 1, & z \geq b \end{cases}
\end{array}$$

for parameters $a > 0$, b, and in the case of families II and III, $\alpha > 0$.

This is the first EVT result (also known as the Fisher-Tippett-Gnedenko Theorem) which characterizes the asymptotic distribution of the sample maxima. The theorem states that the asymptotic distribution G of the maximum of a sample of i.i.d. random variables after proper renormalization can converge in distribution to only one of three possible distributions: Gumbel, Féchet, or Weibull. Collectively, these three classes of distributions are termed as the extreme value distribution (EVD).

Early applications of EVT are based on characterization of "maximum domains of attraction". By definition, a random variable X (the distribution function F of X) belongs to the maximum domain of attraction of the extreme value distribution G if there exists constants $a_n > 0, b_n$ such that 1.2 holds. It was usually assumed the underlying distribution F belongs to one of the three "maximum domains of attraction" (MDA) of extreme value distribution G ($F \in \mathcal{D}(G)$) and adopt the associated distribution family $\mathcal{D}(G)$. Then the relevant parameters of the EVD can be estimated, the quantiles or tail probabilities can be derived based on the estimators. In the estimation process, one selects the k largest order statistics to obtain a fraction of the sample which represents the distribution tail. Then estimators, depending on the number of upper order statistics, are used in the estimation. There are two issues related to the implementation of this approach. First, a technique is required to choose which of the three families is most appropriate for the data at hand. Second, once a decision is made, subsequent inference presumes this choice to be correct, and does not allow for the uncertainty which such a selection involves.

A reformulation of Theorem 1.1 combines the three distributions into a single family of models called the generalized extreme value (GEV) distribution. The modified version of the first theorem characterizes the asymptotic distribution of a series of maxima, and states that under certain conditions the distribution of the standardized maximum of the series is shown to converge to one of the Gumbel, Féchet, or Weibull distributions.

Theorem 1.2. *Let X_n be a sequence of i.i.d. random variables. If there exist constants $a_n > 0, b_n \in \Re$ and some non-degenerate distribution function G such that*

$$\frac{M_n - b_n}{a_n} \xrightarrow{d} G,$$

then G is a member of the GEV family:

$$G(z) = \exp\left\{ -\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi} \right\}$$

defined on $\{z : 1 + \xi(x-\mu)/\sigma > 0\}$ where $-\infty < \mu < \infty$ is the location parameter, $\sigma > 0$ the scale and $\xi \neq 0$ the shape parameter.

The above one parameter representation of the three standard cases into one family is also known as the Jenkinson-von Mises representation (Embrechts et al.) [14] which allows one to write the c.d.f. of G as a function $G_\gamma(x) \equiv \exp(-(1+\xi x)^{(-1/\xi)})$ depending on the *extreme value index* ξ. The limiting case $\xi \to 0$, corresponds to the Gumble distribution, the case $\xi > 0$ corresponds to the Féchet distribution and the case $\xi < 0$ to the Weibull distribution. The three types of distributions correspond to the different tail behaviors for the distribution of the original population. Gumbel is related to light-tailed distributions such as normal, gamma or exponential distributions; Féchet is related to heavy-tailed distributions such as Pareto, Cauchy or Student-distribution and Weibull is related to distributions with finite support such as Uniform and Beta. This result is very significant. The theorem states that M_n can be stabilized with properly chosen a_n and b_n, the corresponding transformed sample maxima $(M_n - b_n)/a_n$ converges to a variable having a distribution within the generalized extreme value (GEV) distribution families, regardless of the underlying distribution F of the population.

For an illustrative example on the choice of centering and normalizing constants a_n and b_n, suppose $X_1, \ldots, X_n$ is a sequence of exponential variables with parameter $\lambda = 1$. Then, for $a_n = 1$ and $b_n = \log n$, the limiting distribution of M_n is the Gumbel distribution as $n \to \infty$, as shown:

$$\begin{aligned} \mathrm{P}\left(\tfrac{M_n-b_n}{an}\right) &= F^n(z+\log n) \\ &= [1-\exp(-(z+\log n))]^n \\ &= [1-n^{-1}\exp(-z)]^n \\ &\to \exp(-\exp(-z)) \end{aligned}$$

The above calculation show the resulting distribution is indeed Gumbel, corresponding to $\xi \to 0$ in the GEV family. For a summary of maximum domains of attraction and derivation of normalizing constants a_n and b_n, the reader is referred to Embrechts et al. (1997) [14] Chapter 3.

Theorem 1.2 plays a fundamental role when modeling the maxima of random variables which is analogous to the Central Limit Theorem when modeling sums of random variables. Through inference of the shape parameter ξ, the most appropriate type of tail behavior is determined by the data themselves. This allows the risk modelers to avoid making a subjective selection about which individual extreme value family to adopt which may lead to substantial uncertainty. The unification of the original three families greatly simplifies statistical implementation and leads to the so called Block Maxima approach.

Figure 1.1 illustrates the shape of the probability density functions of the standard Féchet, Weibull, and Gumbel distributions.

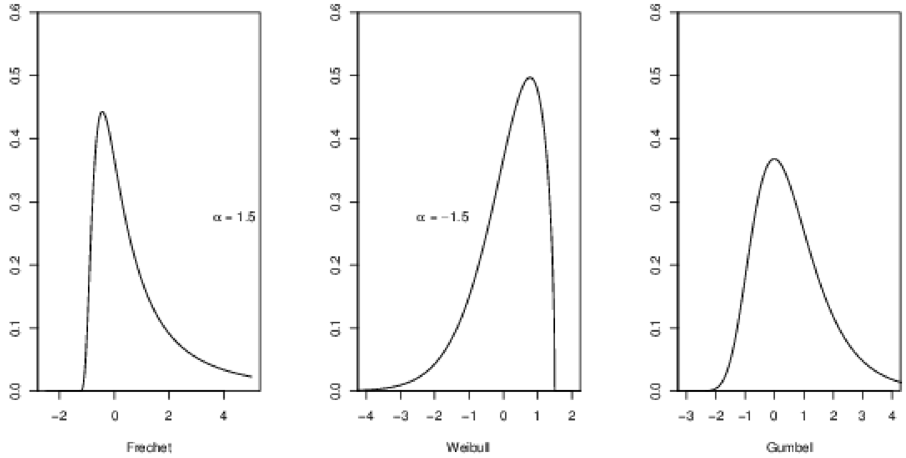

Figure 1.1: Density Plots for distributions from GEV Families

1.3 Block Maxima

The Block Maxima approach considers the maximum the variable takes in successive observations. More precisely, a sample is divided into sub-samples or blocks first. Then, the largest observations in each block (block maximum) are taken as extreme data points which will be used for fitting the GEV. An alternative version of this approach called rth largest order model is suggested to deal with cases when the underlying sample is not sufficiently large enough. In this alternative approach, not only the maximal observation in a given block is used as a data point for fitting the GEV, but the r largest observations are taken as well.

The block maxima and r largest order data selection methods are better illustrated in Figure 1.2. In the figure, 100 points that follows Gamma(1, 0.1) distribution are randomly generated and subdivided into 10 blocks of equal size. The

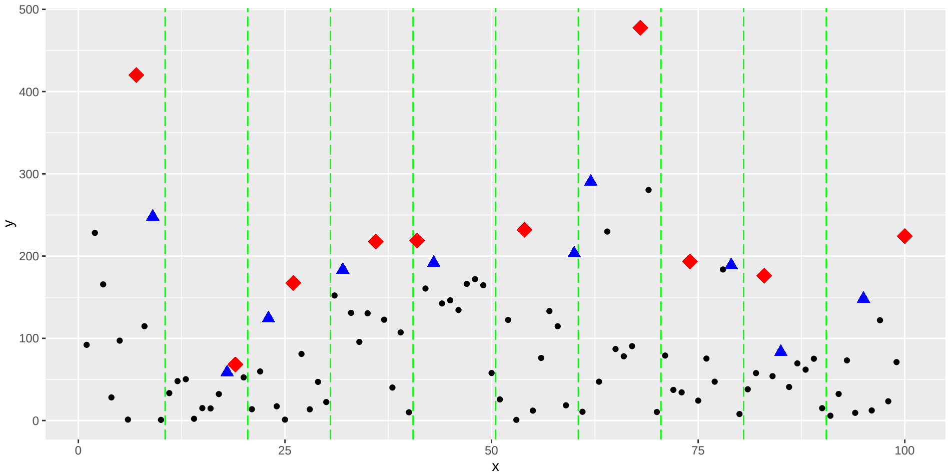

Figure 1.2: Block Maxima Approach Illustration

maximum observation in each block is marked by a red diamond, and the second largest observation is indicated by a blue triangle. The diamonds would be employed by the block maxima approach, and both diamonds and triangles would be utilized if one were to fit a second largest order model. Several issues arise when the block maxima approach is adopted in a real data application. First, in practical situations, it is very common that the sample size is not large enough, so that the unknown distribution parameters are subject to great uncertainty. The confidence intervals for the unknown distribution parameters and the derived risk measures would be too wide to make any practical sense. Of course, the rth largest order approach might be employed to reduce the variance by increasing the sample size. However, another problem is that the rth largest observations probably do not qualify as extreme events, and the inclusion of such points would lead to a biased sample. Furthermore, if a given sample is a time series, it is very natural that the series fluctuates more in volatile periods than it does in tranquil periods. The inclusion of data points during tranquil periods as block maxima values would also lead to substantial estimation bias.

1.3.1 Parameter Estimation

For notational convenience, let $Z_1, \ldots, Z_m$ be the block maxima when a sequence of random variables X_i are divided into m blocks. It is reasonable to assume that $Z_1, \ldots, Z_m$ are independent variables following the GEV distribution according to Theorem 1.2. Then, based on $Z_1, \ldots, Z_m$, the parameters (μ, σ, ξ) can be estimated by fitting the GEV distribution in a variety of ways. These include graphical techniques based on (versions of) probability plots, moment based techniques in which functions of model moments are set equal to their empirical equivalents, procedures in which the parameters are estimated as specified functions of order statistics and likelihood based techniques. Each technique has its pros and cons; however, in this dissertation we will focus on the likelihood based method due to its wide adaptability. The reader may refer to Beirlant et al. (2004) [3] for a detailed introduction on other estimation methods.

In general, maximum likelihood estimators of the parameters for the GEV distribution are obtainable. The the log-likelihood for the GEV parameters when $\xi \neq 0$ is

$$\ell(\mu, \sigma, \xi) = -m \log \sigma - (1+1/\xi) \sum_{i=1}^{m} \log \left[1 + \xi \left(\frac{z_i - \mu}{\sigma} \right) \right] - \sum_{i=1}^{m} \left[1 + \xi \left(\frac{z_i - \mu}{\sigma} \right) \right]^{-1/\xi}$$

given that $1 + \xi \left(\frac{z_i - \mu}{\sigma} \right) > 0, \quad \text{for} \quad i = 1, \ldots, m.$

When $\xi = 0$, the Gumbel limit of the GEV distribution leads to the log-likelihood of the following form

$$\ell(\mu, \sigma) = -m \log \sigma - \sum_{i=1}^{m} \left(\frac{z_i - \mu}{\sigma} \right) - \sum_{i=1}^{m} \exp \left[-\frac{z_i - \mu}{\sigma} \right]$$

Maximization of the log-likelihood function with respect to the parameters (μ, σ, ξ) yields the maximum likelihood estimates. Unfortunately, there is no closed form solution. For any given dataset, the maximization can be done by standard numerical optimization algorithms. The approximate distribution of the estimators $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ follows a multivariate normal with mean (μ, σ, ξ) and variance-covariance matrix equal to the inverse of the observed information matrix evaluated using the obtained maximum likelihood estimates.

The support of G depends on the unknown parameter values: $\mu - \sigma/\xi$ is an upper endpoint of the distribution when $\xi < 0$, and a lower endpoint when $\xi > 0$. This violates the regularity conditions so the standard asymptotic likelihood results do not hold. Smith (1985) [43] studied this problem in depth and concludes the following:

$$\sqrt{m}\big((\hat{\mu}, \hat{\sigma}, \hat{\xi}) - (\mu, \sigma, \xi) \xrightarrow{d} N(0, V)\big), \qquad \xi > -0.5$$

where V is the inverse of the Fisher information matrix. In words, when $\xi > -0.5$, the usual properties of consistency, asymptotic efficiency and asymptotic normality hold.

1.3.2 Tail Probability and Extreme Quantile Estimation

The main focus of this dissertation is the estimation of threshold exceeding probabilities and extreme quantiles. Substituting the parameters (μ, σ, ξ) by their estimates $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$, we can obtain the estimated tail probability:

$$\hat{p}_z = \hat{\mathrm{P}}(Z > z) = 1 - \hat{G}(z) = 1 - \exp\Big\{ - \Big[1 + \hat{\xi}\Big(\frac{z - \hat{\mu}}{\hat{\sigma}}\Big)\Big]^{-1/\hat{\xi}}\Big\},$$

as well as the p quantile:

$$\hat{z}_p = \hat{G}^{-1}(p) = \begin{cases} \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}}[y_p^{-\hat{\xi}} - 1], & \xi \neq 0 \\ \\ \hat{\mu} + \hat{\sigma} \log y_p, & \xi = 0 \end{cases}$$

where $y_p = -\log(p)$.

In this dissertation, z_p is defined as the p quantile for the block maxima data, meaning that $G(z_p) = p$. In EVT literature, z_p is commonly referred to as the return level associated with a return period. For the block maxima approach, the return period is $1/(1-p)$. In other words, the level z_p is expected to be exceeded on average once every $1/(1-p)$ blocks with probability $1-p$.

Let q_p denote the p quantile of the original data. In many applications, an estimation of this extreme quantile of the original data is desired. Suppose that the original sample of size n is divided into m blocks, then q_p corresponds to the $m/((1-p)n)$ blocks return level or the $[m-n(1-p)]/m$ quantile of the block maxima data.

Standard errors and confidence intervals for the exceedance probability p and the pth quantile z_p can be derived by the delta method,

$$\begin{aligned}\mathrm{Var}(\hat{p}_z) &\approx \nabla p_z^T V \nabla p_z \\ \mathrm{Var}(\hat{z}_p) &\approx \nabla z_p^T V \nabla z_p\end{aligned} \tag{1.3}$$

where

$$\begin{aligned}\nabla p_z^T &= \left[\frac{\partial p_z}{\partial \mu}, \frac{\partial p_z}{\partial \sigma}, \frac{\partial p_z}{\partial \xi}\right] \\ &= \Big[(1/\sigma)(1-p_z)(\xi(z-\mu)/\sigma+1)^{-(1+\xi)/\xi}, \\ &\qquad (z-\mu)/\sigma^2(1-p_z)(\xi(z-\mu)/\sigma+1)^{-(1+\xi)/\xi}, \\ &\qquad \frac{(1-p_z)\log(1-p_z)((\xi(z-\mu)+\sigma)\log(\xi(z-\mu)/\sigma+1)+\xi(\mu-z))}{\xi^2(\xi(z-\mu)+\sigma)}, \\ \nabla z_p^T &= \left[\frac{\partial z_p}{\partial \mu}, \frac{\partial z_p}{\partial \sigma}, \frac{\partial z_p}{\partial \xi}\right] \\ &= \Big[1, \quad \xi^{-1}(y_p^{-xi}-1), \quad \sigma\xi^{-2}(y_p^{-\xi}-1)+\sigma\xi^{-1}y_p^{-xi}\log y_p\Big],\end{aligned}$$

evaluated at $(\hat{\sigma}, \hat{\xi}, \hat{\zeta}_u)$. Coles (2001) [10] suggested caution is required in the interpretation of tail probabilities and quantile inferences using EVT, especially for extreme tail probabilities and quantiles for few reasons. First, the normal approximation to the distribution of the maximum likelihood estimators could be poor and better approximations are generally obtained from appropriate profile likelihood functions. More fundamentally, estimates and their measures of precision are based on the assumption that the EVT model is valid. Though the EVT models are supported by mathematical argument, the inference for small tail probability or large quantile requires extrapolation based on unverifiable assumptions.

1.3.3 Model Diagnostics

When EVT is applied to tail probabilities and extreme quantile estimation, the analysis is essentially an extrapolation outside of the sample. Although EVT is well supported by mathematical argument, it is not possible to check the validity of an extrapolation based on a GEV model in general. However, assessment can be made with reference to the observed data through several graphical means: probability plots, quantile plots, return level plots and density plots. These are not sufficient to justify extrapolation, but may serve as model diagnostic tools to check the quality of a fitted GEV model. Let $z_{(1)} \leq \cdots \leq z_{(m)}$ denote ordered block maxima data, $\hat{G}$ be the estimated GEV distribution function, and $\tilde{G}(z_{(i)}) = i/(m+1)$ be the empirical distribution function evaluated at $z_{(i)}$.

If the GEV fit works reasonably well, then $\hat{G}(z_{(i)}) \approx \tilde{G}(z_{(i)})$ for each i. The probability plot consists of the pairs

$$(\hat{G}(z_{(i)}), \tilde{G}(z_{(i)})), \qquad i = 1, \ldots, m$$

where $\hat{G}(z_{(i)}) = \exp\{-[1 + \hat{\xi}/\hat{\sigma}(z_{(i)} - \hat{\mu})]^{-1/\hat{\xi}}\}$, should lie close to the unit diagonal. Substantial departures from linearity indicates lack of goodness of fit in the GEV model.

Usually, the accuracy of the model for large values of z are of greatest interest. However, in the probability plot for extreme value models, $\hat{G}(z_{(i)})$ and $\tilde{G}(z_{(i)}))$ are both bound to approach 1 as $z_{(i)}$ increases. In other words, the probability plot provides the least information in the region of most interest. This deficiency is complemented by the quantile plot which consists of the pairs

$$(\hat{G}^{-1}(\frac{i}{m+1}), z_{(i)}), \qquad i = 1, \ldots, m$$

where

$$\hat{G}^{-1}(\frac{i}{m+1}) = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left[1 - \big\{-\log\big(\frac{i}{m+1}\big)\big\}^{-\hat{\xi}}\right].$$

Departures from linearity indicates lack of goodness-of-fit of the GEV model. If the GEV fit is reasonable for modeling the block maxima, then both the probability and quantile plot should consist of points that are approximately linear.

A return level plot consists of the locus of points $(1/(1-p), \hat{z}_p)$ for small values of p, where $\hat{z}_p$ is the estimated $(1/(1-p)$-block return level. It is usual to plot the return level curve on a logarithmic scale to reflect the effect of extrapolation. Confidence bounds and empirical estimates of the return levels also help to make goodness of fit judgments when added to the plot.

For completeness, the density plot of the fitted GEV can be compared against the histogram of the block maxima data.

1.4 Peaks Over Threshold

The peaks over threshold (POT) method is an alternative approach that ameliorates the above mentioned issues to some extent. It considers all observations above a certain threshold value as extreme observations as illustrated in Figure 1.3. Again, the figure shows 100 randomly generated points from Gamma(1, 0.1) distribution. The threshold $u = 200$ is represented by a horizontal dashed line in the plot; all points above this threshold which are marked as diamonds would be considered as extreme observations for the POT method. Confidence intervals and inference in other forms follow from the approximate normality of the estimators.

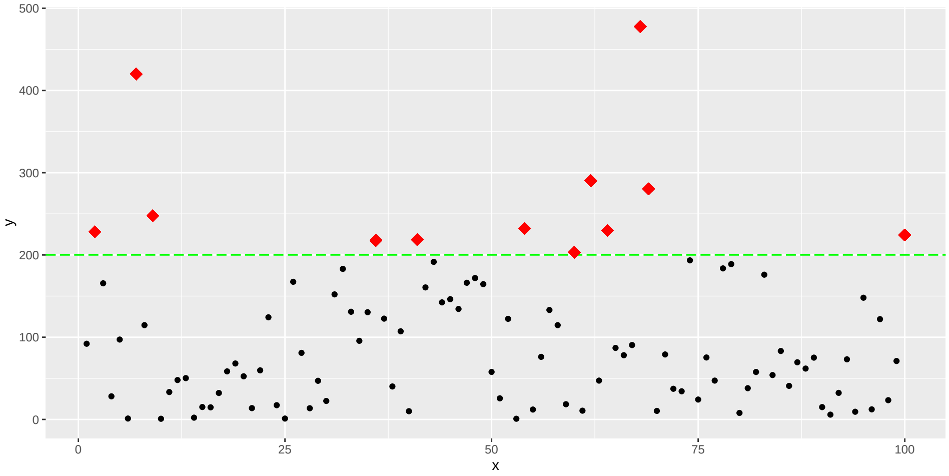

Figure 1.3: Peaks Over Threshold

The conditional distribution functions of values of x above the threshold u is denoted as F_u. How to estimate this conditional excess distribution function is a question of interest. By definition, F_u can be written in the following form:

$$\begin{aligned} F_u(y) &= \mathrm{P}(X - u \leq y | X > u), \qquad y \geq 0 \\ &= \tfrac{F(u+y)-F(u)}{1-F(u)} \\ &= \tfrac{F(x)-F(u)}{1-F(u)} \end{aligned}$$

The second EVT result (Picklands-Baikema-de Haan theorem) provides a very helpful theoretical results that help the modeling of the conditional excess distribution.

It is needed for the Peaks Over Threshold approach. The theorem states the following:

Theorem 1.3 (Picklands-Baikema-de Haan)**.** *Let X_n be a sequence of i.i.d. random variables with common distribution function F and let*

$$M_n = \max(X_1, \dots, X_n).$$

Suppose that F satisfies Theorem 1.1, so that for large n, $(M_n - b_n)/(a_n) \xrightarrow{d} G$, where

$$G(z) = \exp\Big\{ - \Big[1 + \xi\Big(\frac{z-\mu}{\sigma}\Big)\Big]^{-1/\xi}\Big\}$$

for some $\mu, \sigma > 0$. Then, for large enough u, the distribution function F_u of $X - u$, conditional on $X > u$, is approximately

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi}$$

defined on $\{y : y > 0, \text{and} \quad (1 + \xi y)/\tilde{\sigma} > 0\}$, where $\tilde{\sigma} = \sigma + \xi(u - \mu)$.

The family of distributions determined by H is called the generalized Pareto distribution (GPD). Theorem 1.3 states that, if the limiting distribution of block maxima approximates the GEV distribution G, then the threshold exceedances could be approximated by the generalized Pareto distribution for sufficiently large threshold u. The shape parameter ξ determines the tail behavior, the larger ξ, the heavier the tail. Furthermore, ξ is identical for both GEV and GPD which implies that ξ is invariant to block size. In summary, ξ plays the same role as it does for GEV: for $\xi < 0$ the distribution of exceedances possesses an upper bound of $u - \tilde{\sigma}/\xi$; for $\xi > 0$

the distribution is unbounded to the right. As $\xi \to 0$, the distribution function becomes

$$H(y) = 1 - \exp\left(-\frac{y}{\tilde{\sigma}}\right), \qquad y > 0$$

which corresponds to exponential distribution with parameter $1/\tilde{\sigma}$.

When POT approach is adopted in a practical application, it is necessary to properly choose the threshold u. If u is too small, a biased sample is obtained. Observations that do not qualify as extreme values would be included in the sample and violate the GPD approximation. On the other hand, if this value is chosen too large, the sample size would be too small leading to large estimation errors for the unknown distribution parameters. This is the extreme value version of the bias variance trade-off. A sufficiently large threshold can be determined, and a graphical tool called the mean residual life (MRL) plot may serve this purpose. Let Y denote excess of a threshold u_0 generated by X and Y follows a generalized Pareto distribution with parameters σ and ξ. Then the expected mean of Y equals $\sigma/(1-\xi)$ when $\xi < 1$ and infinite when $\xi \geq 1$. The plot is based on the expected value $E(Y) = \sigma/(1-\xi)$ of the GPD. Then we have

$$E(X - u_0 | X > u_0) = \frac{\sigma_{u_0}}{1-\xi},$$

given $\xi < 1$. In the above equation, σ_{u_0} denotes the scale parameter when the threshold is chosen to be u_0. If the GPD is valid as a model for excesses of the threshold u_0, it should equally be valid for all $u > u_0$.

$$E(X-u|X>u)=\frac{\sigma_u}{1-\xi}=\frac{\sigma_{u_0}+\xi u}{1-\xi}$$

Therefore, for a range of threshold values u, the conditional expected values are plotted against u. The above equation implies that at levels of u for which the generalized Pareto model is appropriate, the mean excess values should be linear with respect to u. Therefore, an appropriate value for u is given when the MRL plot of points $\{u, 1/n_u \sum_{i=1}^{n_u}(x_{(i)}-u)\}$ starts to become linear. The sign of the gradient in the linear part of the MRL plot also suggests the shape of the tail. A negative slope is associated with short-tailed distributions; a horizontal line (zero gradient) suggests exponential type tail; and a positive slope indicates a heavy-tailed distribution.

An illustrative MRL plot is shown in Figure 1.4. The underlying sample is randomly generated from a Normal(0,3) distribution with a sample of size 10000. The mean excess becomes linear for u ranging from 2 to 6. Therefore, a wide range of choices of u seem to be reasonable as suggested by the MRL plot. Empirical studies show that often the $50th$ to $70th$ percentile may serve as a suitable threshold.

1.4.1 Parameter Estimation

Once the threshold is determined, and the GPD is fitted to the reduced subsample consisting of threshold exceedances, the parameters of the GPD can be estimated by a variety of methods: the ML method, the method of probability

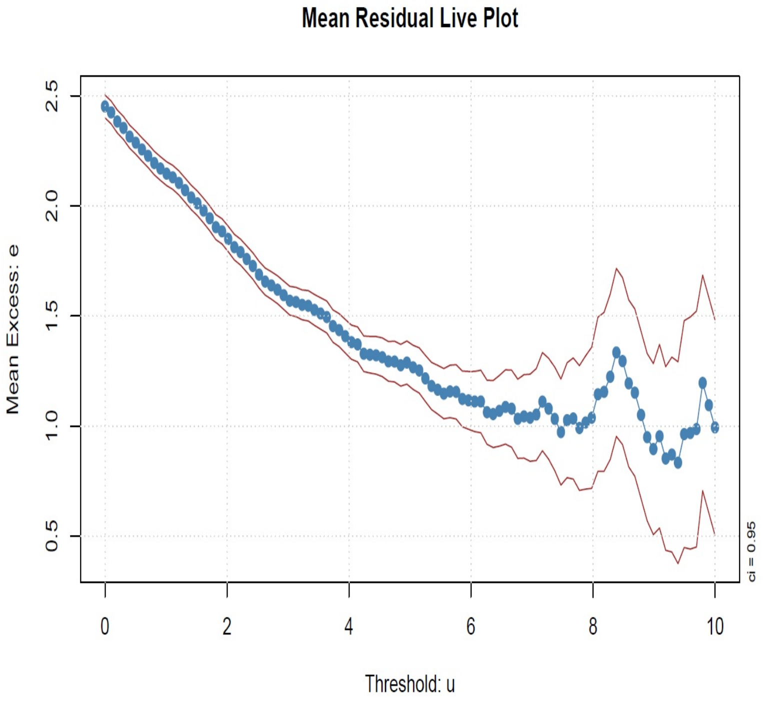

Figure 1.4: MRL Plot for Random Normal(0,3) Sample of Size 10000

weighted moments (PWM) and the elemental percentile method (EPM) etc. In this dissertation, we only use the likelihood based method. Let $y_1, \ldots, y_k$ be k excesses of a predetermined threshold u. Then the log-likelihood derived from the distribution function is:

$$\begin{aligned}\xi \neq 0, \quad \ell(\sigma,\xi) &= \begin{cases} -k\log\sigma - (1+1/\xi)\sum_{i=1}^{k}\log(1+\xi y_i/\sigma), & 1+\xi y_i/\sigma > 0 \\ 0, \quad otherwise & \end{cases} \\ \xi = 0, \quad \ell(\sigma) &= -k\log\sigma - \sigma^{-1}\sum_{i=1}^{k} y_i\end{aligned}$$

Again, closed-form solutions of the maximum likelihood estimators do not exist and numerical optimization routines are required.

In real data applications in Food and Drug safety, the sample size is typically small. With small sample sizes, the GPD log-likelihood function may become flat and the optimizer could fail to converge. This problem could be overcome through penalizing the likelihood by some function of the parameters. Empirical studies suggest that such problems are often overcome by putting a moderate penalty on ξ^2. That is, instead of maximizing the log-likelihood $\ell(\sigma, \xi)$, one seeks to maximize $\ell(\sigma, \xi) - \lambda\xi^2$ for some λ. This expression can be exponentiated and rewritten as $L(\sigma, \xi)e^{-\xi^2/2\theta^2}$, where $\theta = \sqrt{1/2\lambda}$. This factor term of this expression is proportional to a Gaussian distribution with zero mean. In this sense, the penalized likelihood estimation has a Bayesian interpretation and corresponds to the mode of the posterior distribution. In practice, MLE is attempted first. When convergence issues arise, penalized MLE with a diffuse prior is used.

1.4.2 Tail Probability and Extreme Quantile Estimation

It should be noticed that the tail probability for the POT method is computed differently. Theorem 1.3 states:

$$\mathrm{P}(X > x | X > u) = \left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi}$$

It follows that the tail probability is then:

$$\mathrm{P}(X > x) = \mathrm{P}(X > u)\left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi}$$

Let ζ_u denote $\mathrm{P}(X > u)$, the natural estimator of this probability is $\hat{\zeta}_u = k/n$. In other words, the estimator of $\mathrm{P}(X > u)$ is simply the proportion of the observations in the sample that exceeds the predetermine threshold u. Denote the estimated tail probability by $\hat{p}_x$; this can be obtained by substituting the parameters (σ, ξ, ζ_u) by their estimates $\hat{\sigma}, \hat{\xi}, \hat{\zeta}_u$:

$$\hat{p}_x = \hat{\mathrm{P}}(X > x) = \hat{\zeta}_u(1 - \hat{H}(x)) = \frac{k}{n}\left[1 + \hat{\xi}\left(\frac{x-u}{\hat{\sigma}}\right)\right]^{-1/\hat{\xi}}$$

Similarly, the estimated p quantile can be obtained:

$$\hat{z}_p = \hat{H}^{-1}(p) = \begin{cases} u + \frac{\hat{\sigma}}{\hat{\xi}}\left[\left(\frac{k}{np}\right)^{-\xi} - 1\right], & \xi \neq 0 \\ u + \hat{\sigma}\log(\frac{k}{np}), & \xi = 0 \end{cases}$$

By construction, z_p is the $1/(1-p)$-observations return level. This is equivalent to the pth quantile q_p of the original data. Standard errors or confidence intervals for the exceedance probability p and the pth quantile z_p can be derived by the delta method. For the POT method, the uncertainty in the estimate of ζ_u should also be considered in the calculation. By standard properties of the binomial distribution, $\mathrm{Var}(\hat{\zeta}_u) \approx \hat{\zeta}_u(1 - \hat{\zeta}_u)$. The complete variance-covariance matrix for $(\hat{\sigma}, \hat{\xi}, \hat{\zeta}_u)$ is approximately:

$$V = \begin{bmatrix} \hat{\zeta}_u(1 - \hat{\zeta}_u) & 0 & 0 \\ 0 & v_{1,1} & v_{1,2} \\ 0 & v_{2,1} & v_{2,2} \end{bmatrix}$$

where $v_{i,j}$ denotes the (i,j) term of the covariance matrix of $\hat{\sigma}$ and $\hat{\xi}$. By the delta method,

$$\begin{aligned} \mathrm{Var}(\hat{p}_x) &\approx \nabla p_x^T V \nabla p_x \\ \mathrm{Var}(\hat{z}_p) &\approx \nabla z_p^T V \nabla z_p \end{aligned} \tag{1.4}$$

where

$$\begin{aligned}
\nabla p_x^T &= \left[\frac{\partial p_x}{\partial \zeta_u}, \frac{\partial p_x}{\partial \sigma}, \frac{\partial p_x}{\partial \xi}\right] \\
&= \Big[[1+\xi(x-u)/\sigma]^{-1/\xi}, \quad \zeta(x-u)/\sigma^2(\xi(x-u)/\sigma+1)^{-(1+\xi)/\xi}, \\
&\qquad \frac{p_x((\xi(x-u)+\sigma)\log(\xi(x-u)/\sigma+1)+\xi(u-x))}{\xi^2(\xi(x-u)+\sigma)}\Big], \\
\nabla z_p^T &= \left[\frac{\partial z_p}{\partial \zeta_u}, \frac{\partial z_p}{\partial \sigma}, \frac{\partial z_p}{\partial \xi}\right] \\
&= \Big[\sigma\zeta_u^{\xi-1}/p, \quad \xi^{-1}\{(\zeta_u/p)^{\xi}-1\}, \\
&\qquad -\sigma\xi^{-2}\{(\zeta_u/p)^{\xi}-1\}+\sigma\xi^{-1}(\zeta_u/m)^{\xi}\log(\zeta_u/p)\Big],
\end{aligned}$$

evaluated at $(\hat{\sigma}, \hat{\xi}, \hat{\zeta}_u)$.

1.4.3 Model Diagnostics

Probability plots, quantile plots, return level plots and density plots are useful graphical model diagnostic tools for assessing the quality of a fitted GPD model. Let $y_{(1)} \leq \cdots \leq y_{(k)}$ denote ordered threshold excesses for a threshold u, and $\hat{H}$ be an estimated GPD fit, the probability plot consists of the pairs

$$(\frac{i}{k+1}, \hat{H}(y_{(i)})), \qquad i = 1, \ldots, k$$

where $\hat{H}(y) = 1-(1+\hat{\xi}y/\hat{\sigma})^{-1/\hat{x}i}$ for $\hat{\xi} \neq 0$ and $\hat{H}(y) = 1-\exp(-y/\hat{\sigma})$ for $\hat{\xi} = 0$.

The quantile plot consists of the pairs

$$(\hat{H}^{-1}(\frac{i}{k+1}), y_{(i)}), \qquad i = 1, \ldots, k$$

where $\hat{H}^{-1}(y) = u+\hat{\sigma}/\hat{\xi}(y^{-\hat{\xi}}-1)$ for $\hat{\xi} \neq 0$.

If the GPD fit is reasonable for modeling excesses of u, then both the probability and quantile plots should consist of points that are approximately linear.

A return level plot consists of the locus of points $(1/(1-p), \hat{x}_p)$ for small values of p, where $\hat{x}_p$ is the estimated $(1/(1-p)$-observation return level. It is usual to plot the return level curve on a logarithmic scale to reflect the effect of extrapolation. Confidence bounds and empirical estimates of the return levels also help to make goodness of fit judgments when added to the plot.

Finally, the density plot of the fitted GPD can be compared against the histogram of the threshold exceedances.

1.5 Other Methods in Tail Probability and Extreme Quantile Estimation

In this section, two other methods in computing threshold exceeding probabilities and the associated confidence intervals will be introduced. One method is the widely known empirical method, and the other is called the Agresti-Coull method.

1.5.1 Empirical Method

Let $X_1, \ldots, X_n$ be a sequence of i.i.d. variables with a common distribution function G. Then the empirical distribution function can be defined as:

$$\tilde{G}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i \leq t\}$$

where $\mathbb{1}\{x_i \leq t\}$ is the indicator that follows a Bernoulli distribution with parameter $p = G(t)$ for some fixed t. Therefore, $\tilde{G}$ is a binomial random variable with mean

$G(t)$ and standard deviation $\sqrt{nG(t)(1-g(t)}$. The empirical distribution converges weakly to the true distribution function for every t:

$$\sqrt{n}(\tilde{G}(t) - G(t)) \xrightarrow{d} \mathcal{N}(0, G(t)(1 - G(t)))$$

Thus the tail probability that a random variable exceeds a fixed threshold t is simply

$$\hat{p} = \hat{\mathrm{P}}(X > t) = 1 - \tilde{G}(t)$$

If the significance level α is specified, due to the above convergence result, the confidence intervals for this probability can also be constructed:

$$\{1 - \tilde{G}(t) z_{1-\alpha/2} \sigma_{EP}(t)/\sqrt{n}, 1 - \tilde{G}(t) z_{1-\alpha/2} \sigma_{EP}(t)/\sqrt{n}\}$$

where $\sigma_{EP} = \sqrt{nG(t)(1-g(t)}$. This method is widely known and taught as the standard method in all introductory statistical text books. If exceeding a fixed threshold t for any observation in the sample is considered as a success, and X denotes the number of successes in a sample of size n, then X simply follows the binomial distribution, and the point estimator of the success proportion is $\hat{p} = X/n$. It is a nonparametric method that does not require any distributional assumptions, and is relatively robust. However, the confidence interval constructed by this approach often does not have the nominal coverage when the threshold t is large. In other words, the performance of this method is poor when $G(t)$ is close to 0 or 1, especially when the sample size is not large enough. Some remedial approaches are proposed to overcome this difficulty. The Agresti-Coull method which is one of the remedial methods will be described next.

1.5.2 Agresti-Coull Method

The Agresti-Coull method has been proposed to improve the coverage of the empirical confidence intervals when the sample size is small and the threshold probability is close to 0 and 1. In the Agresti-Coull method, the usual point estimator of the threshold exceeding probability $\hat{p}$ is replaced by the modified estimator:

$$\tilde{p} = \frac{X + z_{1-\alpha/2}^2/2}{\tilde{n}}$$

where $\tilde{n} = n + z_{1-\alpha/2}^2$ is the modified sample size which replaces the true sample size n. Similar to the empirical method, the Agrest-Coull method also uses the normal approximation to obtain the confidence interval.

$$\left(\tilde{p} - z_{1-\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}, \tilde{p} + z_{1-\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}\right)$$

Note that $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of a standard normal distribution. When a 95% confidence interval for the tail probability is desired, $\alpha = 0.05$ and $z_{1-\alpha/2} = 1.96$. If 2 is used instead of 1.96, then the Agresti-Coull method simply adds approximately two successes and two failures in the computation of $\tilde{p}$ and the associated confidence interval. Therefore, the resulting confidence interval is a very conservative one and usually wider than the confidence interval of the empirical method.

Chapter 2: The Density Ratio Model

2.1 Overview

In this chapter, some necessary background about the density ratio model will be introduced. The ideas of Out of Sample Fusion (OSF)and Repeated Out of Sample Fusion (ROSF)are reviewed.

The estimation of a distribution function is one of the most fundamental problems in probability and statistics. Given an observed sample, one may directly estimate the underlying distribution function through the empirical distribution function which is a step function that jumps by $1/n$ at each of the n data points. The details of Density Ratio Models are discussed in Sec. 2.2. The DRM is by no mean a new discipline - its roots can be traced back to the 1980s. An early form of DRM was suggested by Patil and Rao (1978) [37], and Vardi (1982) [48]. In Vardi's study, the length of an object is assumed to be distributed according to the cdf G, and the selection probability for any particular object is proportional to its length. The distribution of the length of sampled objects is given by the model,

$$F(y) = \frac{1}{\mu} \int_0^y x dG(x), \quad y \geq 0$$

where $\mu = \int_0^\infty x dG(x) < \infty$ is the normalization constant. Here the cdf G is unknown and is to be estimated. The cdf F, the length-biased distribtuion corresponding to G, turns out to be a weighted version of G in terms of the weight function x. Vardi later generalized the two sample model to allow for $s+1$ different biased samples:

$$F_i(y) = W_i(G)^{-1} \int_{-\infty}^{y} w_i(x) dG(x), \quad i = 1, \ldots, s$$

where w_i's are given the nongegative selection bias weight function and

$$W_i(G) = \int_{-\infty}^{\infty} w_i(x) dG(x).$$

A simple way to estimate G is to use the empirical distribution of the reference sample X_0 only, but this ignores the other s samples. A bias-corrected estimator which corrects for the biasing involved in the distributions F_i is desired. Vardi [49] developed methodology for obtaining a nonparametric maximum likelihood estimate (NPMLE) by using all the $n = n_0 + n_1 + \cdots + n_s$ observations from the $s+1$ samples. In Vardi's original treatment, the weight functions were assumed completely known. However, in many practical situations, a complete specification of the weight functions is unrealistic and too restrictive. To overcome this issue, one may assume that the weight function comes from a parametric family. In this case, the model involves two components to be estimated: the unknown reference distribution G and the parameters involved in the weight function. These types of models are called biased sampling semiparametric models, and the logistic regression models in case-control studies is an example.

Case-control is a frequently used tool to study risk factors related to disease incidence, and logistic regression models are commonly used in analyzing case-control data. Let $D = 0$ be the control, $D = 1$ be the case, $\mathbf{x} = (x_1, \ldots, x_p)$ be the regression vector or covariates, and $P(D = i \mid \mathbf{x})$ denote the probability that individual with characteristic $\mathbf{x}$ develops disease $D = i$, the logistic regression model takes the following form:

$$P(D = i \mid \mathbf{x}) = \frac{\exp(\alpha_i + \beta_i' x)}{1 + \sum_{j=0}^{m} \exp(\alpha_j + \beta_j' x)}, \qquad i = 0, 1 \tag{2.1}$$

Let $p(x)$ be the marginal distribution of $\mathbf{x}$, and let$\pi_i = P(D = i)$ (note that the π_i's satisfy $\sum_{i=0}^{m} \pi_i = 1$). Then, by Bayes rule, we have:

$$P(x \mid D = i) = \frac{P(D = i \mid x) p(x)}{\pi_i}, \quad i = 0, 1$$

Therefore,

$$\frac{P(x \mid D = 1)}{P(x \mid D = 0)} = \frac{\pi_0}{\pi_1} \frac{P(D = 1 \mid x)}{P(D = 0 \mid x)} \tag{2.2}$$

Substituting 2.1 into 2.2, and notice that $\alpha_0 = \beta_0 = 0$, we get the density ratio setup:

$$\frac{P(x \mid D = 1)}{P(x \mid D = 0)} \quad = \quad \exp(\alpha_1^* + \beta_1' x)$$

where $\alpha_1^* = \log(\pi_0/\pi_1) + \alpha$. If we let $g_i(x)$ denote the conditional density function $P(x \mid D = i)$, $i = 0, 1$. We can rewrite the previous formula as:

$$g_1(x) = \exp(\alpha_1^* + \beta_1' x) g_0(x)$$

The case pdf become a weighted version of the control pdf. This is a density ratio model. The exponential function is the weight, x is called the distortion function, and the function $g_0(x)$ is regarded as the density of the reference sample X_0. The parameters α_i, β_i, and the density g_0 are to be estimated. This shows that the logistic regression model for a case-control study is equivalent to the biased sampling model with weight function $\exp(\alpha + \beta\mathbf{X})$. Later we will show that the multiple-sample semiparametric density ratio model is equivalent to the generalized logistic regression model.

2.2 Density Ratio Models

Motivated by biased sampling models and case-control studies, density ratio models were studied in Qin and Lawless (1994) [39], Qin and Zhang (1997) [40], Fokianos et al. (2001) [20], Kedem et al. (2008) [29], Voulgaraki et al. (2012) [51], Zhou et al. (2013) [53]. For the two-sample case:

$$\begin{aligned} X_0 &= (x_{01}, \ldots, x_{0n_0})' \sim g_0(x) \\ X_1 &= (x_{11}, \ldots, x_{1n_1})' \sim g_1(x) \end{aligned}$$

the density ratio model is:

$$\frac{g_1(x)}{g_0(x)} = e^{\alpha + \beta' \cdot h(x)} \tag{2.3}$$

where $h(x)$ is the so called tilt function, which can be regarded as distortion of sample x_1's pdf from the reference sample x_0's pdf. The model is intuitive since the density ratio of two pdfs' has the form 2.3 if both of them come from the exponential family. Now let's consider two cases.

2.2.1 When g_0 and g_1 are from the same exponential family

Suppose X_0 and X_1 are from the same exponential family with pdf $g_i(x|\theta), \quad i = 0, 1$ expressed in the following form:

$$\begin{aligned} g_i(x|\boldsymbol{\theta}) &= h(x)c(\boldsymbol{\theta}) \exp\left[\sum_{j=1}^{k} w_j(\boldsymbol{\theta}) t_j(x)\right] \\ &= h(x)c(\boldsymbol{\theta}) \exp\left[(w_1(\boldsymbol{\theta}), \ldots, w_k(\boldsymbol{\theta})) \begin{pmatrix} t_1(x) \\ \vdots \\ t_k(x) \end{pmatrix}\right] \\ &= h(x)c(\boldsymbol{\theta}) \exp\left[\mathbf{w}(\boldsymbol{\theta}) \cdot \mathbf{t}(x)\right], \quad x \in \chi \subset \mathbb{R}^q, \end{aligned}$$

where $w_1, \ldots, w_k$ and c are real-valued functions of $\boldsymbol{\theta}$, and real-valued functions $t_1, \ldots, t_k$ and h have their supports on $\mathbb{R}^q$. Then:

$$\begin{aligned} \frac{g_1(x)}{g_0(x)} &= \frac{c(\boldsymbol{\theta}_1)}{c(\boldsymbol{\theta}_0)} \cdot \exp\left\{\sum_{j=1}^{k} \big[w_j(\boldsymbol{\theta}_1) - w_j(\boldsymbol{\theta}_0)\big] \cdot t_j(x)\right\} \\ &= \exp\left\{\sum_{j=1}^{k} \big[w_j(\boldsymbol{\theta}_1) - w_j(\boldsymbol{\theta}_0)\big] \cdot t_j(x) + \log\frac{c(\boldsymbol{\theta}_1)}{c(\boldsymbol{\theta}_0)}\right\} \\ &= \exp\left\{\alpha + \boldsymbol{\beta} \cdot \mathbf{h}(x)\right\} \end{aligned}$$

where,

$$\begin{aligned} \alpha &= \log\frac{d(\boldsymbol{\theta}_1)}{d(\boldsymbol{\theta}_0)} \\ \boldsymbol{\beta} &= \big(w_j(\boldsymbol{\theta}_1) - w_j(\boldsymbol{\theta}_0)\big), \quad j = 1, \ldots k \\ \mathbf{h}(x) &= \mathbf{t}(x) = \big(t_i(x)\big)', \quad j = 1, \ldots, k \end{aligned}$$

A list of the one-to-one correspondence between the tilt function h(t) and pdf follows:

$h(x)$	distribution
$h(x) = x$	$g(x) \sim \exp(\lambda)$
$h(x) = x, x^2$	$g(x) \sim \mathrm{N}(\mu, \sigma^2)$
$h(x) = x, \log(x)$	$g(x) \sim \Gamma(k, \lambda)$
$h(x) = \log(x), \log(1 - x)$	$g(x) \sim \mathrm{Beta}(\alpha, \beta)$
$h(x) = \log(x), (\log(x))^2$	$g(x) \sim \text{Log-Normal}(\mu, \sigma^2)$
$h(x) = x, \log(\delta^2 + (x - \mu)^2)$	$g(x) \sim \mathrm{GHD}(\lambda, \alpha, \beta, \delta, \mu)$

When the underlying distribution of the reference sample approximately follows the normal distribution, the tilt function $h(x) = x, x^2$ would be valid. In risk analysis, however, the distribution of the underlying sample is often skewed with long and possibly heavy tails. In such cases, the tilt function should be chosen with discretion. Typically, the “gamma” tilt and the “log-normal” tilt are good choices when the reference sample takes only positive values and has a long tail, such as pathogen counts and contamination intake. When the sample takes negative values (eg. stock returns), it is appropriate to use the ”GHD” (Generalized Hyperbolic Distribution) tilt.

For example, let us consider the ratio of two gamma probability densities with shapes r_1, r_2 and rates λ_1, λ_2. Then the parameters take on the form

$$\begin{aligned} \alpha &= \log \frac{\lambda_1^{r_1} \Gamma(r_2)}{\lambda_2^{r_2} \Gamma(r_1)} \\ (\beta_1, \beta_2) &= (\lambda_2 - \lambda_1, r_1 - r_2) \end{aligned}$$

and the tilt function is $h(x) = (x, \log(x))$

2.2.2 When g_0 and g_1 come from different exponential families

$$\begin{aligned}
\frac{g_1(x)}{g_0(x)} &= \frac{c(\boldsymbol{\theta}_1)h_1(x)}{c(\boldsymbol{\theta}_0)h_0(x)} \cdot \exp\left\{\sum_{j=1}^{k}\big[w_{1j}(\boldsymbol{\theta}_1)\cdot t_{1j}(x) - w_{0j}(\boldsymbol{\theta}_0)\cdot t_{0j}(x)\big]\right\} \\
&= \exp\left\{\sum_{j=1}^{k}\big[w_{1j}(\boldsymbol{\theta}_1)\cdot t_{1j}(x) - w_{0j}(\boldsymbol{\theta}_0)\cdot t_{0j}(x)\big] + \log\frac{c_1(\boldsymbol{\theta}_1)}{c_0(\boldsymbol{\theta}_0)} + \log\frac{h_1(x)}{h_0(x)}\right\} \\
&= \exp\{\alpha + \phi(x,\boldsymbol{\beta})\}
\end{aligned}$$

where,

$$\begin{aligned}
\alpha &= \log\frac{c(\boldsymbol{\theta}_1)}{c(\boldsymbol{\theta}_0)} \\
\phi(x,\boldsymbol{\beta}) &= \sum_{j=1}^{k}\big[w_{1j}(\boldsymbol{\theta}_1)\cdot t_{1j}(x) - w_{0j}(\boldsymbol{\theta}_0)\cdot t_{0j}(x) + \log\left.\frac{h_1(x)}{h_0(x)}\right\}\big]
\end{aligned}$$

The semiparametric density ratio model establishes relationships between a reference distribution and its tilted versions. The multiple sample semiparametric density ratio model considers the following $m+1$ independent samples:

$$\begin{aligned}
X_0 &= (x_{01},\ldots,x_{0n_0})' \sim g(x) \\
X_1 &= (x_{11},\ldots,x_{1n_1})' \sim g_1(x) \\
&\vdots \\
X_m &= (x_{m1},\ldots,x_{mn_m})' \sim g_m(x)
\end{aligned}$$

where $g_j(x)$ is the probability density of the jth sample of size n_j. We denote X_0 as the reference sample, its distribution $G(x)$ is assumed to be unknown. To estimate g and G, we assume there are additional samples from related distributions, or in some sense, similar with regard to the regions of values of the variable(s) of interest.

In particular, the samples may be computer generated. The out of sample fusion idea is to combine or fuse the reference real data X_0 with computer generated data using the density ratio model. See Katzoff et al. (2014) [28] and Zhou (2013) [53].

The density ratio model assumes that the reference distribution $g(x)$ and its tilted versions $g_j(x)$ are related by the ratios,

$$\begin{aligned}\frac{g_1(x)}{g(x)} &= exp(\alpha_1 + \beta_1^{'} h(x)) \\ &\vdots \\ \frac{g_m(x)}{g(x)} &= exp(\alpha_m + \beta_m^{'} h(x))\end{aligned} \tag{2.4}$$

This in turn gives the tilt model:

$$g_j(x) = e^{\alpha_j + \beta_j' h(x)} g(x), \quad j = 1, \ldots, m \tag{2.5}$$

where β_j is $p \times 1$ parameter vectors, α_j are scalar parameters and $h(x)$ is a vector valued distortion or tilt function. The probability densities g, g_1, ... , g_m and the parameters α's and β's are unknown, but h is assumed to be a known function. The relationship 2.4 is called the density ratio model, and allows semiparametric inference about all the parameters and distributions from the fused $m + 1$ samples,

$$\mathbf{t} = (\mathrm{t}_1, ..., \mathrm{t}_\mathrm{n})' = (\mathrm{X}_0^{'}, \mathrm{X}_1^{'}, ..., \mathrm{X}_\mathrm{m}^{'})^{'} \tag{2.6}$$

of sizes $n = n_0 + n_1 + ... + n_m$. Since $n_0 \lll n$, the reference G, under 2.4, is estimated with much more data. For a rigorous treatment of the semiparametric inference under 2.4, see, for example, Lu (2007) [34] and Qin (1997) [40].

2.3 Estimation for Density Ratio Model

Maximum likelihood estimates for all the parameters and $G(x)$ can be obtained by maximizing the empirical likelihood over the class of step cumulative distribution functions with jumps at the observed values $t_1, ..., t_n$ (see [36]). Note that from 2.6 the reference distribution function G is supported at all the n observed values $t_1, ...t_n$ and not just at the n_0 values from the reference sample X_0. Thus, if we let $p_i = dG(t_i)$ be the mass at t_i, for $i = 1, ..., n$, the empirical likelihood becomes

$$\mathcal{L}(\boldsymbol{\theta},\mathrm{G}) = \prod_{i=1}^{n} p_i \prod_{j=1}^{n_1} \exp(\alpha_1 + \beta_1^{'} h(x_{1j})) \times \cdots \times \prod_{j=1}^{n_m} \exp(\alpha_m + \beta_m^{'} h(x_{mj})) \tag{2.7}$$

Maximizing $\mathcal{L}(\boldsymbol{\theta},\mathrm{G})$ subject to the constraints $\sum_{i=1}^{n} p_i = 1$ and

$$\sum_{i=1}^{n} p_i[w_1(t_i) - 1] = 0, ..., \sum_{i=1}^{n} p_i[w_m(t_i) - 1] = 0$$

where $w_j(x) = \exp(\alpha_j + \beta_j^{'} h(x))$, $j = 1, ..., m$, we obtain the desired estimates through the method of Lagrange multiplier. First, set up the objective function

$$\log \mathcal{L}(\theta, G) - \lambda_0(1 - \sum_{i=1}^{n} p_i) - \lambda_1 \sum_{i=1}^{n} p_i[w_1(t_i) - 1] - \cdots - \lambda_m \sum_{i=1}^{n} p_i[w_m(t_i) - 1], \quad i = 1, \ldots, n$$

and obtain $\lambda_0 = n$ and $\lambda_j = n_j \quad j = 1, \ldots, m$ and

$$p_i = \frac{1}{n_0} \cdot \frac{1}{1 + \rho_1 w_1(t_i) + \cdots + \rho_q w_q(t_i)}$$

Back substitute p_i's into $\mathcal{L}(\boldsymbol{\theta},\mathrm{G})$ to get the profile log-likelihood as a function of $\boldsymbol{\theta}$ only:

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= -n\log n_0 - \sum_{i=1}^{n}\log[1+\rho_1 w_1(t_i)+\cdots+\rho_m w_m(t_i)] \\ &+ \sum_{j=1}^{n_1}(\alpha_1+\beta_1' h(x_{1j}))+\cdots \\ &+ \sum_{j=1}^{n_m}(\alpha_m+\beta_m' h(x_{mj}))\end{aligned}$$

Then, differentiate the objective function $\log\boldsymbol{\mathcal{L}}$ w.r.t. α's and β's to get the score equations

$$\begin{aligned}\frac{\partial l}{\partial \alpha_j} &= -\sum_{i=1}^{n}\frac{\rho_j w_j(t_i)}{1+\rho_1 w_1(t_i)+\cdots+\rho_q w_q(t_i)} + n_j = 0 \\ \frac{\partial l}{\partial \beta_j} &= -\sum_{i=1}^{n}\frac{\rho_j h(t_i) w_j(t_i)}{1+\rho_1 w_1(t_i)+\cdots+\rho_q w_q(t_i)} + \sum_{i=1}^{n_j} h(x_j i) = 0\end{aligned}$$

The solution of the score equations gives the maximum likelihood estimators $\hat{\alpha}, \hat{\beta}$. Consequently, by substitution,

$$\hat{p}_i = \frac{1}{n_0}\cdot\frac{1}{1+\sum_{j=1}^{q}\rho_j\exp(\hat{\alpha}_j+\hat{\boldsymbol{\beta}}'{}_j\mathbf{h}(t_i))} \tag{2.8}$$

In particular, the maximum likelihood estimate $\hat{G}$ of G is given in 2.9 for relative sample sizes $\rho_j = n_j/n_0$,

$$\hat{G}(t) = \frac{1}{n_0}\sum_{i=1}^{n}\frac{I(t_i\le t)}{1+\rho_1\hat{w}_1(t_i)+\ldots+\rho_m\hat{w}_m(t_i)} \tag{2.9}$$

where $\hat{w}_j(x) = \exp(\hat{\alpha}_j + \hat{\beta}_j' h(x))$, $j = 1, ..., m$, and $I(t_i \le t)$ equals one for $t_i \le t$ and is zero otherwise. Similarly, $\hat{G}_j$ can be estimated by accumulating $\exp(\hat{\alpha}_j + \hat{\beta}_j' h(t_i))dG(t_i)$. The asymptotic results for $\hat{G}$ are given in the next section.

2.4 Asymptotic Theory for $\hat{G}$

The multiple sample asymptotic behavior of $\hat{G}$ is obtained by Lu (2007) from which we obtain semiparametric (SP) confidence intervals for using the covariance matrix given below. The following quantities must be defined first before we give the asymptotic behavior of $\hat{G}(t)$:

$$\begin{aligned} A_j(t) &= \int \frac{w_j(y)I(y \le t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \\ B_j(t) &= \int \frac{w_j(y)h(y)I(y \le t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \end{aligned}$$

$$\bar{A}(t) = (A_1(t), ..., A_m(t))' \quad , \quad \bar{B}(t) = (B_1'(t), ..., B_m'(t))'$$

$$\boldsymbol{\rho} = \text{diag}\{\rho_1, \ldots, \rho_m\}_{m \times m}, \quad \mathbf{1}_p = (1, ..., 1)'$$

Then the asymptotic distribution of $\hat{G}(t)$ for $m \ge 1$ is given in the following theorem.

Theorem 2.1 (Lu)**.** *The process $\sqrt{n}(\hat{G}(t) - \tilde{G}(t))$ converges weakly to a zero-mean Gaussian process in the space of real right continuous functions that have left limits with covariance matrix given by*

$$\begin{aligned} Cov\left\{\sqrt{n}(\hat{G}(t) - \tilde{G}(t)),\, \sqrt{n}(\hat{G}(s) - \tilde{G}(s))\right\} &= \sum_{k=0}^m \rho_k \sum_{j=1}^m \rho_j A_j(t \wedge s)) \\ &\quad - \left(\bar{A}'(t)\boldsymbol{\rho}, \bar{B}'(t)(\boldsymbol{\rho} \otimes \mathbf{1}_p)\right) S^{-1} \begin{pmatrix} \boldsymbol{\rho}\bar{A}(s) \\ (\boldsymbol{\rho} \otimes \mathbf{1}_p)\bar{B}(s) \end{pmatrix} \end{aligned}$$

Theorem 2.2 (Lu)**.** *The process $\sqrt{n}(\hat{G}(t) - G(t))$ converges to a zero-mean Gaussian process in the space of real right continuous functions that have left limits with covariance matrix given by*

$$\begin{aligned} Cov\left\{\sqrt{n}(\hat{G}(t) - G(t)),\, \sqrt{n}(\hat{G}(s) - G(s))\right\} &= (\sum_{k=0}^m \rho_k)(G(t \wedge s) - G(t)G(s) - \sum_{j=1}^m \rho_j A_j(t \wedge s)) \\ &\quad + \left(\bar{A}'(s)\boldsymbol{\rho}, \bar{B}'(s)(\boldsymbol{\rho} \otimes \mathbf{1}_p)\right) S^{-1} \begin{pmatrix} \boldsymbol{\rho}\bar{A}(t) \\ (\boldsymbol{\rho} \otimes \mathbf{1}_p)\bar{B}(t) \end{pmatrix} \end{aligned}$$

where $\mathbf{1}_p$ is the $p \times p$ identity matrix, and $\otimes$ denotes Kronecker product.

The complete derivation of the theory is quite technical and only the main steps of the proof are given here. First express $\sqrt{n}(\hat{G}(t) - G(t))$ as the sum of two parts:

$$\sqrt{n}(\hat{G}(t) - G(t)) = \sqrt{n}(\hat{G}(t) - \tilde{G}(t)) + \sqrt{n}(\tilde{G}(t) - G(t))$$

where $\tilde{G}(t) = \frac{1}{n_m}\sum_{i=1}^{n_m} I[x_{mi} < t]$ is the empirical distribution of the reference sample X_m only. The asymptotic properties of $\sqrt{n}(\tilde{G}(t) - G(t))$ are well known. Thus, the objective is to prove the weak convergence of $\sqrt{n}(\hat{G}(t) - \tilde{G}(t))$. By the strong consistency of the estimators $\hat{\boldsymbol{\theta}}$, the Taylor series expansion of $\hat{G}(t)$ at the true parameter θ_0 approximates $\hat{G}$ uniformly in t:

$$\hat{G}(t) = H_1(t) - H_2(t) + R_n(t)$$

where

$$\begin{aligned} H_1(t; \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \frac{1}{n_m}\sum_{i=1}^{n} \frac{\mathrm{I}(t_i \leq t)}{\sum_{k=0}^{m} \rho_k w_k(t_i; \alpha_k, \beta_k)} \\ H_2(t; \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \frac{1}{n}\Big(\bar{A}'(t)\boldsymbol{\rho}, \bar{B}'(s)(\boldsymbol{\rho} \otimes \mathbf{1}_p)\Big) S^{-1} \begin{pmatrix} \frac{\partial l(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\alpha}} \\ \frac{\partial l(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \end{pmatrix} \end{aligned}$$

In this case, the asymptotic behavior of $\sqrt{n}(\hat{G}(t) - \tilde{G}(t))$ is equivalent to that of $\sqrt{n}(H_1(t) - H_2(t) - \tilde{G}(t))$ which involves the true parameter $\boldsymbol{\theta}_0$ only. The weak convergence of the finite-dimensional distributions of $\sqrt{n}(H_1(t) - H_2(t) - \tilde{G}(t))$ follows from the multivariate central limit theorem after obtaining the variance covariance structure. Tightness of the process is proved by noting that both $\sqrt{n}(H_1(t) - \tilde{G}(t))$ and $\sqrt{n}H_2(t)$ can be decomposed into sums of empirical processes. Each empirical process is evaluated at a function $f(\cdot)$ in a Donsker class

$(P_n f = n^{-1}\sum_{i=1}^{n} f(T_i),\quad$ where $P_n = n^{-1}\sum_{i=1}^{n}\delta_{T_i}$ is an empirical measure defined on i.i.d. observations $T_1,\dots,T_n)$. The weak convergence of each empirical process follows from the classical Donsker theory.

2.5 Goodness-of-Fit

Goodness-of-fit tests are needed to justify the density ratio model applicability. Let $\hat{G}(t)$ be the estimated reference cdf and $\tilde{G}(t)$ be the corresponding emipircal cdf from X_0. Most goodness-of-fit tests measure the discrepancy between $\hat{G}(t)$ and $\tilde{G}(t)$. A simple graphical method is to plot $\hat{G}(t)$ versus $\tilde{G}(t)$, see Voulgaraki et al. (2012) [51]. A more prudent numerical method is propposed by Qin and Zhang (1997) [40]. Define the difference between $\hat{G}(t)$ and $\tilde{G}(t)$ as:

$$\Delta_n(t) = \sqrt{n}\,|\hat{G} - \tilde{G}|, \qquad \Delta_n = \sup_{-\infty<t<\infty} \Delta_n(t)$$

Δ_n can be used to measure the departure from the assumption of the Semiparametric Density Ratio Model. Theorem 2.1 shows that $\sqrt{n}(\hat{G}(t) - \tilde{G}(t))$ converges weakly to a Gaussian process W. Let w_α denote the α-quantile of the distribution of $sup_{-\infty<t<\infty}|W(t)|$. By Theorem 2.1,

$$\begin{aligned}\lim_{n\to\infty} P(\Delta_n &\geq w_{1-\alpha}) = \lim_{n\to\infty} P(\sup_{-\infty<t<\infty} \sqrt{n}\,|\hat{G} - \tilde{G}| \geq w_{1-\alpha}) \\ &= P(\sup_{-\infty<t<\infty} \sqrt{n}\,|W(t)| \geq w_{1-\alpha}) = \alpha\end{aligned}$$

The density ratio model is rejected at level α if $\Delta_n \geq w_{1-\alpha}$. However, there is no analytic expression available for the distribution of the supremum of a Gaussian process $W(t)$ and its corresponding quantile function. A bootstrap procedure must be applied to simulate the distribution of $sup_{-\infty<t<\infty}|W(t)|$ and its quantiles. We follow the following steps to get the distribution of Δ_n :

1. Obtain the estimated SP CDF's $\hat{G}, \hat{G}_1, \ldots, \hat{G}_m$ from combined samples $(X_0, X_1, \ldots, X_m)$.

2. Let $X_0^*, X_1^*, \ldots, X_m^*$ be random samples generated from $\hat{G}, \hat{G}_1, \ldots, \hat{G}_m$.

3. Let $(\hat{\boldsymbol{\alpha}}^*, \hat{\boldsymbol{\beta}}^*)$ and let $\hat{G}^*$ be the estimates for the parameters and the reference distribution obtained from $(X_0^*, X_1^*, \ldots, X_m^*)$.

4. Let $\tilde{G}^*$ be the empirical (EP) reference CDF from X_0^*.

5. The bootstrap version of the test statistic Δ_n is given by: $\Delta_n^*(t) = \sqrt{n}\,|\hat{G}^* - \tilde{G}^*|, \quad \Delta_n^* = \sup_{-\infty<t<\infty} \Delta_n^*(t)$.

6. Repeat Step 2 to Step 4 to generate many bootstrap replications of Δ_n^* and calculate the empirical p-value.

It turns out that as $n \to \infty$, $\sqrt{n}(\hat{G}^* - \tilde{G}^*) \xrightarrow{d} W$ in $D[-\infty, +\infty]$, where W is the Gaussian process defined in Theorem 2.1 [34]. This shows that the asymptotic behavior of $\sqrt{n}(\hat{G}^* - \tilde{G}^*)$ agrees with that of $\sqrt{n}(\hat{G} - \tilde{G})$, and $\Delta_n^*(t) = \sup_{-\infty\leq+\infty} \sqrt{n}\,|\hat{G}^* - \tilde{G}^*|$ has the same limiting behavior as $\Delta_n(t) = \sup_{-\infty\leq+\infty} \sqrt{n}\,|\hat{G} - \tilde{G}|$. Thus, it is legitimate to approximate the quantiles ofΔ_n by those of Δ_n^*.

Chapter 3: Threshold Exceedance Probabilities Using DRM

In this chapter, two new methods based on density ratio model in computing threshold exceedance probabilities and its associated confidence intervals will be introduced. One method is called the Out of Sample Fusion (OSF) and the other is called Repeated Out of Fusion (ROSF). See Katzoff et al. (2014) [28] and Zhou (2013) [53]. The performance of the existing methods (mentioned in Chapter 1) and DRM based methods will be compared through extensive simulations.

3.1 Out of Sample Fusion in Estimation of Threshold Probabilities

Let X_0 denote an i.i.d. sample from some given population

$$X_0 = (x_{01}, \ldots, x_{0n_0})' \sim g(x)$$

The distribution function $G(x)$ of X_0 is assumed to be unknown, and the threshold exceedance probability $\hat{p} = \hat{\mathrm{P}}(X_0 > t)$ for some fixed threshold t is of interest. X_0 is referred to as the *reference sample*. Let X_j denote a computer generated i.i.d. sample with sample sizes $n_j, j = 1, \ldots, m$

$$X_j = (x_{j1}, \ldots, x_{jn_j})' \sim g_j(x)$$

The computer generated samples X_j will be referred to as the *fusion samples*. Then under the density ratio model, we have

$$\frac{g_j(x)}{g(x)} = \exp\left(\alpha_j + \boldsymbol{\beta}_j'\mathbf{h}(x)\right), \qquad j = 1, \ldots, m$$

where α_j is a scalar parameter, β_j is a known $p \times 1$ parameter vector, and $h(x)$ is a $p \times 1$ vector valued distortion or tilt function. All probability distributions are connected under the density ratio model as shown in Chapter 2. Thus the semiparametric statistical inference about all the parameters and the probability distribution of the reference X_0 can be obtained from the combined data from the $m+1$ samples $X_0, X_1, \ldots,$. The combined data now has the size of $n = n_0 + n_1 + \cdots + n_m$. Therefore, the reference distribution function G is estimated from the fused data with n observations and not just from the reference sample itself with n_0 observations. The maximum likelihood estimator of the reference distribution function can be obtained from equation 2.8. So the corresponding threshold exceedance probability is:

$$\hat{p} = \hat{\mathrm{P}}(X_0 > t) = 1 - \hat{G}(t) = 1 - \frac{1}{n_0}\sum_{i=1}^{n} \frac{I(t_i \leq t)}{1 + \rho_1\hat{w}_1(t_i) + ... + \rho_m\hat{w}_m(t_i)}$$

where $\hat{w}_j(x) = \exp(\hat{\alpha}_j + \hat{\beta}_j' h(x))$, $j = 1, ..., m$.

For a large threshold T, the $100(1-\alpha)\%$ confidence intervals for $\hat{p} = 1 - G(T)$ can be constructed based on the asymptotic results given in Theorem 2.2:

$$\left(1 - \hat{G}(t) - z_{1-\alpha/2}\sqrt{\hat{V}(t)}, 1 - \hat{G}(t) + z_{1-\alpha/2}\sqrt{\hat{V}(t)}\right)$$

where $\hat{V}(t)$ denotes the estimated variance of $\hat{G}(t)$ as given in Theorem 2.2. When the tail probability $p = 1 - G(t)$ of interest becomes very small as the threshold t

becomes large, only the upper bound is used in the construction of the confidence intervals and lower bound is set to 0.

3.2 Repeated Out of Sample Fusion

Based on OSF, Repeated Out of Sample Fusion (ROSF) is a modified algorithm to estimate the tail probabilities and its confidence intervals where a given reference sample is fused or combined repeatedly with computer generated data. In statistics, bootstrapping is an extremely powerful idea that is widely used to compute the standard deviation of a quantity of interest when direct computation of such quantity is hard or even infeasible. Bootstrapping can be also used in a completely different context to improve measures of accuracy (defined in terms of variance, standard error, confidence intervals, etc) of sample estimates. In machine learning, there is also a list of techniques for model averaging and improvements based on the idea of bootstrapping. Bootstrapping aggregation or bagging is a general purpose algorithm for reducing the variance of a statistical learning method. Given a set of n independent observations $Z_1, \ldots, Z_n$, each with variance σ^2, the variance of the mean $\bar{Z}$ is then σ^2/n. Thus, averaging a set of observations could reduce the variance significantly. Therefore, it is natural to reduce the variance and hence increase the prediction accuracy of any statistical learning method by taking many training datasets from the population, building a separate prediction model using each training set and averaging the resulting predictions. In other word, B separate training sets can be used to calculate $\hat{f}^1(x), \hat{f}^2(x), \ldots, \hat{f}^B(x)$. Averaging the predictions would yield a single low variance statistical learning model,

$$\hat{f}_{avg}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{b}(x).$$

This is not practical since access to multiple training sets is generally not available. However, multiple training datasets can be generated by repeatedly sampling from the single training data set. Suppose B different bootstrapped training datasets are generated. Then the statistical method can be trained on the bth bootstrapped training set to obtain $\hat{f}^{*b}(x)$, and averaging all the predictions to yield:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x).$$

This algorithm is also known as bagging in the machine learning literature.

The general idea of ROSF is closely related to bootstrapping/bagging. While bootstrapping/bagging improves the estimation results by repeatedly sampling within a given sample, ROSF improves the results by repeatedly fusing a given sample with external samples. We shall now describe the implementation of ROSF in the estimation of threshold exceedance probabilities and the associated confidence intervals.

We are interested in estimating a small threshold exceedance probability $p > 0$ for a random sample X_0 from some distribution. X_0 is referred to as the reference sample. A fusion sample X_1 is then generated by the computer and fused together with the reference sample. The point estimate $\hat{p}_1$ and the confidence interval $[0, B_1]$ is then obtained through the semiparametric density ratio model as described in OSF method. The *same* reference sample is then fused with another computer generated sample (from the same distribution of the previous artificial sample and independent of it) to obtain another $\hat{p}_2$ and confidence interval $[0, B_2]$ in the same manner as before. This process is repeated n_r (stands for number of repetition) times to produce a sequence of point estimates $\hat{p}_i$ and confidence intervals $[0, B_i], i = 1, \ldots, n_r$. Conditional on X_0, the sequence of upper bounds B_i are independent and identically distributed from some distribution F_B. Denote the empirical distribution of B_i's by $\hat{F}_B$. By Glivenko-Cantelli theorem, $\hat{F}_B$ converges to F_B almost surely uniformly as n_r increases. Since the process may be repeated many times, a very close approximation of F_B can be obtained. In other words, as the number of fusions becomes very large, $\tilde{F}_B$ is almost the exact F_B.

The final point estimate of the threshold exceedance probability from ROSF algorithm is the average of $\hat{p}_i$'s from n_r OSF runs:

$$\hat{p} = \hat{\mathrm{P}}(X_0 > t) = \frac{1}{n_r} \sum_{i=1}^{n_r} \hat{p}_i, \quad i = 1, \ldots, n_r,$$

and the associated $100(1-\alpha)\%$ confidence interval is

$$\left[0, F_B^{-1}\left(\alpha^{1/N}\right)\right].$$

where N is a large enough positive integer.

Theorem 3.1. *Let $\hat{p}_i$ and B_i be the the sequence of point estimates of the tail probabilities and its upper bounds of associated confidence intervals obtained by ROSF. Let F_B denote the distribution function of the B's. Under the condition*

$$P(B > p) = 1 - F_B(p) > 0, \tag{3.1}$$

there exists N_0 such that for all $N > N_0$, the confidence interval for the tail probability p, $\left[0, F_B^{-1}\left(\alpha^{1/N}\right)\right]$ gives at least $100(1-\alpha)\%$ coverage.

Proof. For an i.i.d. sample $B_1, \ldots, B_N$, denote the maximum by $B_{(N)} = \max(B_i)$. It follows that

$$\mathrm{P}(B_{(N)} > p) = 1 - F_B^N(p)$$

If $\mathrm{P}(B > p) = 1 - F_B(p) > 0$, then from the above equation, the probability that the maximum upper bound covers the desired tail probability increases as the tuning parameter N increases. Conditional on the given sample X_0, for all $N > N_0$, we have the following inequality:

$$1 - F_B^N(p) \geq 0.95$$

for some N_0 sufficiently large. The inequality can be rewritten by inverting the distributon function:

$$0 < p \leq F_B^{-1}(0.05^{1/N})$$

The above relationship implies that the interval $\left[0, F_B^{-1}\left(\alpha^{1/N}\right)\right]$ covers the true tail probability p with at least 95% confidence for sufficiently large N. □

The length of the confidence interval depends on the choice of N. Here, N serves as a tuning parameter. Intuitively, as the number of fusions increases, the number of B_i's grows and the confidence interval $[0, \max(B_i)]$ covers p with a large probability close to one. That is, as $n_r \to \infty$,

$$\mathrm{P}(B_{(n_r)} > p) \to 1.$$

In practice, the exact CDF of B's F_B is unknown. So the corresponding empirical distribution $\hat{F}_B$ is estimated based on B_i's obtained from n_r OSF repetitions. As $n_r \to \infty$, $\hat{F}_B \to F_B$ uniformly almost surely. Therefore, as we control the number of repetitions n_r, F_B is practically known. The only problem here is to determine the tuning parameter N. Let b denote the true upper bound of the confidence interval, then by solving the following equation for the tuning parameter N,

$$b = F_B^{-1}(0.05^{1/N})$$

we have

$$N = \frac{\log 0.05}{\log F_B(b)}$$

Suppose b is close to the median of the B_i's, then $F_B(b) = \mathrm{P}(B \leq b) = 0.5$. From the above expression, $N \approx 5$. In some cases, b is close to the maximum of the B_i's. Let's say $F_B(b) = \mathrm{P}(B \leq b) \approx 0.99$, then $N \approx 300$ would yield the "correct" upper bound that reflect the truth. How to choose the optimal tuning parameter N is still an open problem. From many simulation results, the choice $N = 5$ suffices in many misspecified cases for small tail probabilities. This means that the ROSF confidence intervals yield desired coverage when $N = 5$. The choice $N = 100$ is more prudent across most misspecified cases. In some difficult cases (e.g. when $X_0 \sim$ Pareto), $N = 300$ is needed so that the confidence intervals give the desired coverage. In other words, in difficult cases, it is advisable to use the 99th quantile of the B_i's as the upper bound of the ROSF confidence interval.

3.3 Tilt Function Specification

In Chapter 2, a list the of one-to-one correspondence of the tilt function $h(x)$ and the probability density functions of the reference and fusion was given. Here we revisit the specification of the tilt function.

Consider now the case of two samples ($m = 1$): the reference sample $X_0 \sim g$ and the fusion sample $X_1 \sim g_1$. Suppose both g and g_1 are gamma densities. Then by the density ratio g_1/g, the corresponding tilt function $h(x) = (x, \log x)$. If the densities are right truncated at the same point $b > 0$, the resulting density ratio still holds with $h(x) = (x, \log x)$. The same holds true if g_1 is replaced by a uniform distribution on $(0, b)$. If g is indeed a gamma density and g_1 is uniform with sufficiently large support, then the density ratio is satisfied approximately with $h(x) = (x, \log x)$. In all simulation studies we present in the next section, the reference sample X_0 is fused with computer-generated uniform random samples X_1, and the threshold exceedance probability $P(X_0 > T)$ is estimated through DRM. Since X_1 must be generated from a uniform distribution with sufficiently large support, an appropriate upper limit of the uniform distribution must be determined. In all cases, the threshold T and the maximum value of X_0 are known. As a rule of thumb, the upper limit of the uniform distribution is chosen to be greater than both T and the maximum of X_0.

Under the mild assumption that 3.1 holds, it will be demonstrated later that in many cases, when g is skewed (not necessarily gamma density), then fusing the given reference sample X_0 repeatedly with computer generated random samples X_1 with uniform density g_1 and with the tilt function $h(x) = (x, \log x)$, the resulting confidence intervals for the threshold exceedance probabilities are surprisingly short and reliable. Furthermore, experiments also show that the log-normal tilt function $h(x) = (\log x, \log^2 x)$ is a useful alternative to the gamma tilt and could be used in the ROSF algorithm to accommodate a wide range of skewed reference distributions.

The normal tilt $h(x) = (x, x^2)$ obtained from the ratio of two normal densities is a reasonable choice when the reference sample X_0 comes from a symmetric or nearly symmetric distribution family. However, it is a poor choice when X_0 is highly skewed and typically causes computational issues when applied in ROSF.

It should also be noticed that, to generate the fusion samples X_j , the corresponding density function g_j must be specified. However, beyond the generating stage, this knowledge is not used in ROSF, and the algorithm proceeds as if the g_j are unknown. Thus, beyond the generating stage, in the semiparametric estimation both g and g_j, as well as the corresponding parameters, are all assumed unknown.

3.4 Simulation Studies: $h(x)$ Correctly Specified

In this section, BM, POT, OSF, ROSF, and other well-known methods of estimation of tail probabilities and associated intervals will be applied and compared in simulation studies.

In the context of quantitative risk assessment, data are usually highly skewed in nature. Furthermore, in many cases (e.g. number of patients in an early stage clinical trial), the size of the given reference sample is limited. Therefore, we focus our attention on moderately large samples that come from highly skewed distributions. In the simulation studies, the reference samples are randomly generated from a range of skewed distributions including: Exponential, Gamma, Log-normal, Weibull, Pareto, etc. Then the reference sample is fused with a computer generated uniform random sample. In all cases, the upper limit of the uniform distribution is chosen to be greater than both the threshold T and the largest value of X_0. As mentioned earlier, the density ratio model with both the gamma tilt $h(x) = (x, \log x)$ and the log-normal tilt $(\log x, \log^2(x))$ hold approximately when the reference sample comes from a skewed distribution. Unless otherwise specified, the gamma tilt $h(x) = (x, \log x)$ will be used as the default. For each case, the tail probabilities $p = 0.01$ and $p = 0.001$ will be considered. For each given tail probability, the correspond theoretical quantile will be used as the threshold T. For example, the 0.99 theoretical quantile for Exp(1.2) is 3.8376. Then the threshold exceedance probability $P(X > 3.8376) = 0.01$ is estimated based on the given sample using the methods we covered so far. Mean absolute error (MAE) which calculates the mean absolute value of the differences between the estimated threshold exceedance probability $\hat{p}$ and the true proababilty p is the metric that will be used to measure the precision of the point estimates of the tail probabilities. More importantly, the 95% confidence interval for this tail probability is calculated. The confidence intervals for OSF, AC and EP can be obtained through explicit expressions; however, computing the confidence intervals for BM and POT is quite involved.

To obtain the confidence intervals for the BM and POT method, it is natural to first consider using the *delta method*. However, when the sample size of the reference is small, the normal approximation to the distribution of the maximum likelihood estimator may be poor. Since the estimator of the tail probability p relies on the estimated parameters from fitting the GEV or GP distributions, the confidence interval of the estimated tail probability through the delta method is not reliable. Therefore, for the BM and POT methods, confidence intervals are bootstrapped with 500 replications. Since many of the bootstrapped estimates hit the 0 limit, the confidence intervals are constructed from the 0th to 95th percentiles.

To obtain reliable coverage results, in each study, five hundred confidence intervals are computed for each method. The performance of the methods are evaluated based on coverage, mean length of the confidence intervals, and mean absolute error which is the mean absolute difference between the estimated tail probability and the true tail probability.

The tilt function is correctly specified if it corresponds to the distribution of the reference sample correctly. For example, if the reference sample comes from the gamma family, the gamma tilt $h(x) = (x, \log x)$ is appropriate. Likewise, if the reference sample comes from the log-normal, family then the log-normal tilt would be appropriate. In this section, we will see, as a check that under correctly specified tilt functions where the density ratio model holds, the OSF method gives very precise point estimates for the threshold exceedance probability, and short yet reliable confidence intervals as well.

3.4.1 $X_0 \sim \text{Exp}(1.2)$

In our first example, $X_0 \sim \text{Exp}(1.2)$. In each simulation run, a sample of size 100 that follows Exp(1.2) is generated. The goal is to predict the threshold exceeding probability $\text{P}(X > t)$ for some large threshold t and construct its confidence interval using all methods introduced so far. For tail probability $p = 0.01$, the theoretical quantile $t = 3.8376$ is used as the threshold t. So the true probability $\text{P}(X > 3.8376)$

is 0.01. The performance is evaluated based on 500 such runs.

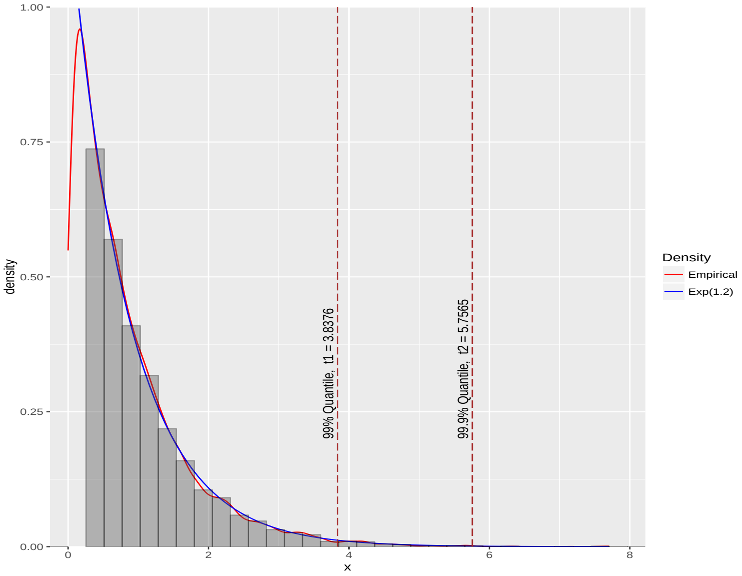

Figure 3.1: Density Plot for Exp(1.2)

For the OSF method, the fusion samples X_1 of size 100 are generated from Uniform(0,5). Exp(λ) distribution is a special case of the gamma distribution when the shape parameter $\alpha = 1$ and the scale parameter $\beta = \lambda$. The gamma tilt $h(x) = (x, \log x)$ is appropriate in this case and therefore adopted.

In this example, detailed procedures on how the point estimates and confidence intervals are obtained through Block Maxima (BM) and Peak Over Threshold (POT) approach will be given. Recall that, in Chapter 1, BM and POT are introduced as the two primary approaches to analyzing extremes of a data set. Block Maxima reduces the data by taking maxima of blocks of data of fixed length (e.g. annual maxima). Peak Over Threshold restricts attention to large observations that exceed a high threshold. Both methods reduces the size of the data, while OSF on the other hand *augments* the data. In practice, difficulties arise if one adopts two traditional approaches, BM and POT, based on the Extreme Value Theorem. For BM, the number of blocks must be determined before fitting GEV to the reduced data set, while for POT an adequate threshold has to be selected before GP is fitted to the reduced data set. If the number of blocks in BM or the threshold in POT are chosen too small, a biased sample results. In this case, small observations which are not qualified as extreme observations are included in the sample, and the EVT approximations would be violated. On the contrary, if the number of block in BM or the threshold in POT is chosen too large, the sample size would be too small to yield reliable estimates for the tail probabilities. The trade-off between a biased estimate and a large estimation error often puts practitioners in a dilemma when trying to apply EVT analysis.

For POT, these problems can be solved to some extent through graphical means such as a thresh range plot or a MRL plot. These graphical approaches do not pick a threshold for the user, and one may still have to rely on subjective decisions with the plots. For BM, if the underlying sample is a time series, the sample is typically blocked by some specified regular period (week, quarter, annual). Unfortunately, there is no general rule or guideline on how to choose the appropriate number of blocks when the underlying sample is not a time series. The researchers would need to choose a reasonable number of blocks with discretion.

In each run, we subdivide the sample of size 100 into 20 blocks of five observations in each block. The maximum observation from each block is taken to form a Block Maxima sample of size 20. Then the GEV distribution is fitted to the BM sample to obtain the distribution parameters and to derive the estimated threshold exceedance probability. The choice of the number of blocks is arbitrary. Different choices of this number are tested in many simulation studies, and the outcomes produced are quite similar.

The mean residual life (MRL) plot is used to help to determine an appropriate choice for the threshold u in the POT method. Recall that MRL plots the mean excess values for a range of threshold choices with uncertainty. For the reference sample of size 100 from Exp(1.2), we will restrict our attention to the range of zero to four. The idea behind a MRL plot is to select a threshold whereby the graph becomes linear as the threshold increases.

In Figure 3.2, we have the MRL plot for a random sample of size 100 from an Exp(1.2) distribution in the top panel, and the MRL plot for a random sample of size 1000 from the same distribution in the bottom panel. Both plots in this case are not very informative, since the mean excess stays linear for a wide range (0 to 2.25) in both cases. Therefore, to ensure we have enough data to produce relatively stable estimates, u is chosen to be 1.25; that is, approximately 20% of the sample is used for the fitting of the GP distribution.

Table 3.1: OSF Interval Coverage and Length for $p = 1 - G(T) = 0.01$, $T = 3.8376$, $X_0 \sim \text{Exp}(1.2)$, $X_1 \sim \text{Unif}(0, 5)$, $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

	OSF	AC	EP	POT	BM
Coverage	94.6%	96.8%	66.6%	76.4%	96.6%
CI Length	0.02456218	0.0600026	0.02729541	0.01973923	0.10109796
MAE	0.00533125	0.0192253	0.00782000	0.00598448	0.04208432

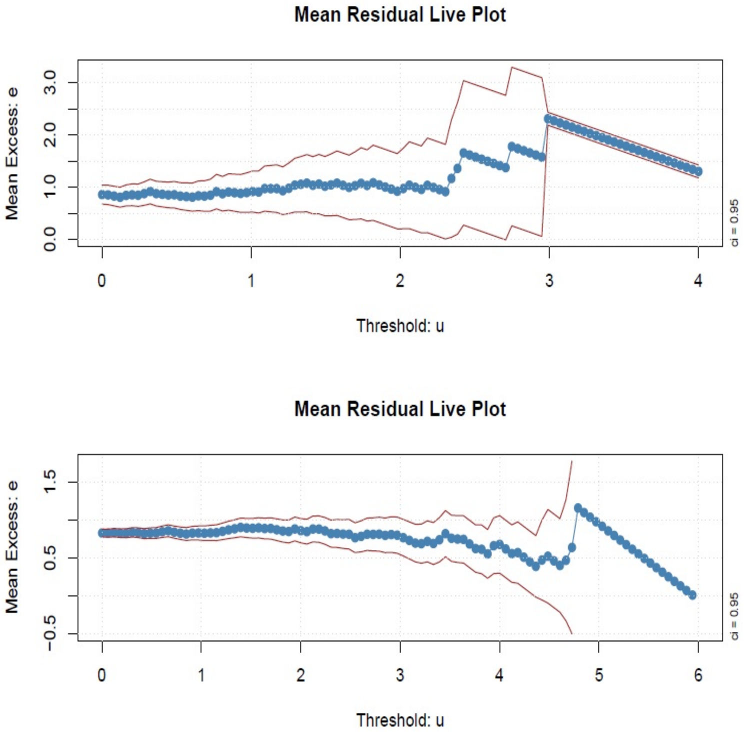

Figure 3.2: MRL Plot for Exp(1.2). Top sample size is 100; bottom sample size is 1000.

Figure 3.3 shows the coverage and histograms of the CI length of the first 100 confidence intervals obtained from various methods, when the true probability P($X > 3.8376$) is 0.01.

From Table 3.1, it can be observed that OSF outperformed the other methods in all aspect. Among all methods, OSF gives the smallest MAE which shows that the point estimate $\hat{p}_{OSF}$ has the best accuracy. Furthermore, the mean length of the confidence intervals associated with OSF is the shortest while the nominal coverage is maintained. In other words, the confidence intervals produced by OSF are short yet reliable.

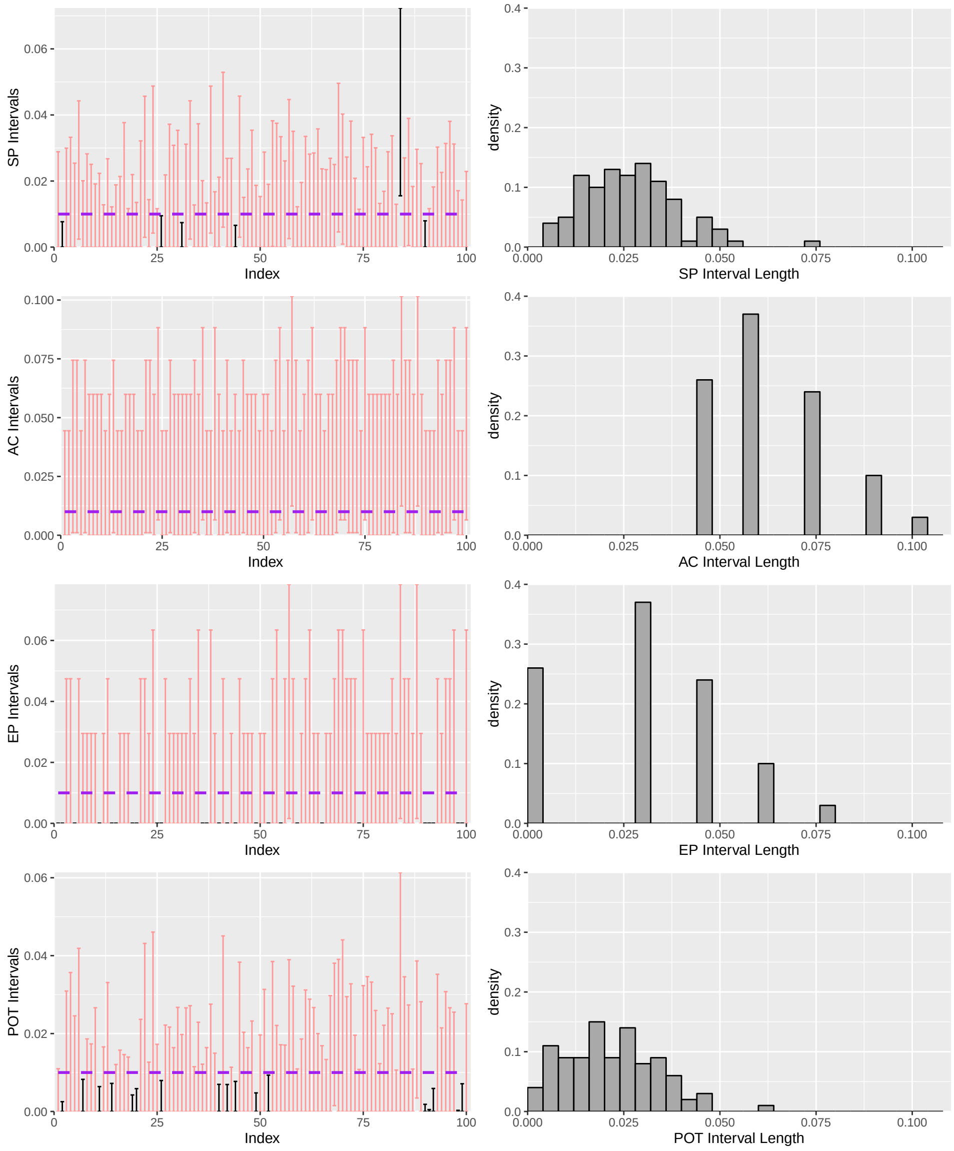

Figure 3.3: Coverage and Histograms of the Length of CI. The true probability $\mathrm{P}(X > 3.8376) = 0.01$.

For a tail probability of $p = 0.001$, the theoretical quantile $t = 5.7565$ is used as the threshold. Fusion samples X_1 are generated from Uniform(0,6.5).

This is a much smaller tail probability, the threshold $t = 5.7565$ is quite large for randomly generated samples of size 100 from an Exp(1.2) distribution. Numeri-

Table 3.2: OSF Interval Coverage and Length for $p = 1 - G(T) = 0.001$, $T = 5.7565$ $X_0 \sim \text{Exp}(1.2)$, $X_1 \sim \text{Unif}(0, 6.5)$, $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

	OSF	AC	EP	POT	BM
Coverage	95.8%	100%	8.5%	72.6%	94.6%
CI Length	0.00502374	0.0457728	0.00252677	0.00524260	0.03416001
MAE	0.00081316	0.0183449	0.00171600	0.00145271	0.00991796

cal experiments show that out of 100,000 randomly generated Exp(1.2) samples of size 100, 90.432% samples do not contain observations greater than the threshold $t = 5.7565$. This means that over 90% of the samples contains no success at all. When the sample contains no success, the estimator of the binomial probability $\sum X_i/n$ is simply 0. Therefore, there's no surprise that the EP method has the worst performance. Only 4 EP intervals cover the true tail probability $p = 0.001$. AC intervals yield 100% coverage, but this method produces the same estimates all the time. The resulting point estimate and the confidence intervals are too large to have any practical meaning. POT intervals also become worse for $p = 0.001$. In this particular simulation, only about 73% intervals covered the true tail probability. OSF method shows its true potential when the tail probability gets this small. The point estimates $\hat{p}_{OSF}$ has the smallest MAE. The associated CI not only has the shortest length but a high coverage rate as well. The nominal 95% coverage is achieved in this simulation. The coverage and histograms of the CI length of the first 100 confidence intervals obtained from various methods when the true probability $\text{P}(X > 5.7565) = 0.001$ is shown in Figure 3.4.

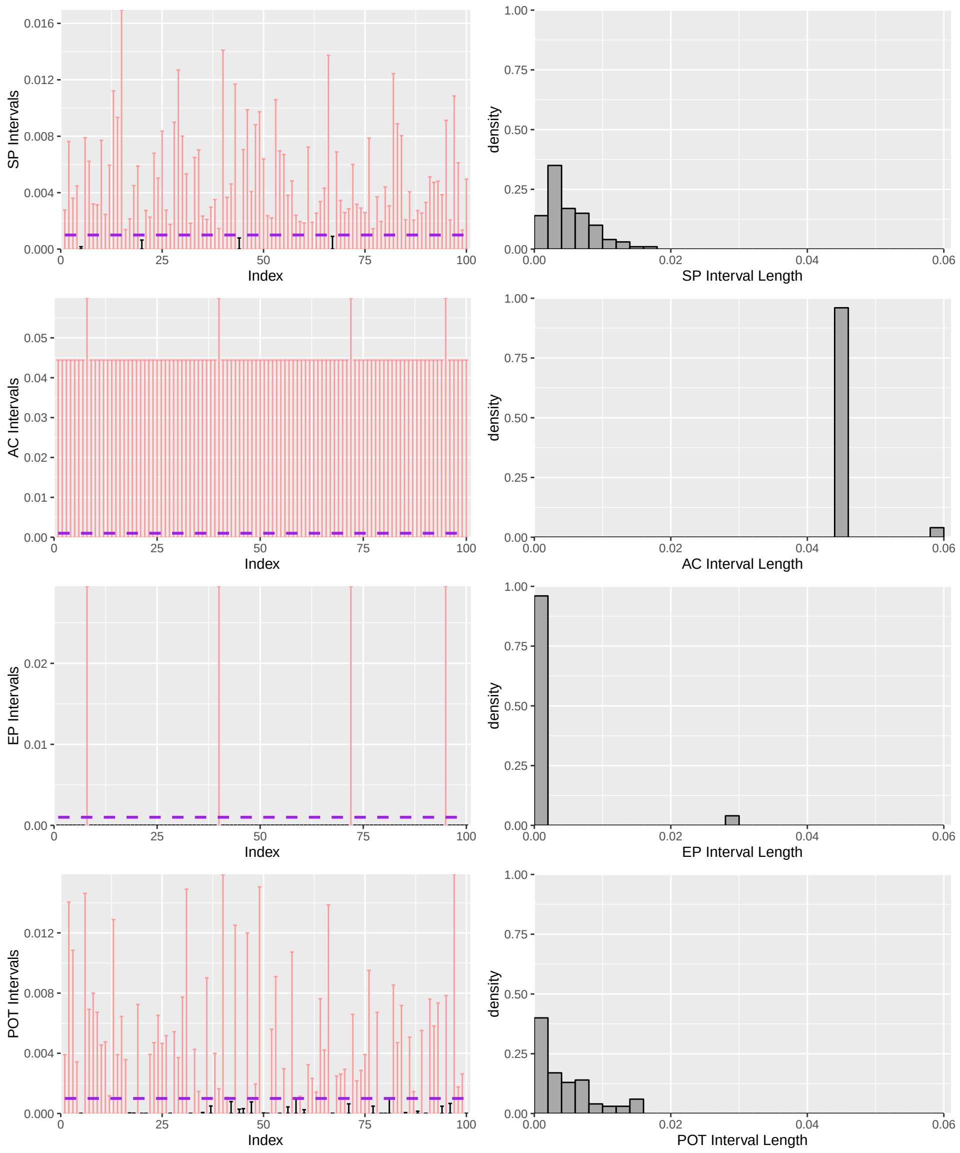

Figure 3.4: Coverage and Histograms of the Length of CI. The true probability $\mathrm{P}(X > 5.7565) = 0.001$.

3.4.2 $X_0 \sim \mathrm{Gamma}(5, 3)$

In this and the next subsection, we demonstrate the results when the reference sample X_0 is simulated from two different gamma distributions. The histograms and the density plots are given in Figure 3.5. The correctly specified gamma tilt function $h(x) = (x, \log x)$ is adopted for both cases; thus, the density ratio model holds.

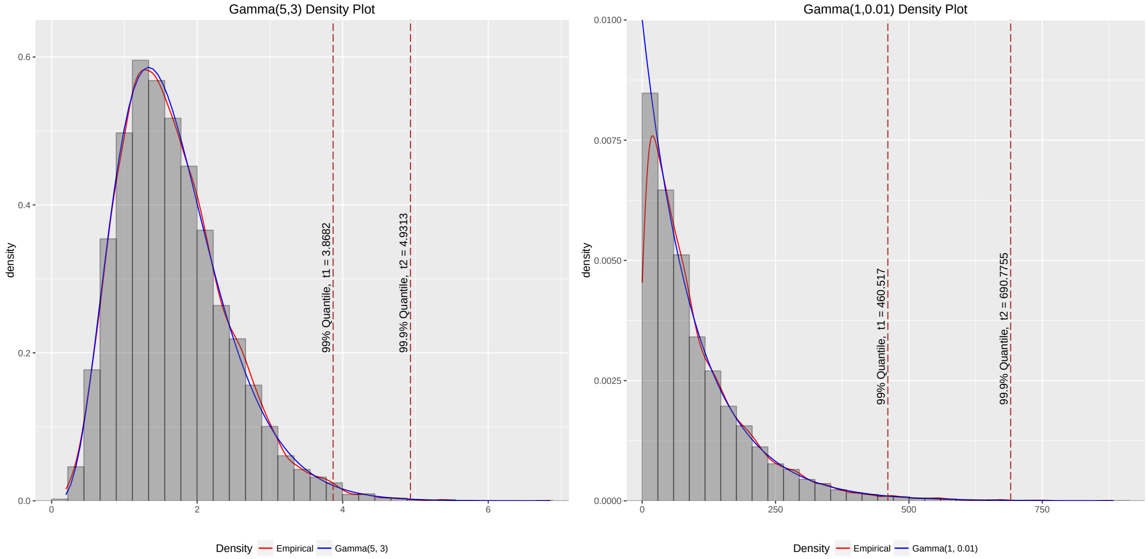

Figure 3.5: Density Plot for Gamma(5,3) and Gamma(1,0.01)

In this example, $X_0 \sim \text{Gamma}(5, 3)$. For tail probability $p = 0.01$, the theoretical quantile $t = 3.8682$ is used as the threshold. Fusion samples X_1 are generated from a Uniform(0,6). For tail probability $p = 0.001$, the theoretical quantile $t = 4.9314$ is used as the threshold. Fusion samples X_1 are generated from a Uniform(0,7). The gamma tilt function $h(x) = (x, \log x)$ is appropriate and therefore adopted.

3.4.3 $X_0 \sim \text{Gamma}(1, 0.01)$

In this example, $X_0 \sim \text{Gamma}(1, 0.01)$. For tail probability $p = 0.01$, the theoretical quantile $t = 460.517$ is used as the threshold. Fusion samples X_1 are

Table 3.3: OSF Interval Coverage and length for $p = 1 - G(T) = 0.01/T = 3.8682$ and $p = 1 - G(T) = 0.001/T = 4.9314$, $X_0 \sim \text{Gamma}(5, 3)$, $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

p/Threshold	Method	Fusion Sample X_1	Coverage	CI Length	MAE
$p = 0.01$, $T = 3.87$	OSF	Unif(0, 6)	94.2%	0.02351581	0.00536009
	AC	-	97.6%	0.05834720	0.01797340
	EP	-	63.4%	0.02465786	0.00716000
	POT	-	65.4%	0.01789809	0.00664344
	BM	-	94.2%	0.09298356	0.03764912
$p = 0.001$, $T = 4.93$	OSF	Unif(0, 7)	98.2%	0.00549161	0.00082905
	AC	-	99.4%	0.04613860	0.01857600
	EP	-	10.6%	0.00323482	0.00190800
	POT	-	61.0%	0.00414608	0.00134398
	BM	-	91.4%	0.02869033	0.00826318

generated from Uniform(0,600). For tail probability $p = 0.001$, the theoretical quantile $t = 690.7755$ is used as the threshold. Fusion samples X_1 are generated from Uniform(0,800). The gamma tilt function $h(x) = (x, \log x)$ is appropriate and therefore adopted.

3.4.4 $X_0 \sim \text{Log-Normal}(1, 1)$

In this and the next subsection, we demonstrate the results when the reference sample X_0 is simulated from two different log-normal distributions. The histograms and the density plots are given in Figure 3.6. The correctly specified gamma tilt function $h(x) = (\log x, \log^2 x)$ is adopted for both cases; thus the density ratio model holds.

In this example, $X_0 \sim \text{Log-Normal}(1, 1)$. For tail probability $p = 0.01$, the theoretical quantile $t = 27.83649$ is used as the threshold. The fusion samples X_1

Table 3.4: OSF Interval Coverage and length for $p = 1 - G(T) = 0.01/T = 460.517$ and $p = 1 - G(T) = 0.001/T = 690.7755$, $X_0 \sim$ Gamma$(1, 0.01)$, $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

p/Threshold	Method	Fusion Sample X_1	Coverage	CI Length	MAE
$p = 0.01$, $T = 460.72$	OSF	Unif(0, 600)	95.4%	0.02407956	0.00510845
	AC	-	97.4%	0.05918750	0.01849340
	EP	-	65.4%	0.02590928	0.00730000
	POT	-	75.6%	0.01935038	0.00581125
	BM	-	98.2%	0.09923025	0.04095700
$p = 0.001$, $T = 690.78$	OSF	Unif(0, 800)	97.4%	0.00508180	0.00075557
	AC	-	99.8%	0.04577470	0.01834490
	EP	-	8.6%	0.00257303	0.00170800
	POT	-	80.0%	0.00555302	0.00143851
	BM	-	97.6%	0.03638597	0.01113462

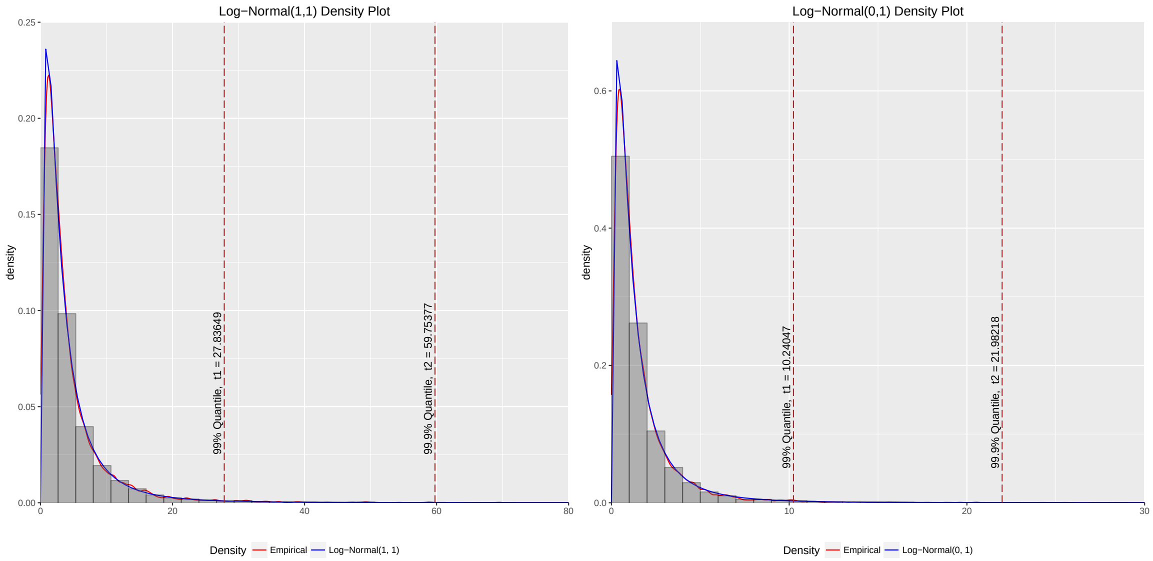

Figure 3.6: Density Plot for Log-Normal(1,1) and Log-Normal(0,1)

are generated from Uniform(1,65). For tail probability $p = 0.001$, the theoretical quantile $t = 59.75377$ is used as the threshold. Fusion samples X_1 are generated from Uniform(1,85). In this case, the log-normal tilt function $h(x) = (\log x, \log^2 x)$

is appropriate and therefore adopted.

Table 3.5: OSF Interval Coverage and length for $p = 1 - G(T) = 0.01/T = 27.83649$ and $p = 1 - G(T) = 0.001/T = 59.75377$, $X_0 \sim$ Log-Normal$(1,1)$, $n_0 = n_1 = 100$, $h(x) = (\log x, (\log x, \log^2 x)$

p/Threshold	Method	Fusion Sample X_1	Coverage	CI Length	MAE
$p = 0.01$, $T = 27.84$	OSF	Unif(1, 65)	96.6%	0.02838072	0.00619843
	AC	-	98.6%	0.05842430	0.01787710
	EP	-	64.6%	0.02473896	0.00682000
	POT	-	77.6%	0.02030155	0.00574118
	BM	-	99.0%	0.10742792	0.04359354
$p = 0.001$, $T = 59.75$	OSF	Unif(1, 85)	98.8%	0.00672108	0.00114244
	AC	-	99.8%	0.04571260	0.01830630
	EP	-	8.2%	0.00245502	0.00167600
	POT	-	80.2%	0.00565705	0.00139721
	BM	-	99.0%	0.03996896	0.01116874

3.4.5 $X_0 \sim$ Log-Normal$(0, 1)$

In this example, $X_0 \sim$ Log-Normal$(0, 1)$. For tail probability $p = 0.01$, the theoretical quantile $t = 10.24047$ is used as the threshold. Fusion samples X_1 are generated from Uniform(1,20). For tail probability $p = 0.001$, the theoretical quantile $t = 21.98218$ is used as the threshold. Fusion samples X_1 are generated from Uniform(1,35). In this case, the log-normal tilt function $h(x) = (\log x, \log^2 x)$ is appropriate and therefore adopted.

3.5 Simulation Studies: $h(x)$ Misspecified

The key assumption of the density ratio model is that the log ratio of two unknown probability density functions takes some known linear form which depends

Table 3.6: OSF Interval Coverage and length for $p = 1 - G(T) = 0.01/T = 10.24047$ and $p = 1 - G(T) = 0.001/T = 21.98218$, $X_0 \sim$ Log-Normal$(0, 1)$, $n_0 = n_1 = 100$, $h(x) = (\log x, \log^2 x)$

p/Threshold	Method	Fusion Sample X_1	Coverage	CI Length	MAE
$p = 0.01$, $T = 10.24$	OSF	Unif(1, 65)	96.2%	0.02916668	0.00667020
	AC	-	98.8%	0.05851740	0.01789630
	EP	-	65.6%	0.02492600	0.00664000
	POT	-	79.0%	0.01976141	0.00554119
	BM	-	99.0%	0.10284490	0.04184385
$p = 0.001$, $T = 21.98$	OSF	Unif(1, 85)	98.2%	0.00783214	0.00149434
	AC	-	99.8%	0.04586780	0.01840260
	EP	-	9.2%	0.00275004	0.00175600
	POT	-	82.8%	0.00553023	0.00131581
	BM	-	98.2%	0.03714942	0.01096847

on finite dimensional parameters. In the previous section, it was shown that when the tilt function $h(x)$ is correctly specified, the tail probability estimated by the OSF method is precise and the associated confidence intervals are short and reliable. Fokianos and Kaimi (2006) [21] studied the effect of misspecified tilt functions by embedding the unknown linear form to some parametric transformation family which leads ultimately to its identification. In this section, how the misspecification issue could be addressed by Repeated Out of Sample Fusion (ROSF) will be presented in simulation examples. Random samples generated from various types of skewed distributions will be treated as the reference sample. Skewed distributions we considered in the simulation studies include: Pareto, F, Inverse Gaussian, Weibull, and Truncated Gumbel. See Figure 3.7 for thedensity plot for various types of skewed distributions we considered in this section. Instead of using the correct tilt function for each case, the gamma tilt function $h(x) = (x, \log x)$ will be used

Table 3.7: OSF Interval Coverage and Length for $p = 1 - G(T) = 0.01$, $T = 3.1623$, $X_0 \sim \text{Pareto}(1, 4)$, $X_1 \sim \text{Unif}(1, 5.5)$, $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

	OSF	AC	EP	POT	BM
Coverage	78.4%	98.8%	64.8%	77.0%	97.8%
CI Length	0.02291682	0.0583229	0.02458624	0.01962555	0.10122122
MAE	0.00702798	0.0177615	0.00666000	0.00600362	0.04120398

throughout.

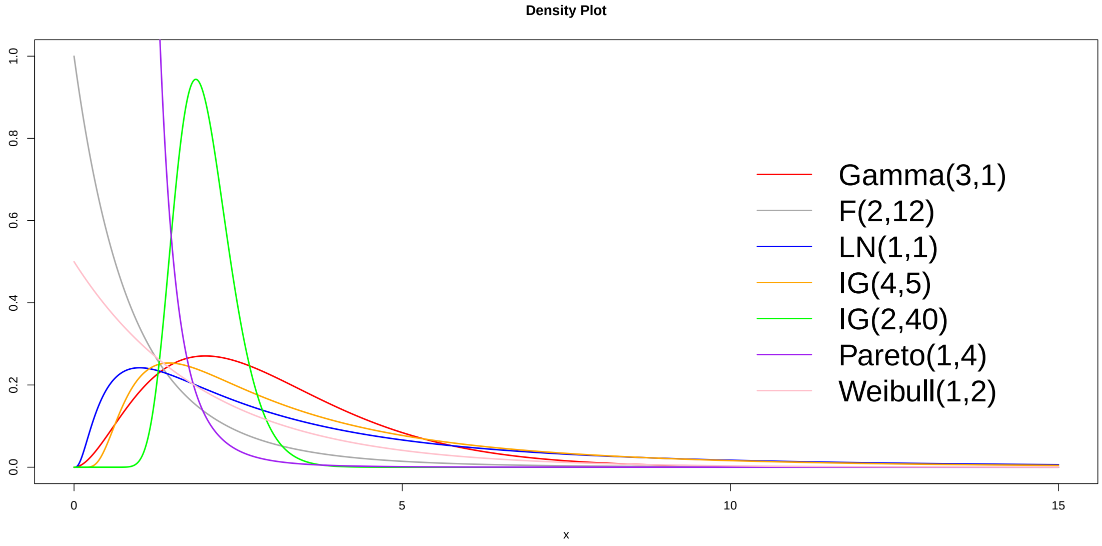

Figure 3.7: Density Plot for Various Types of Skewed Distributions

3.5.1 $X_0 \sim \text{Pareto}(1, 4)$

In this example, $X_0 \sim \text{Pareto}(1, 4)$. For tail probability $p = 0.01$, the theoretical quantile $t = 3.1623$ is used as the threshold. Fusion samples X_1 are generated from Uniform(1,5.5). In this case, the gamma tilt function $h(x) = (x, \log x)$ is inappropriate and the density ratio model is misspecified.

From Table 3.7, the OSF confidence interval no longer has the nominal 95% coverage when the tilt function is misspecified. Nevertheless, the estimated tail probabilities given by the non-parametric methods (AC and EP) and EVT based methods(POT and BM) are not satisfactory either. But, ROSF fixes the situation.

The ROSF procedure to estimate the tail probability and confidence interval is described in detail in section 3.2. Here we summarize the ROSF algorithm:

1. Fuse the given reference sample X_0 with the computer generated sample (fusion sample) X_1, and obtain the estimated tail probability $\hat{p}$ and its confidence interval $[0, B_1]$ through the Density Ratio Model.

2. Fuse the given reference sample X_0 again with another computer generated sample (of the same type of the previous artificial sample and independent of it) to get another $\hat{p}$ and confidence interval $[0, B_2]$ in the same manner as in step one.

3. Repeat the process n_r times to produce a sequence of $\hat{p}_i$ and confidence intervals $[0, B_i], i = 1, \ldots, n_r$.

4. Obtain the empirical distribution for the upper bounds B_i and denote it as $\hat{F}_B$. Based on computational efficiency consideration, in our simulation studies $\hat{F}_B$ is obtained from $n_r = 200$ B's.

5. Obtain the ROSF point estimate of the tail probability by $\hat{p} = \hat{\mathrm{P}}(X_0 > t) = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_i$ and the confidence interval by $[0, F_B^{-1}(\alpha^{1/N})]$.

Table 3.8: ROSF Interval Coverage and Length for $p = 1-G(T) = 0.01$, $T = 3.1623$, $X_0 \sim \text{Pareto}(1, 4)$, $X_1 \sim \text{Unif}(1, 5.5)$, $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

	N=5	N=20	N=50	N=100	N=300
Coverage	79.8%	86.8%	90.0%	90.6%	92.8%
CI Length	0.02336956	0.0259955	0.02730116	0.02822144	0.02958183
MAE	0.00686935				

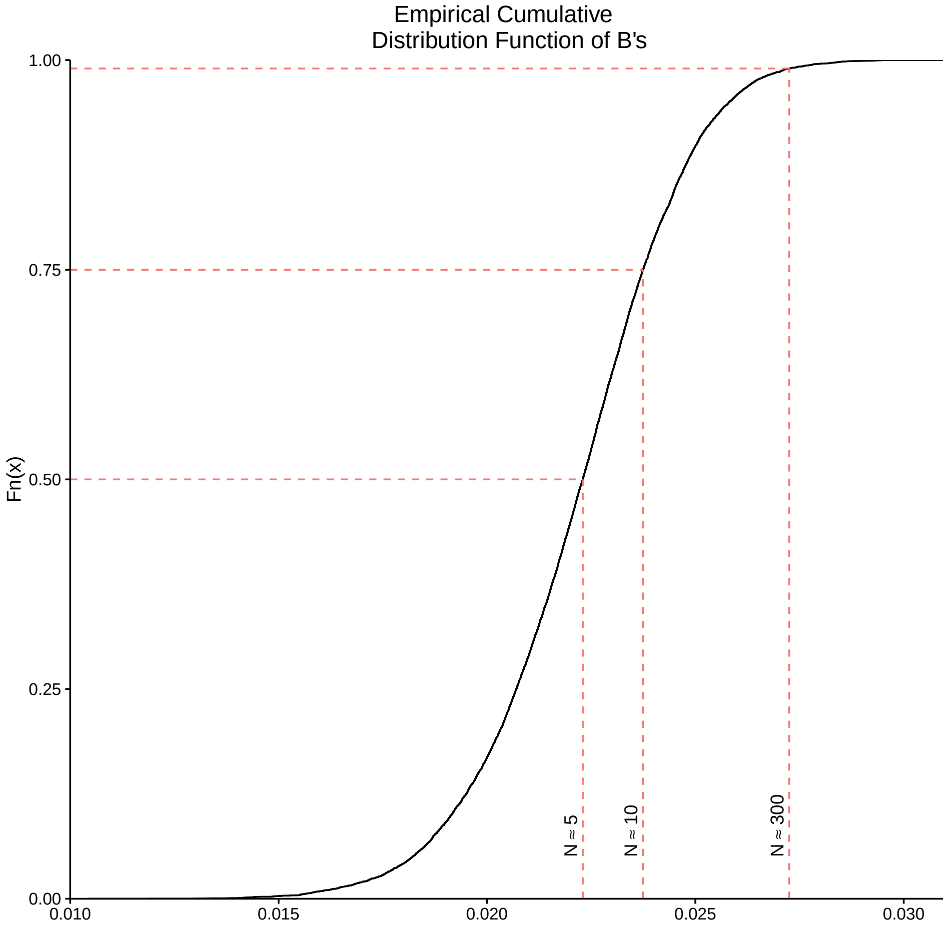

Figure 3.8: Typical ECDF Plot of the ROSF Upper Bounds B's

Figure 3.8 shows the empirical CDF of the ROSF upper bounds. The ECDF is obtained based on $n_r = 10,000$ B's for illustrative purposes. Note that the choice $N = 5$ approximately corresponds to the median of the B's, $N = 10$ roughly gives the third quantile of the B's, and $N = 300$ yields the 99th quantile. The coverage increases as N increases. However, the length of the confidence intervals also increases with N. The subject on how to determine the optimal N is discussed in section 3.2. In the simulation studies results from $N = 5$ to $N = 300$ are provided. As we will see soon, in many misspecified cases, the choice $N = 5$ suffices, as the resulted point estimate is precise (low MAE) and the coverage is high. This case when X_0 follows a Pareto distribution is a difficult case, and the choice of $N = 300$ yields confidence intervals with the targeted coverage. Since $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} B_i$, the point estimate does not depend on the choice of N. The mean absolute deviation (MAE) is also independent of N. The MAE obtained from 500 simulation runs in this case is 0.006869346. The ROSF coverage, CI length, and MAE for various choices of N can be found in Table 3.8 Note that the ROSF point estimate of the tail probability is $\hat{p} = \sum_i \hat{p}_i / n_r$ and does not depend on the tuning parameter N. Therefore, the MAE stays the same across all N values and we report it only once.

For tail probability $p = 0.001$, the theoretical quantile $t = 5.6234$ is used as the threshold. Fusion samples X_1 are generated from Uniform(1,8). Detailed results are given in Table 3.9

From Table 3.9, the coverage for OSF does not reach the desired 95 percent. However, it is clear that the condition $\mathrm{P}(B > p) > 0$ holds so the ROSF can be applied. As N becomes larger, the ROSF algorithm gradually improves the interval coverage and the MAE of the point estimate at the expense of slightly increased interval length. The desired coverage is reached when $N = 300$. Again, the ROSF point estimate does not depend on the tuning parameter N and we report it only

Table 3.9: OSF and ROSF interval coverage and length for $p = 1 - G(T) = 0.001/T = 5.6234$, $X_0 \sim \text{Pareto}(1,4)$, $X_1 \sim \text{Unif}(1,8)$, $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

Method	N	Coverage	CI Length	MAE
OSF	-	70.6%	0.00511820	0.00154342
AC	-	99.6%	0.04567740	0.01828710
EP	-	7.8%	0.00237289	0.00166400
POT	-	77.8%	0.00527968	0.00138423
BM	-	96.4%	0.03476016	0.00990250
ROSF	5	72.8%	0.00511630	0.00143907
	25	85.6%	0.00666390	-
	50	89.6%	0.00748020	-
	100	92.4%	0.00806978	-
	300	96.4%	0.00892499	-

once.

3.5.2 $X_0 \sim \text{F}(2,12)$

In this example, $X_0 \sim \text{F}(2,12)$. For tail probability $p = 0.01$, the theoretical quantile $t = 6.9266$ is used as the threshold. Fusion samples X_1 are generated from Uniform(0,10). In this case, the gamma tilt function $h(x) = (x, \log x)$ is inappropriate and the density ratio model is misspecified.

From Table 3.10, the OSF confidence interval no longer has the nominal 95% coverage when the tilt function is misspecified. Likewise, the estimated tail probabilities obtained by the non-parametric methods (AC and EP) and EVT based methods(POT and BM) are not satisfactory either. As N becomes larger, the ROSF algorithm gradually improves the interval coverage and the MAE of the point estimate at the expense of slightly increased interval length. The desired coverage is

Table 3.10: OSF and ROSF interval coverage and length for $p = 1 - G(T) = 0.01/T = 6.9266$, $X_0 \sim \text{F}(2, 12)$, $X_1 \sim \text{Unif}(0, 10)$, $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

Method	N	Coverage	CI Length	MAE
OSF	-	79.0%	0.02041963	0.00618648
AC	-	98.2%	0.05942880	0.01828710
EP	-	65.8%	0.02638146	0.00748000
POT	-	75.6%	0.01976238	0.00624685
BM	-	97.0%	0.10230150	0.04257300
ROSF	5	82.8%	0.02070051	0.00583096
	25	90.6%	0.02536070	-
	50	93.6%	0.02779773	-
	100	95.0%	0.02945160	-
	300	96.6%	0.03184255	-

reached when $N = 100$. The ROSF point estimate does not depend on the tuning parameter N and we report it only once.

For tail probability $p = 0.001$, the theoretical 0.999 quantile $t = 12.9737$ is used as the threshold. Fusion samples X_1 are generated from Uniform(1,16). Detailed results are given in Table 3.11.

Similar results are obtained in this case. The coverage for OSF does not reach the desired 95 percent. However, as N becomes larger, the ROSF algorithm gradually improves the interval coverage and the MAE of the point estimate at the expense of slightly increased interval length. The desired coverage is reached when $N = 25$. The ROSF point estimate does not depend on the tuning parameter N and we report it only once.

Table 3.11: OSF and ROSF interval coverage and length for $p = 1 - G(T) = 0.001/T = 12.9737$, $X_0 \sim \mathrm{F}(2, 12)$, $X_1 \sim \mathrm{Unif}(0, 16)$, $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

Method	N	Coverage	CI Length	MAE
OSF	-	91.2%	0.00647695	0.00134798
AC	-	99.6%	0.04573940	0.01832560
EP	-	8.2%	0.00249090	0.00169600
POT	-	76.0%	0.00530864	0.00143199
BM	-	96.0%	0.03557303	0.01037317
ROSF	5	94.2%	0.00660943	0.00080944
	25	97.4%	0.00805930	-
	50	99.0%	0.00881673	-
	100	99.2%	0.00935866	-
	300	99.4%	0.01020506	-

3.5.3 $X_0 \sim$ Inverse-Gaussian$(4, 5)$

In this example, $X_0 \sim \mathrm{IG}(4, 5)$. For tail probability $p = 0.01$, the theoretical quantile $t = 17.87176$ is used as the threshold. Fusion samples X_1 are generated from Uniform(1,30). In this case, the gamma tilt function $h(x) = (x, \log x)$ is inappropriate and the density ratio model is misspecified.

From Table 3.12, the coverage for OSF does not reach the desired 95 percent. However, as N becomes larger, the ROSF algorithm gradually improves the interval coverage and the MAE of the point estimate at the expense of slightly increased interval length. The desired coverage is reached when $N = 25$. The ROSF point estimate does not depend on the tuning parameter N and we report it only once.

For tail probability $p = 0.001$, the theoretical quantile $t = 28.95409$ is used as the threshold. Fusion samples X_1 are generated from Uniform(1,35). Detailed

Table 3.12: OSF and ROSF interval coverage and length for $p = 1 - G(T) = 0.01/T = 17.87176$, $X_0 \sim \mathrm{IG}(4, 5)$, $X_1 \sim \mathrm{Unif}(1, 30)$, $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

Method	N	Coverage	CI Length	MAE
OSF	-	87.2%	0.02416190	0.00582463
AC	-	98.4%	0.05848490	0.01799260
EP	-	64.2%	0.02486776	0.00702000
POT	-	77.8%	0.01948964	0.00575958
BM	-	99.2%	0.10354190	0.04137860
ROSF	5	89.6%	0.02435503	0.00558762
	25	95.8%	0.02839780	-
	50	98.0%	0.03049135	-
	100	98.6%	0.03193300	-
	300	99.2%	0.03398256	-

results are given in Table 3.13.

Table 3.13: OSF and ROSF interval coverage and length for $p = 1 - G(T) = 0.001/T = 28.95409$, $X_0 \sim \mathrm{IG}(4, 5)$, $X_1 \sim \mathrm{Unif}(1, 35)$, $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

Method	N	Coverage	CI Length	MAE
OSF	-	88.4%	0.00468428	0.00950715
AC	-	99.6%	0.04580150	0.01836410
EP	-	8.6%	0.00260890	0.00172800
POT	-	81.4%	0.00602746	0.00160986
BM	-	98.8%	0.04122780	0.01284510
ROSF	5	91.4%	0.00462069	0.00087914
	25	98.4%	0.00631720	-
	50	99.6%	0.00728733	-
	100	99.6%	0.00803363	-
	300	99.8%	0.00922135	-

Similar results are obtained in this case. The coverage for OSF does not reach the desired 95 percent. However, as N becomes larger, the ROSF algorithm gradually improves the interval coverage and the MAE of the point estimate at the expense of slightly increased interval length. The desired coverage is reached when $N = 25$. The ROSF point estimate does not depend on the tuning parameter N and we report it only once.

3.5.4 $X_0 \sim$ Inverse-Gaussian$(2, 40)$

In this example, $X_0 \sim \text{IG}(2, 40)$. For tail probability $p = 0.01$, the theoretical quantile $t = 3.257718$ is used as the threshold. Fusion samples X_1 are generated from Uniform(0,5). In this case, the gamma tilt function $h(x) = (x, \log x)$ is inappropriate and the density ratio model is misspecified.

Table 3.14: OSF and ROSF interval coverage and length for $p = 1 - G(T) = 0.01/T = 3.257718$, $X_0 \sim \text{IG}(2, 40)$, $X_1 \sim \text{Unif}(0, 5)$, $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

Method	N	Coverage	CI Length	MAE
OSF	-	85.0%	0.02035902	0.00518535
AC	-	98.4%	0.05942890	0.01864750
EP	-	67.8%	0.02647355	0.00698000
POT	-	67.0%	0.01857640	0.00656944
BM	-	95.8%	0.09546870	0.03841630
ROSF	5	90.4%	0.02034459	0.00451952
	25	96.6%	0.02566740	-
	50	98.6%	0.02850117	-
	100	99.2%	0.03055052	-
	300	99.4%	0.03372030	-

From Table 3.14, the coverage for OSF does not reach the desired 95 percent. However, as N becomes larger, the ROSF algorithm gradually improves the interval coverage and the MAE of the point estimate at the expense of slightly increased interval length. The desired coverage is reached when $N = 25$. The ROSF point estimate does not depend on the tuning parameter N and we report it only once.

For tail probability $p = 0.001$, the theoretical quantile $t = 3.835791$ is used as the threshold. Fusion samples X_1 are generated from Uniform(0,6). Detailed results are given in Table 3.15.

Table 3.15: OSF and ROSF interval coverage and length for $p = 1 - G(T) = 0.001/T = 3.835791$, $X_0 \sim \mathrm{IG}(2, 40)$, $X_1 \sim \mathrm{Unif}(0, 6)$, $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

Method	N	Coverage	CI Length	MAE
OSF	-	89.6%	0.00375765	0.00776180
AC	-	99.8%	0.04611600	0.01855670
EP	-	10.8%	0.00322207	0.00188400
POT	-	59.4%	0.00404488	0.00135171
BM	-	90.6%	0.02792973	0.00797250
ROSF	5	94.8%	0.00362604	0.00065085
	25	98.6%	0.00530770	-
	50	99.4%	0.00635649	-
	100	99.6%	0.00720078	-
	300	99.8%	0.00862340	-

Similar results are obtained in this case. The coverage for OSF does not reach the desired 95 percent. However, as N becomes larger, the ROSF algorithm gradually improves the interval coverage and the MAE of the point estimate at the expense of slightly increased interval length. The desired coverage is reached when $N = 25$. The ROSF point estimate does not depend on the tuning parameter N and we report it only once.

3.5.5 $X_0 \sim \text{Weibull}(1,2)$

In this example, $X_0 \sim \text{Weibull}(1,2)$. For tail probability $p = 0.01$, the theoretical quantile $t = 9.21034$ is used as the threshold. Fusion samples X_1 are generated from Uniform(0,12). In this case, the gamma tilt function $h(x) = (x, \log x)$ is inappropriate and the density ratio model is misspecified.

Table 3.16: OSF and ROSF interval coverage and length for $p = 1 - G(T) = 0.01/T = 9.21034$, $X_0 \sim \text{Weibull}(1,2)$, $X_1 \sim \text{Unif}(0,12)$, $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

Method	N	Coverage	CI Length	MAE
OSF	-	92.4%	0.02257618	0.00489417
AC	-	97.8%	0.05882050	0.01826230
EP	-	64.8%	0.02537867	0.00718000
POT	-	73.8%	0.01871685	0.00582358
BM	-	98.4%	0.09990350	0.03900090
ROSF	5	95.4%	0.02284684	0.00422050
	25	97.8%	0.02776200	-
	50	98.4%	0.03034152	-
	100	99.8%	0.03215609	-
	300	99.8%	0.03498779	-

From Table 3.16, the coverage for OSF does not reach the desired 95 percent. However, as N becomes larger, the ROSF algorithm gradually improves the interval coverage and the MAE of the point estimate at the expense of slightly increased interval length. The desired coverage is reached when $N = 5$. The ROSF point estimate does not depend on the tuning parameter N and we report it only once.

For tail probability $p = 0.001$, the theoretical quantile $t = 13.81551$ is used as the threshold. Fusion samples X_1 are generated from Uniform(0,16). Detailed results are given in Table 3.17.

Table 3.17: OSF and ROSF interval coverage and length for $p = 1 - G(T) = 0.001/T = 13.81551$, $X_0 \sim \text{Weibull}(1, 2)$, $X_1 \sim \text{Unif}(0, 16)$, $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

Method	N	Coverage	CI Length	MAE
OSF	-	96.4%	0.00481890	0.00072284
AC	-	99.6%	0.04586350	0.01840260
EP	-	9.0%	0.00272691	0.00176000
POT	-	77.2%	0.00530155	0.00140549
BM	-	98.0%	0.04307054	0.01043342
ROSF	5	98.2%	0.00476868	0.00059276
	25	99.8%	0.00653690	-
	50	100%	0.00754933	-
	100	100%	0.00830616	-
	300	100%	0.00954166	-

Similar results are obtained in this case. The coverage for OSF actually reaches the desired 95 percent. As N becomes larger, the ROSF algorithm improves the interval coverage and the MAE of the point estimate at the expense of slightly increased interval length. The desired coverage is reached when $N = 5$. The ROSF point estimate does not depend on the tuning parameter N and we report it only once.

3.5.6 $X_0 \sim$ Truncated Gumbel$(0, 5)$

In this example, $X_0 \sim$ Truncated Gumbel$(0, 5)$. For tail probability $p = 0.01$, the 99% quantile $t = 25.36103$ is used as the threshold. Fusion samples X_1 are generated from Uniform(0,28). In this case, the gamma tilt function $h(x) = (x, \log x)$ is inappropriate and the density ratio model is misspecified.

Table 3.18: OSF and ROSF interval coverage and length for $p = 1 - G(T) = 0.01/T = 25.36103$, $X_0 \sim$ Trunc. Gumbel$(0, 5)$, $X_1 \sim$ Unif$(0, 28)$, $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

Method	N	Coverage	CI Length	MAE
OSF	-	86.4%	0.01766805	0.00463530
AC	-	98.4%	0.05855340	0.01812750
EP	-	63.6%	0.02502152	0.00728000
POT	-	69.6%	0.01828719	0.00641836
BM	-	98.0%	0.12222860	0.03820720
ROSF	5	91.4%	0.01790894	0.00425766
	25	97.6%	0.02232830	-
	50	99.2%	0.02476274	-
	100	99.2%	0.02652814	-
	300	99.6%	0.02926242	-

From Table 3.18, the coverage for OSF does not reach the desired 95 percent. As N becomes larger, the ROSF algorithm gradually improves the interval coverage and the MAE of the point estimate at the expense of slightly increased interval length. The desired coverage is reached when $N = 25$. The ROSF point estimate does not depend on the tuning parameter N and we report it only once.

For tail probability $p = 0.001$, the 99.9% quantile $t = 37.20$ is used as the threshold. Fusion samples X_1 are generated from Uniform(0,40). Detailed results are given in Table 3.19.

Similar results are obtained in this case. The coverage for OSF actually reaches the desired 95 percent. As N becomes larger, the ROSF algorithm gradually improves the interval coverage and the MAE of the point estimate at the expense of slightly increased interval length. The desired coverage is reached when $N = 25$. The ROSF point estimate does not depend on the tuning parameter N and we

Table 3.19: OSF and ROSF interval coverage and length for $p = 1 - G(T) = 0.001/T = 37.20$, $X_0 \sim$ Trunc. Gumbel$(0, 5)$, $X_1 \sim$ Unif$(0, 40)$, $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

Method	N	Coverage	CI Length	MAE
OSF	-	96.4%	0.00390412	0.00059924
AC	-	99.4%	0.04582830	0.01838340
EP	-	8.6%	0.00264478	0.00174800
POT	-	72.2%	0.00474003	0.00140497
BM	-	96.6%	0.05526941	0.00940537
ROSF	5	99.2%	0.00387261	0.00053174
	25	100%	0.00528320	-
	50	100%	0.00611575	-
	100	100%	0.00675228	-
	300	100%	0.00781973	-

report it only once.

3.6 Discussion

For cases when the tilt function is correctly specified, the density ratio model holds approximately when the reference samples are fused with uniform samples. In these cases, the point estimates of the tail probability given by the OSF method are precise. More importantly, under correctly specified cases, the OSF confidence intervals reach the 95% nominal coverage for the tail probabilities. For misspecified cases, the confidence intervals produced by the OSF method no longer have 95% coverage. The ROSF method provides a way to relax the density ratio model assumption by repeatedly fusing the reference sample with artificial samples. The resulted point estimate can be obtained by averaging many OSF point estimates from repeated fusions. They are more accurate than OSF estimates and closer to the true tail probability as indicated by smaller MAE. ROSF method provides a way to balance the coverage and the length of confidence interval. By choosing appropriate N, the ROSF method provides in general a wider interval (relative to OSF) with significantly improved coverage. For the conservative choice $N = 100$, ROSF intervals reach the 95% coverage almost in all misspecified cases (except for the Pareto case), while the length of the confidence intervals are still much shorter than the AC and BM methods. The confidence intervals given by EP and POT methods are short but unreliable. The coverage for both EP and POT intervals is poor.

The AC method almost always gives the same point estimates and confidence interval. For $p = 0.01$, AC upper bounds are almost always 0.06; while for $p = 0.001$, the AC upper bounds are almost always 0.04. When $p = 0.01$, the EP method gives reasonable point estimates and confidence intervals approximately 60% of the time. However, when $p = 0.001$, almost none of the simulation samples contain "successes", due to the relatively small size of the sample ($n_0 = 100$). The EP method gives mostly 0 point estimates and upper bounds, as none of the data points in the reference sample exceeds the high threshold associated with the small tail probability. The performance of the two EVT methods (POT and BM) are not satisfactory either. The BM method gives confidence intervals with high coverage. However, the BM estimates can be highly variable leading to wide confidence intervals. The BM method gives the widest interval among all methods. The POT method gives good point estimates for the tail probabilities, however the POT confidence intervals are too short. In general, POT confidence intervals are about the same length as the

OSF intervals, and are slightly shorter than the ROSF intervals. However, the coverage of the POT intervals are around 60% to 70% in most cases. Such low coverage suggests that the POT estimates and intervals are unreliable and should be treated with caution.

EVT based modeling approaches characterize the behavior of distribution tails by reducing the sample and fitting GEV or GDP to the extreme observations. On the other hand, the DRM based approach *augments* the sample by combining real and artificial samples. DRM connects the distributions and produces estimated distribution functions and tail probabilities. The EVT data reduction and the DRM data augmentation idea are the major differences between EVT based and DRM based approaches. Given samples of limited size, EVT based approaches further reduce the sample size, thus the results are highly variable. In such scenarios, the DRM data fusion mechanism produces precise and reliable estimates and confidence intervals.

Chapter 4: Quantile Estimation Using DRM

4.1 Overview

Quantile estimation is an important task in many applications. For example, Value at Risk (VaR) which is an extreme upper quantile serves as a crucial risk indicator in the field of finance. In this Chapter, traditional quantile estimation based on empirical distribution function and the extreme value theorem (EVT)will be reviewed. New methods in quantile estimation based on the density ratio model will be introduced.

The estimation of quantiles and quantile functions is one of the most fundamental problems in probability and statistics. Let X be a random variable with distribution function F, and let q_p be the pth quantile for some $p \in (0, 1)$. Then

$$q_p = F^{-1}(p) = \inf\{x : F(x) \geq p\}. \tag{4.1}$$

For instance, $q_{0.5}$ would be the median and $q_{0.99}$ would be the 99th percentile, etc. Furthermore, we define the 0th quantle as $q_0 = \lim_{p\to 0} q_p$ and q_1 is defined as $q_1 = \lim_{p\to 1} q_p$. From the definition, it is clear that if F is strictly increasing in a neighborhood of q_p, then q_p is the inverse of the CDF F at p. If F happens to consist of flat sections (that is, an interval of points x satisfy $F(x) = p$), then q_p is the smallest x in the interval (see 4.1).

The sample quantile is a widely used estimator of q_p. Let $X_1 \ldots, X_n$ be a random sample of size n with common CDF F; then the sample quantile $\hat{q}_p, \in (0,1)$ is defined by:

$$\hat{q}_p = \hat{F}^{-1}(p)$$

where $\hat{F} = 1/n \sum_{i=1}^{n} I\{X_i \leq x\}$ is the empirical CDF.

As implied by the Glivenko-Cantelli Theorem, for a large sample of size n, $\hat{F} \approx F(x)$ for all x, therefore $\hat{q}_p \approx q_p$. Furthermore, $p = F(q_p)$, and we have the following:

$$p \approx F(\hat{q}_p) \approx F(q_p) + f(q_p)(\hat{q}_p - q_p) \approx \hat{F}(q_p) + f(q_p)(\hat{q}_p - q_p).$$

The second step follows from a Taylor approximation and the last step holds as $\hat{F} \approx F(x)$ for all x. Rearranging terms gives

$$\hat{q}_p \approx q_p - \frac{\hat{F}(q_p) - p}{f(q_p)} + R_n$$

where $R_n = O(n^{-3/4}(\log n)^{1/2}(\log\log n)^{1/4})$ almost surely (a.s.) as $n \to \infty$.

This result is known as the Bahadur representation of the sample quantile. Bahadur (1966) [2] had a rigorous development of the above argument.

The following theorem shows that the asymptotic distribution of the sample quantiles $\hat{q}_p$ for $p \in (0,1)$ are normally distributed.

Theorem 4.1 (Bahadur)**.** *Let $X_1, \ldots, X_n$ be i.i.d. drawn from a CDF F with continuous density f. If $f(q_p) > 0$ for $0 < p < 1$, then*

$$\sqrt{n}(\hat{q}_p - q_p) = \sqrt{n}\big(\hat{F}^{-1}(p) - F^{-1}(p)\big) \xrightarrow{d} N(0, \sigma^2),$$

where $\sigma^2 = p(1-p)/f^2(q_p)$.

To estimate the variance of a quantile, it is required to estimate the density f at the unknown point q_p. In practice, bootstrapping provides a simpler way for the estimation of the variance. It is also important to note that the two cases $p = 0$ and $p = 1$ were excluded in the above theorem as the asymptotic distribution of these extreme value statistics is very different and generally non-normal. In fact, the asymptotic behaviors of these extreme quantiles were described by EVT as introduced in Chapter 1. Let us briefly recap the methods for quantile estimation in EVT.

For the Block Maximum (BM) method, the data are blocked into a number of blocks of equal length from which a series of block maxima points are obtained. GEV distribution are fitted to the block maxima sample. Let z_p denote the p quantile of the block maxima data. Then z_p can be obtained by inverting the fitted GEV distribution function:

$$\hat{z}_p = \hat{G}^{-1}(p) = \begin{cases} \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}}[y_p^{-\hat{\xi}} - 1], & \hat{\xi} \neq 0 \\ \hat{\mu} + \hat{\sigma}\log y_p, & \hat{\xi} = 0 \end{cases}$$

where $y_p = -\log(p)$. Recall that z_p is also known as the $1/(1-p)$-blocks return level. Let q_p denote the p quantile of the original sample. Suppose now the estimated extreme quantile of the original data is desired, and further assume that the original sample of size n is divided into m blocks. Then q_p corresponds to the $m/((1-p)n)$ blocks return level or the n-observation return level. When using the BM method to estimate the extreme quantile for the original data, one should pay close attention to the fact that the p quantile of the original data is equivalent to the $[m-n(1-p)]/m$ quantile of the block maxima data.

For the Peaks Over Threshold (POT) method, a generalized Pareto distribution is fitted to the subsample whose values exceed a large threshold. The estimates of extreme quantiles can be then obtained by inverting the fitted GPD distribution function:

$$\hat{q}_p = \hat{H}^{-1}(p) = \begin{cases} u + \frac{\hat{\sigma}}{\hat{\xi}}\left[\left(\frac{k}{np}\right) - 1\right], & \hat{\xi} \neq 0 \\ u + \hat{\sigma}\log(\frac{k}{np}), & \hat{\xi} = 0 \end{cases}$$

.

4.2 ROSF in Extreme Quantile Estimation

In this section, we study the quantile estimator based on the density ratio model and ROSF. Let X_0 denote an i.i.d. sample from some given population

$$X_0 = (x_{01}, \ldots, x_{0n_0})' \sim g(x)$$

The distribution function $G(x)$ of X_0 is assumed to be unknown, and the p quantile $\hat{q} = inf\{x : \hat{G}(x) \geq p\}$ is the statistic of interest. X_0 is referred to as the *reference sample.* Let X_j denote a computer generated i.i.d. sample with sample size $n_j, j = 1, \ldots, m$

$$X_j = (x_{j1}, \ldots, x_{jn_j})' \sim g_j(x)$$

The computer generated samples X_j will be referred to as the *fusion samples.* Consider density ratio model 2.4, and let $\hat{G}$ denote the estimated distribution function for the reference sample. For any $p \in (0, 1)$, define the pth quantile of G as

$q_p = inf\{x : G(x) \geq p\}$, then the density ratio model based estimator is:

$$\hat{q}_p^{\mathrm{DRM}} = inf\{x : \hat{G}(x) \geq p\}.$$

The $100(1-\alpha)\%$ confidence interval for the p quantile $\hat{q}_p$ can be constructed based on the asymptotic results given in Appendix A:

$$\big(\hat{q}_p - z_{\alpha/2}\sqrt{\hat{\mathrm{var}}(\hat{q}_p)}, \hat{q}_p + z_{\alpha/2}\sqrt{\hat{\mathrm{var}}(\hat{q}_p)}\big).$$

A consistent and effective estimator of the estimated variance of $\hat{q}_p$, $\hat{\mathrm{var}}(\hat{q}_p)$, is needed. Based on asymptotic results, plug-in consistent variance estimator can be obtained. Two necessary ingredients are consistent estimation of $v_{ij}(x,y)$ and $g_i(x)$. Note that v_{ij} is the i,jth component in the covariance matrix of the process $\sqrt{n}(\hat{G}(t) - G(t))$ given in Theorem 2.2. Thus, v_{ij} can be obtained through DRM. To obtain $\hat{g}_i$ can be a bit more involved due to the fact that $\hat{G}$ is discrete. However, the idea of kernel density estimation can be used to produce a density estimate. Let $K(.) \geq 0$ be a commonly used kernel function (e.g. standard normal density function) such that $\int K(x)dx = 1$, $\int xK(x)dx = 0$ and $\int x^2K(x)dx < \infty$. For some bandwidth b¿0, let $K_b(x) = (1/b)K(x/b)$. Then a kernel estimate of g can be obtained by

$$\hat{g}(x) = \int K_b(x-y)d\hat{G}(y).$$

For detailed introduction about kernel density estimation on DRM, see Voulgaraki et al. (2012) [51]. The variance of $\hat{q}_p$ becomes large as $p \to 1$ or $p \to 0$. Therefore, we propose the ROSF algorithm to construct short confidence interval for extreme quantiles.

We shall now describe the implementation of ROSF in the estimation of extreme quantiles and the associated confidence interval. We are interested in estimating quantile q_p for a large p based a random sample X_0 from some distribution. X_0 is referred to as the reference sample. A fusion sample X_1 is then generated by the computer and fused together with the reference sample. The point estimate $\hat{q}_1$ is then obtained through the OSF as described above. The same reference sample is then fused with another computer generated sample (from the same distribution of the previous artificial sample and independent of it) to obtain another $\hat{q}_2$ in the same manner as before. This process is repeated n_r (n_r stands for number of repetition) times to produce a sequence of OSF point estimators of the quantile $\hat{q}_i, i = 1, \ldots, n_r$.

The final point estimate of the p quantile $\hat{q}_p$ from ROSF algorithm is the average of $\hat{q}_i$'s from n_r OSF runs:

$$\hat{q} = inf\{x : \hat{G}(x) \geq p\} = \frac{1}{n_r}\sum_{i=1}^{n_r} \hat{q}_i, \quad i = 1, \ldots, n_r$$

and the associated $100(1-\alpha)\%$ confidence interval is

$$\left[F_Q^{-1}\left(1-\alpha^{1/N}\right), F_Q^{-1}\left(\alpha^{1/N}\right)\right].$$

where N is a large enough positive integer.

Conditional on X_0, the sequence of the OSF quantile estimators $\hat{q}_i$ are independent and identically distributed from some distribution F_Q. Denote the empirical distribution of Q_i's by $\hat{F}_Q$. By Glivenko-Cantelli theorem, $\tilde{F}_Q$ converges to F_Q almost surely uniformly as n_r increases. Since the process may be repeated many times, a very close approximation of F_Q can be obtained. In other words, as the number of fusions becomes very large, $\hat{F}_Q$ is almost the exact F_Q.

The ROSF point estimate of the quantile is the average of B OSF point estimates. The ROSF confidence interval for the estimated quantile $\hat{q}_p$ depends on the estimated cumulative distribution function $\hat{F}_Q$ and a tuning parameter N.

The construction of the ROSF confidence interval follows from similar reasoning as presented in Chapter 3. For an i.i.d. sequence of the estimated OSF quantiles $\hat{q}_1, \ldots, \hat{q}_N$, denote the distribution of the sequence by Q. Furthermore, let us denote the maximum of the sequence by $q_{(N)} = \max(\hat{q}_i)$ and the true quantile by q. It follows that

$$\mathrm{P}(q_{(N)} > q) = 1 - F_Q^N(q)$$

If $\mathrm{P}(Q > q) = 1 - F_Q(q) > 0$, then from the above equation, the probability that the maximum of OSF quantile estimates covers the true quantile increases as the tuning parameter N increases. Conditional on the given sample X_0, for all $N > N_0$, we have the following inequality:

$$1 - F_Q^N(q) \geq 0.95$$

for some N_0 sufficiently large. The inequality can be rewritten by inverting the distributon function:

$$q \leq F_Q^{-1}(0.05^{1/N})$$

The above relationship implies that the constructed ROSF upper bound (for the estimated quantile) $F_Q^{-1}\left(\alpha^{1/N}\right)$ covers the true quantile q with at least 95% confidence for sufficiently large N. The ROSF lower bound for the estimated quantile can be constructed by similar arguments. As N becomes larger, the length of the confidence interval also becomes larger. In the next section, it will be shown that the ROSF confidence intervals reach the desired coverage for sufficiently large N in simulation studies.

4.3 Simulation Studies: $h(x)$ Correctly Specified

In this section, BM, POT, ROSF, and empirical methods of estimation of extreme quantiles and associated confidence intervals will be applied and compared in simulation studies.

Following Chapter 3, the reference samples in simulation studies are randomly generated from a range of skewed distributions including: Exponential, Gamma, Log-normal, Weibull, Pareto, F and Truncated Gumbel. The reference sample is then fused repeatedly with computer generated uniform random samples. In all cases, a sufficiently large upper limit of the uniform distribution needs to be determined. Recall that, when the reference sample comes from a skewed distribution, the gamma tilt $h(x) = (x, \log x)$ and the log-normal tilt $(\log x, \log^2(x))$ make the density ratio model hold approximately. The analysis is conducted in two parts: in the first part, we consider scenarios when the tilt function is correctly specified and in the second part when the tilt function is misspecified. For each case, the 0.99 and 0.999 extreme tail quantiles are estimated. Mean absolute error (MAE) which calculates the mean absolute value of the differences between the estimated quantiles $\hat{q}$ and the true quantile q is the metric that is used to measure the precision of the point estimates. More importantly, the 95% confidence intervals for the estimated quantiles are calculated.

The BM and POT confidence intervals are obtained from the delta method and bootstrapping. Delta method intervals are calculated from explicit expressions given in 1.3 and 1.4. Bootstrapped intervals are constructed from the 2.5th to 97.5th percentiles based on 500 replications. The ROSF confidence intervals are obtained by setting the tuning parameter N to 3000 for all cases.

To obtain reliable coverage results, in each study, five hundred confidence intervals are computed for each method. The performance of the methods is evaluated based on the coverage, mean length of the confidence intervals, and the mean absolute error.

The tilt function is correctly specified, if it corresponds to the distribution of the reference sample correctly. For example, if the reference sample comes from the gamma family, the gamma tilt $h(x) = (x, \log x)$ is appropriate; if the reference sample comes from the log-normal family then the log-normal tilt would be appropriate. In this section, we will see that under a correctly specified tilt function where the density ratio model holds exactly, the ROSF methods give very precise point estimates for the extreme quantiles and short yet reliable confidence intervals as well.

4.3.1 $X_0 \sim \text{Exp}(1.2)$

In this example, $X_0 \sim \text{Exp}(1.2)$. The true theoretical 0.99 and 0.999 quantiles are $q_{0.99} = 3.837642$ and $q_{0.999} = 5.756463$ respectively. For the 0.99 quantile, fusion samples X_1's are generated from Uniform(0,20), and F_Q is based based on $n_r = 10000$ repetitions; for the 0.999 quantile, fusion samples X_1's are generated from Uniform(0,25) and F_Q is obtained based on $n_r = 2000$ repetitions. In this case, the gamma tilt function $h(x) = (x, \log x)$ is appropriate and hence adopted. See Table 4.1 for the performance of various methods.

Table 4.1: ROSF Interval Coverage and Length for 0.99 and 0.999 quantiles, $X_0 \sim \text{Exp}(1.2)$, $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

		ROSF	BM_{Delta}	BM_{Boot}	$\text{POT}_{\text{Delta}}$	POT_{Boot}	EP
$q_{0.99}$	Coverage	95.2%	84.2%	85.0%	93.0%	74.4%	65.8%
	CI Length	2.17	3.28	2.89	3.35	2.31	2.26
	MAE	0.44	0.59	0.53	0.54	0.53	0.56
$q_{0.999}$	Coverage	99.2%	85.4%	94.0%	93.2%	79.2%	9.6%
	CI Length	5.56	16.21	18.84	15.25	15.71	2.05
	MAE	0.78	2.43	1.60	1.66	1.57	1.87

4.3.2 $X_0 \sim \text{Gamma}(5, 3)$

In this example, $X_0 \sim \text{Gamma}(5, 3)$. The true theoretical 0.99 and 0.999 quantiles are $q_{0.99} = 3.868209$ and $q_{0.999} = 4.931383$ respectively. For the 0.99 quantile, fusion samples X_1's are generated from Uniform(0,20), and F_Q is obtained based on $n_r = 3500$ repetitions; for the 0.999 quantile, fusion samples X_1's are generated from Uniform(0,25) and F_Q is obtained based on $n_r = 2500$ repetitions.

In this case, the gamma tilt function $h(x) = (x, \log x)$ is appropriate and hence adopted. See Table 4.2 for the performance of various methods.

Table 4.2: ROSF Interval Coverage and Length for 0.99 and 0.999 quantiles, $X_0 \sim \text{Gamma}(5, 3)$, $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

		ROSF	BM_{Delta}	BM_{Boot}	$\text{POT}_{\text{Delta}}$	POT_{Boot}	EP
$q_{0.99}$	Coverage	97.8%	81.4%	79.6%	97.0%	69.6%	66.2%
	CI Length	1.29	1.70	1.49	2.47	1.30	1.42
	MAE	0.26	0.33	0.31	0.32	0.31	0.35
$q_{0.999}$	Coverage	99.6%	76.6%	86.6%	97.8%	74.0%	11.2%
	CI Length	3.00	6.63	6.96	9.25	5.90	1.27
	MAE	0.40	1.07	0.80	0.91	0.87	1.04

4.3.3 $X_0 \sim \text{Gamma}(1, 0.01)$

In this example, $X_0 \sim \text{Gamma}(1, 0.01)$. The true theoretical 0.99 and 0.999 quantiles are $q_{0.99} = 460.517$ and $q_{0.999} = 690.7755$ respectively. For the 0.99 quantile, fusion samples X_1's are generated from Uniform(0,600), and F_Q is obtained based on $n_r = 10000$ repetitions; for the 0.999 quantile, fusion samples X_1's are generated from Uniform(0,800) and F_Q is obtained based on $n_r = 3500$ repetitions. In this case, the gamma tilt function $h(x) = (x, \log x)$ is appropriate and hence adopted. See Table 4.3 for the performance of various methods.

4.3.4 $X_0 \sim \text{Log-Normal}(1, 1)$

In this example, $X_0 \sim \text{Log-Normal}(1, 1)$. The true theoretical 0.99 and 0.999 quantiles are $q_{0.99} = 27.83649$ and $q_{0.999} = 59.75377$ respectively. For the 0.99 quantile, fusion samples X_1's are generated from Uniform(0,95), and F_Q is obtained

Table 4.3: ROSF Interval Coverage and Length for 0.99 and 0.999 quantiles, $X_0 \sim$ Gamma$(1, 0.01)$, $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

		ROSF	BM_{Delta}	BM_{Boot}	$\text{POT}_{\text{Delta}}$	POT_{Boot}	EP
$q_{0.99}$	Coverage	93.2%	85.8%	85.6%	91.4%	74.8%	63.4%
	CI Length	219.99	404.39	363.72	380.84	287.47	272.38
	MAE	45.72	71.06	64.27	64.58	64.14	67.04
$q_{0.999}$	Coverage	97.6%	85.2%	94.4%	89.2%	79.6%	9.0%
	CI Length	321.97	1001.68	1592.45	930.69	799.63	246.21
	MAE	61.55	181.49	151.98	155.30	161.27	223.40

based on $n_r = 4000$ repetitions; for the 0.999 quantile, fusion samples X_1's are generated from Uniform(0,120) and F_Q is obtained based on $n_r = 3500$ repetitions. In this case, the log-normal tilt function $h(x) = (\log x, (\log x)^2)$ is appropriate and hence adopted. See Table 4.4 for the performance of various methods.

Table 4.4: ROSF Interval Coverage and Length for 0.99 and 0.999 quantiles, $X_0 \sim$ lnorm$(1, 1)$, $n_0 = n_1 = 100$, $h(x) = (\log x, (\log x)^2)$

		ROSF	BM_{Delta}	BM_{Boot}	$\text{POT}_{\text{Delta}}$	POT_{Boot}	EP
$q_{0.99}$	Coverage	95.4%	87.8%	90.6%	83.2%	77.6%	63.6%
	CI Length	23.88	49.14	37.61	37.06	37.41	26.05
	MAE	5.84	8.00	6.77	6.88	6.59	6.96
$q_{0.999}$	Coverage	96.4%	86.8%	96.0%	76.6%	81.2%	8.8%
	CI Length	66.45	206.50	383.00	127.17	148.05	24.33
	MAE	12.35	26.71	23.19	24.92	24.98	30.40

4.3.5 $X_0 \sim$ Log-Normal$(0, 1)$

In this example, $X_0 \sim$ Log-Normal$(0, 1)$. The true theoretical 0.99 and 0.999 quantiles are $q_{0.99} = 10.24047$ and $q_{0.999} = 21.98218$ respectively. For the 0.99 quantile, fusion samples X_1's are generated from Uniform(0,50), and F_Q is obtained based on $n_r = 3500$ repetitions; for the 0.999 quantile, fusion samples X_1's are generated from Uniform(0,65) and F_Q is obtained based on $n_r = 2000$ repetitions. In this case, the log-normal tilt function $h(x) = (\log x, (\log x)^2)$ is appropriate and hence adopted. See Table 4.5 for the performance of various methods.

Table 4.5: ROSF Interval Coverage and Length for 0.99 and 0.999 quantiles, $X_0 \sim$ lnorm$(0, 1)$, $n_0 = n_1 = 100$, $h(x) = (\log x, (\log x)^2)$

		ROSF	BM_{Delta}	BM_{Boot}	$\text{POT}_{\text{Delta}}$	POT_{Boot}	EP
	Coverage	95.6%	86.0%	89.0%	81.8%	79.2%	62.6%
$q_{0.99}$	CI Length	9.95	18.89	19.11	14.34	14.23	9.99
	MAE	2.29	3.19	2.74	2.83	2.67	2.74
	Coverage	97.4%	87.0%	94.8%	78.2%	82.6%	12.4%
$q_{0.999}$	CI Length	33.29	82.20	147.83	49.33	59.41	9.34
	MAE	6.67	10.87	9.47	9.33	9.49	10.98

4.4 Simulation Studies: $h(x)$ Misspecified

The key assumption of the density ratio model is that the log ratio of two unknown probability density functions takes some known linear form which depends on finite dimensional parameters. In the previous section, it was shown that when the tilt function $h(x)$ is correctly specified, the estimated quantile by the ROSF method is precise. The associated confidence intervals are short and reliable. In this section, the performance of the ROSF method under misspecified scenarios will be presented in simulation examples. Random samples generated from various types of skewed distributions will be treated as the reference sample. Skewed distributions we considered in the simulation studies include: Pareto, F, Inverse Gaussian, Weibull, and Truncated Gumbel. Instead of using the correct tilt function for each case, either the gamma tilt $h(x) = (x, \log x)$ or the log-normal tilt $h(x) = (\log x, (\log x)^2)$ will be used.

4.4.1 $X_0 \sim \text{Pareto}(1, 4)$

In this example, $X_0 \sim \text{Pareto}(1, 4)$. The true theoretical 0.99 and 0.999 quantiles are $q_{0.99} = 3.162278$ and $q_{0.999} = 5.623413$ respectively. For the 0.99 quantile, fusion samples X_1's are generated from Uniform(0,160), and F_Q is obtained based on $n_r = 3500$ repetitions; for 0.999 quantile, fusion samples X_1's are generated from Uniform(0,200) and F_Q is obtained based on $n_r = 2000$ repetitions. In this case, the gamma tilt function $h(x) = (x, \log x)$ is adopted. See Table 4.6 for the performance of various methods.

Table 4.6: ROSF Interval Coverage and Length for 0.99 and 0.999 quantiles, $X_0 \sim \text{Pareto}(1, 4)$, $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

		ROSF	BM_{Delta}	BM_{Boot}	$\text{POT}_{\text{Delta}}$	POT_{Boot}	EP
$q_{0.99}$	Coverage	95.8%	85.0%	85.4%	80.2%	76.4%	64.8%
	CI Length	1.93	3.51	3.43	2.67	2.63	1.97
	MAE	0.48	0.61	0.52	0.53	0.51	0.53
$q_{0.999}$	Coverage	99.6%	84.2%	94.0%	72.6%	76.8%	7.8%
	CI Length	6.00	32.30	54.56	21.56	30.98	1.84
	MAE	1.93	4.50	2.60	3.17	2.59	2.37

4.4.2 $X_0 \sim \text{F}(2, 12)$

In this example, $X_0 \sim \text{F}(2, 12)$. The true theoretical 0.99 and 0.999 quantiles are $q_{0.99} = 6.926608$ and $q_{0.999} = 12.97367$ respectively. For the 0.99 quantile, fusion samples X_1's are generated from Uniform(0,8), and F_Q is obtained based on $n_r = 500$ repetitions; for the 0.999 quantile, fusion samples X_1's are generated from Uniform(0,15) and F_Q is obtained based on $n_r = 2000$ repetitions. In this case, the gamma tilt function $h(x) = (x, \log x)$ is adopted. See Table 4.7 for the performance of various methods.

Table 4.7: ROSF Interval Coverage and Length for 0.99 and 0.999 quantiles, $X_0 \sim \text{F}(2, 12)$, $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

		ROSF	BM_{Delta}	BM_{Boot}	$\text{POT}_{\text{Delta}}$	POT_{Boot}	EP
$q_{0.99}$	Coverage	97.2%	82.0%	85.6%	84.8%	77.8%	65.8%
	CI Length	1.73	9.19	8.60	7.30	6.57	5.42
	MAE	0.34	1.67	1.45	1.44	1.39	1.43
$q_{0.999}$	Coverage	98.8%	84.2%	92.4%	80.0%	79.4%	8.2%
	CI Length	4.20	31.13	51.15	22.38	23.72	5.03
	MAE	0.72	5.04	4.44	4.37	4.28	5.75

4.4.3 $X_0 \sim \text{Inv-Gauss}(4, 5)$

In this example, $X_0 \sim \text{IG}(4, 5)$. The true theoretical 0.99 and 0.999 quantiles are $q_{0.99} = 17.87176$ and $q_{0.999} = 28.95409$ respectively. For the 0.99 quantile, fusion samples X_1's are generated from Uniform(0,60), and F_Q is obtained based on $n_r = 10000$ repetitions; for 0.999 quantile, fusion samples X_1's are generated from Uniform(0,60) and F_Q is obtained based on $n_r = 2000$ repetitions. In this case, the log-normal tilt function $h(x) = (\log x, (\log x)^2)$ is adopted. See Table 4.8 for the performance of various methods.

Table 4.8: ROSF Interval Coverage and Length for 0.99 and 0.999 quantiles, $X_0 \sim$ invgauss$(4, 5)$, $n_0 = n_1 = 100$, $h(x) = (\log x, (\log x)^2)$

		ROSF	BM_{Delta}	BM_{Boot}	$\text{POT}_{\text{Delta}}$	POT_{Boot}	EP
$q_{0.99}$	Coverage	97.6%	86.2%	90.4%	88.8%	78.2%	64.2%
	CI Length	13.93	16.31	16.97	13.91	11.59	11.79
	MAE	2.41	2.48	2.42	2.38	2.41	2.95
$q_{0.999}$	Coverage	96.6%	86.2%	98.4%	86.4%	82.4%	9.2%
	CI Length	28.77	64.81	126.39	45.81	46.83	10.77
	MAE	4.92	9.34	8.26	7.86	7.41	10.67

4.4.4 $X_0 \sim$ Inv-Gauss$(2, 40)$

In this example, $X_0 \sim \text{IG}(2, 40)$. The true theoretical 0.99 and 0.999 quantiles are $q_{0.99} = 3.257718$ and $q_{0.999} = 3.835791$ respectively. For the 0.99 quantile, fusion samples X_1's are generated from Uniform(0,20), and F_Q is obtained based on $n_r = 5000$ repetitions; for the 0.999 quantile, fusion samples X_1's are generated from Uniform(0,20) and F_Q is obtained based on $n_r = 1000$ repetitions. In this case, the log-normal tilt function $h(x) = (\log x, (\log x)^2)$ is adopted. See Table 4.9 for the performance of various methods.

Table 4.9: ROSF Interval Coverage and Length for 0.99 and 0.999 quantiles, $X_0 \sim$ invgauss$(2, 40)$, $n_0 = n_1 = 100$, $h(x) = (\log x, (\log x)^2)$

		ROSF	BM_{Delta}	BM_{Boot}	$\text{POT}_{\text{Delta}}$	POT_{Boot}	EP
$q_{0.99}$	Coverage	96.8%	78.6%	74.0%	96.6%	65.4%	59.8%
	CI Length	0.77	0.84	0.75	1.36	0.63	0.75
	MAE	0.16	0.18	0.18	0.18	0.18	0.20
$q_{0.999}$	Coverage	99.6%	75.8%	85.2%	97.0%	67.2%	7.4%
	CI Length	1.44	2.96	3.21	4.70	2.17	0.66
	MAE	0.23	0.56	0.45	0.47	0.47	0.59

4.4.5 $X_0 \sim \text{Weibull}(1, 2)$

In this example, $X_0 \sim \text{Weibull}(1, 2)$. The true theoretical 0.99 and 0.999 quantiles are $q_{0.99} = 9.21034$ and $q_{0.999} = 13.81551$ respectively. For the 0.99 quantile, fusion samples X_1's are generated from Uniform(0,30), and F_Q is obtained based on $n_r = 6000$ repetitions; for the 0.999 quantile, fusion samples X_1's are generated from Uniform(0,40) and F_Q is obtained based on $n_r = 500$ repetitions. In this case, the gamma tilt function $h(x) = (x, \log x)$ is adopted. See Table 4.10 for the performance of various methods.

Table 4.10: ROSF Interval Coverage and Length for 0.99 and 0.999 quantiles, $X_0 \sim$ weibull$(1, 2)$, $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

		ROSF	BM_{Delta}	BM_{Boot}	$\text{POT}_{\text{Delta}}$	POT_{Boot}	EP
$q_{0.99}$	Coverage	95.0%	84.6%	90.2%	93.4%	74.6%	64.8%
	CI Length	5.03	8.38	8.87	8.00	6.63	5.44
	MAE	0.99	1.49	1.28	1.32	1.29	1.37
$q_{0.999}$	Coverage	97.4%	83.4%	96.6%	91.0%	79.0%	9.0%
	CI Length	11.12	21.73	35.87	17.72	14.84	4.92
	MAE	1.73	3.67	3.92	4.48	4.12	4.51

4.4.6 $X_0 \sim \text{Truncated Gumbel}(0, 5)$

In this example, $X_0 \sim$ Trunc. Gumbel$(0, 5)$. The true 0.99 and 0.999 quantiles are $q_{0.99} = 25.36103$ and $q_{0.999} = 37.2$ respectively. For the 0.99 quantile, fusion samples X_1's are generated from Uniform(0,40), and F_Q is obtained based on $n_r = 5000$ repetitions; for the 0.999 quantile, fusion samples X_1's are generated from Uniform(0,50) and F_Q is obtained based on $n_r = 1000$ repetitions. In this case, the gamma tilt function $h(x) = (x, \log x)$ is adopted. See Table 4.11 for the performance of various methods.

Table 4.11: ROSF Interval Coverage and Length for 0.99 and 0.999 quantiles, $X_0 \sim$ Trunc. Gumbel$(0, 5)$, $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

		ROSF	BM_{Delta}	BM_{Boot}	$\text{POT}_{\text{Delta}}$	POT_{Boot}	EP
$q_{0.99}$	Coverage	96.2%	82.0%	90.4%	93.8%	72.6%	63.0%
	CI Length	10.43	19.45	26.54	21.88	14.39	13.90
	MAE	2.12	3.68	3.41	3.48	3.41	3.64
$q_{0.999}$	Coverage	96.4%	80.8%	97.4%	92.0%	78.0%	8.6%
	CI Length	15.66	52.63	99.15	47.82	36.57	12.54
	MAE	3.08	9.47	8.35	8.95	8.76	11.78

4.5 Discussion

For 0.99 quantiles, the EP method gives reasonable point estimates and confidence intervals approximately 60% of the time. However, when 0.999 quantiles are under consideration, the EP confidence intervals result in very poor coverage (from roughly 7% to 12%). The coverage of the two EVT methods (POT and BM) are satisfactory in some cases. The BM method gives confidence intervals with high coverage. However, the BM estimates can be highly variable leading to wide confidence intervals which are not informative for risk assessment purposes. The BM method gives the widest interval among all methods. The POT method gives shorter confidence intervals (relative to BM intervals). Also, the coverage of POT confidence intervals is not stable, ranging from 65% to 98%.

The ROSF method demonstrates its advantage in estimation of extreme quantiles across all simulation studies. The ROSF method repeatedly fuses the reference sample with artificial samples. The resulting point estimate is obtained by averaging many OSF point estimates from repeated fusions. In all cases, the ROSF point estimates for quantiles are more accurate than estimates obtained from EVT based methods. The ROSF point estimates are closer to the true tail probability as indicated by smaller MAE. To obtain the ROSF confidence interval for the estimator of quantile, a tuning parameter N needs to be determined. How to choose the optimal N is still an open problem. From many simulation results, the choice $N = 3000$ suffices in many specified and misspecified cases for extreme quantiles. This means that the ROSF confidence intervals give desired coverage when $N = 3000$. Therefore choice $N = 3000$ is prudent and used as default. In some difficult cases, a larger N is needed so that the confidence intervals give the desired coverage. In our simulation studies, the ROSF confidence intervals for extreme quantiles reach the 95% nominal coverage for almost all cases regardless of whether the tilt function $h(x)$ is correctly specified or not. Furthermore, the length of the ROSF confidence intervals is much shorter than intervals obtained from EVT based methods. The advantage of the ROSF method is quite significant when the 0.999 quantile is under consideration. The difference between the ROSF method and EVT based methods are more noticeable, if the underlying sample is from a distribution with a long tail (e.g. Gamma(1,0.01)). In conclusion, the estimation of extreme quantiles based on ROSF is precise and reliable when the size of the underlying sample is moderately large.

Chapter 5: Real Data Applications

In this chapter, the EVT based methods and DRM based methods in the estimation of threshold exceedance probabilities and extreme quantiles are applied in two real data problems. The performance of various methods will be compared.

5.1 Application in Food Safety

Certain foods may contain varying amounts of toxins, chemicals, and/or heavy metals. Exposures to high levels of these contaminants in food may cause severe health problems. Since food is predominantly safe, the outbreak of food-borne illness is rare. However, rare events like food poisoning could be lethal and costly. According to the Centers for Disease Control (CDC) [7], "About one in six Americans gets sick every year from food-borne diseases and $3,000$ die. Salmonella, a bacteria that commonly causes food-borne diseases incurs \$365 million in direct medical costs annually". To establish regulatory standards to prevent the contamination of food and to provide nutritional recommendations, risk assessment measures are required. Risk assessment should be based not only on the detection of the contaminants (qualitative risk assessment), but also on the quantitative evaluation of contaminants in food products (quantitative risk assessment). Joint work by Food and Agriculture Organization (FAO) and World Health Organization (WHO) recommended a stepwise procedure to quantitatively assess consumer exposure to food contaminants [15]. A commonly used measure of risk related to the presence of contaminants in food is the probability that the contaminant intake/exposure exceeds a safe level determined by a FAO/WHO joint expert committee on food additives (JECFA) based on experimental and/or epidemiological studies. This exceedance probability is referred to as the *risk index*, and the safe level of intake dosage is called *provisional tolerable weekly intake* (PTWI) which is defined in micrograms per week per kg of body weight ($\mu g/kgbw$). When both consumption data and contamination data are available, exposure can be defined as the cross product of contamination and consumption for given food items and contaminants. For detailed guidelines and information, the reader may refer to FAO/WHO (1999) [16] and FAO/WHO (2000) [17].

Since the quantity of interest is the probability that the individual intake of contaminants via food exceeds a certain threshold, our goal is to produce precise estimation of this threshold exceedance probability (or risk index) and construct short and reliable confidence intervals. To be more specific, the tail probability P($X > PTWI$) (where X is a random variable that represents individual intake of contaminants through consumption of a certain food) and its associated confidence intervals are desired. Various methods for risk index calculation can be found in existing literature. Tressou et al. (2004) [47] suggested the use of extreme value theory (EVT) to evaluate the risk. They argued that PTWI belongs to the exposure tail distribution, and modeled the exposure tail by a Pareto type distribution characterized by a Pareto index which may be seen as a measure of the risk of exceeding the PTWI. Gauchi and Leblanc (2002) [22] proposed a more empirical approach based on Monte Carlo estimation and a parametric type method of simulation. In their study, confidence intervals of the risk index are constructed by the bootstrap method.

In this section, we focus our attention on heavy metal (lead and mercury) exposures related to the consumption of *fish and seafood products*. How OSF and ROSF methods are used in the estimation of risk index will be presented.

5.1.1 Food Consumption Data

Consumption data come from the National Health and Nutrition Examination Survey (NHANES) which is a program of studies designed to assess the health and nutritional status of adults and children in the United States. A dietary interview component, called *What We Eat in America* (WWEIA 2005-2006), is contained in the survey. The interviews are conducted as joint work between the US Department of Agriculture (USDA) and the US Department of Health and Human Services (DHHS). Under this partnership, DHHS's National Center for Health Statistics (NCHS) is responsible for the sample design and data collection. USDA's Food Surveys Research Group (FSRG) is responsible for the dietary data collection methodology, maintainance of the databases used to code and process data, and data review and processing.

Detailed dietary intake information from NHANES participants are obtained during dietary recall interviews. The dietary intake data are used to estimate the types and amounts of foods and beverages consumed during a 24-hour period prior to the interview (midnight to midnight 24 hour recall). The intakes of energy, nutrients and other food components based on consumed food and beverages are also estimated.

NHANES provides a detailed food list contains 7178 food items clustered in groups including: dairy products, poultry, fish, meat product, and beverages etc. Among this food list, 179 food items belongs to the group "Fish". Among all participants of the food survey, $n = 3021$ individuals who consumed one or more fish items are considered as fish and seafood product consumers. For each seafood product consumer, the total consumption is properly calculated and recorded. Consumer body weights are also available from the survey. The body weight information is crucial to the study since PTWI is expressed as a contaminant unit per kilogram of body weight ($\mu g/kgbw$). The calculation of the normalized consumption (consumption divided by individual weight) relies on the availability of the body weight data.

5.1.2 Contamination Data

Seafood product contamination data are collected by National Oceanic and Atmospheric Administration (NOAA) through their National Status and Trends (NST) program. This is a monitoring program that collects and records data of heavy metal and pesticides residue concentrations present in samples of seafood products. For each of the contaminants of interest (Pb and Hg), there are 327 values expressed in μg per gram of fresh weight of sea product.

Methylmercury (MeHg), the toxic form of mercury, are almost exclusively present in sea products. The amount of Methylmercury in seafood products can be derived from mercury contents. Claisse et al (2001) [9] suggested that methylmercury concentration in seafood can be obtained by simply applying the conversion factor (0.84 for seafood) to the mercury concentration data. This method is adopted in our study.

It is very common that concentration data are subject to left censoring due to detection or quantification limits of analytical methods. By convention, the censored data can be replaced by either the limit of detection (LOD), or by half of LOD, or by zero based on the proportion of left censored data. In our study the proportion of censored lead and mercury concentration data are low, 8.87% and 2.75%, respectively. A conservative method is adopted in our application; that is, censored data are replaced by the limit of detection (LOD).

5.1.3 Exposure Calculation

To our knowledge, heavy metal intakes have not been measured directly. Hence, human heavy metal intake from food consumption is derived from consumption and contamination data. A dataset of 3000 daily heavy metal intakes (in $\mu g/kg$ BW) was constructed bu multiplying the daily seafood consumption data for each individual (in g/kg BW) by heavy metal concentration values (in $\mu g/g$) randomly sampled from the contamination data for seafood.

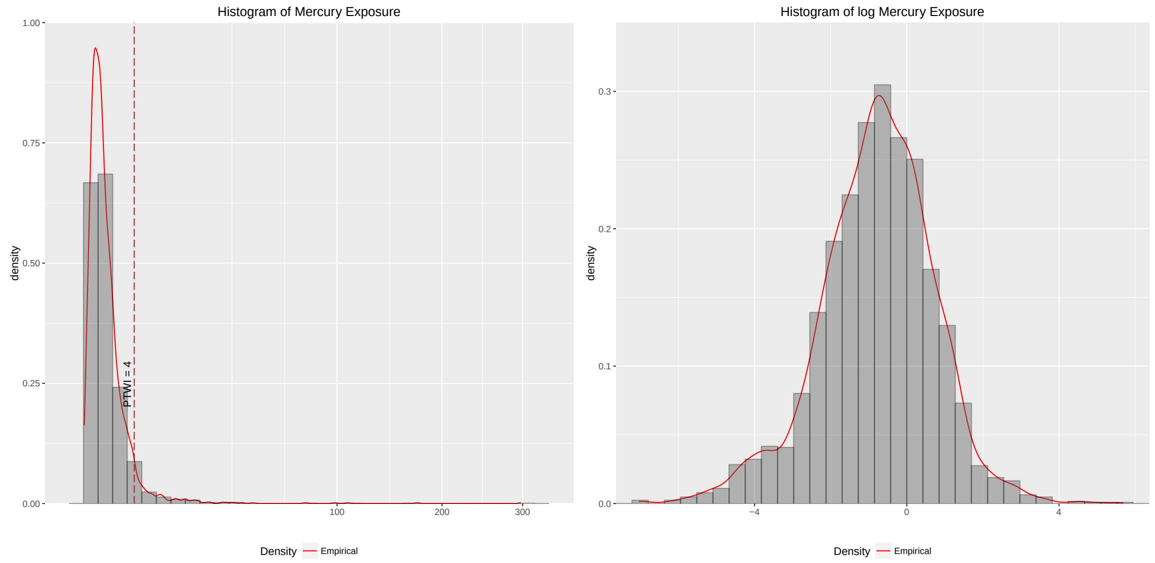

Figure 5.1: Histogram of 3000 Mercury Exposure Measurements

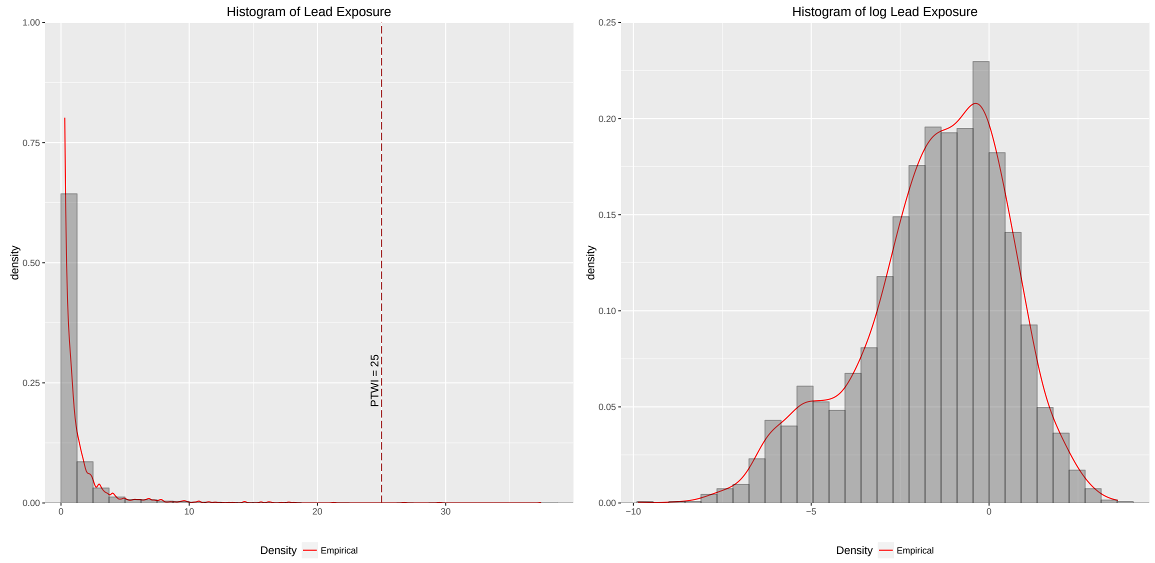

Figure 5.2: Histogram of 3000 Lead Exposure Measurements

The construction of the contamination intake dataset relies on random draws of heavy metal concentration from seafood contamination data. Many contamination intake dataset could be generated and each would be slightly different from the others due to randomness in sampling. The international toxicological references (PTWI) were estabilshed and revised by JECFA. The most recent tolerable intake for heavy metals considered in this study can be found on the WHO website: $25\mu g/kg/week$ for lead, and $4\mu g/kg/week$ for mercury. For each simulation, the contamination intake data are generated and the risk index $\mathrm{P}(X > PTWI)$ is calculated. From many simulation, the risk index is approximately 0.001 for lead and 0.05 for mercury. For illustrative purposes, we record one lead exposure dataset when the risk index is exactly 0.001 and the mercury exposure dataset when the risk index is exactly 0.05. These risk indices are considered as the true risk indices. OSF and ROSF method together with other methods described previously will be applied to samples of size 100 to estimate the risk indices. The performance of each method will be given in the following section. The histograms of the recorded lead and mercury intake data are given in Figure 5.1 and Figure 5.2.

5.1.4 Tail Probability Estimation

To apply the OSF and/or the ROSF methods, the first thing is to specify the tilt function. With a correctly specified tilt function, the density ratio model gives significantly better results (more precise point estimates, better coverage, and shorter confidence intervals). The appropriateness of the tilt function depends on the distribution of the reference sample. In simulation studies, the distribution of the reference sample is known, so whether the tilt function is correct or misspecified is known. In real data applications, however, the distribution of the reference sample is generally unknown. To check if the data follow a certain distribution is essentially a model selection problem. From the histogram of the data, the distribution of heavy metal exposures is skewed with a long tail. From simulation results, skewed distributions can be accommodated by either the gamma or log-normal tilt quite well. Instead of finding the right model among the infinite dimensional set of distributions, we focus our attention on checking whether the data come from the gamma or log-normal distributions. In this application, AIC is used to identify which model is more appropriate for the reference data.

The identification process first starts with fitting both the gamma and log-normal distributions to the heavy metal intake. The AIC values of both models are then obtained. The model with the lower AIC value is favored. For the mercury intake, the log-normal model yields a lower AIC value. Thus the log-normal tilt is more appropriate for the mercury intake.

For the merucry exposure, the threshold $PTWI = 4$ and the true tail probability is $p = 1 - G(T) = 0.05$. Based on AIC, the log-normal tilt function is a more appropriate choice. The performance of the OSF with log-normal tilt is better than the performance of the OSF with gamma tilt. However, the desired coverage is not reached by all OSF methods. Hence, we will use ROSF to fix this situation. The POT method works well when the tail probability is not too small. The POT confidence interval has about 90% coverage, and the mean length of the CI is slightly larger than OSF CI. See Table 5.1 for detailed comparison of the

Table 5.1: OSF Interval Coverage and Length for $p = 1 - G(T) = 0.05$, $T = 4$, X_0 is sampled from Mercury Intake, in all cases $n_0 = n_1 = 100$

Method	Fusion Sample X_1	Tilt $h(x)$	Coverage	CI Length	MAE
OSF	Unif(0, 100)	$(x, \log x)$	418/83.6%	0.0606627	0.0174473
	Unif(1, 100)	$(x, \log x)$	417/83.4%	0.0583135	0.0164046
	Unif(0, 100)	$(\log x, \log^2 x)$	459/91.8%	0.0712286	0.0158632
	Unif(1, 100)	$(\log x, \log^2 x)$	445/89.0%	0.0699552	0.0158091
AC	-	-	496/99.2%	0.1138280	0.0212072
EP	-	-	440/88.0%	0.0912952	0.0168400
POT	-	-	452/90.4%	0.0886092	0.0176988
BM	-	-	498/99.6%	0.3745127	0.1971111

performances. When the tilt function is “misspecified”, then in general, the OSF confidence intervals do not give the desired coverage. In this case, both the gamma and the log-normal tilt could not accommodate the underlying reference sample of mercury intake well. However, the misspecification problems can be overcome by ROSF. The precision of the point estimates and the CI coverage can be improved through repeated fusions. ROSF results for the gamma tilt $h(x) = (x, \log x)$ and the log-normal tilt $(\log x, \log^2 x)$ are given in Table 5.2.

The ROSF CI now reaches the desired coverage at the expense of slightly increased interval length. The OSF and ROSF methods demonstrate advantages in cases when the tail probability is extremely small as we will see in the lead case.

For the lead exposure, the threshold $PTWI = 25$ and true tail probability is $p = 1 - G(T) = 0.001$. The true tail probability is much smaller in this case and a much harder one to estimate. Fitting both gamma and log-normal distributions to the 3000 lead intake measurements give almost identical AIC. No model is preferred

Table 5.2: ROSF Interval Coverage and Length for $p = 1 - G(T) = 0.05$, $T = 4$, X_0 is sampled from Mercury Intake, in all cases $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

Tilt $h(x)$ & X_1	N	Coverage	CI Length	MAE
$(x, \log x)$, $X_1 \sim$ Unif(0,100)	5	451/90.2%	0.08480878	0.01663014
	20	469/93.8%	0.09374276	
	100	477/95.4%	0.10193500	
	300	482/96.4%	0.10723810	
$(\log x, \log^2 x)$, $X_1 \sim$ Unif(0,100)	5	452/90.4%	0.08431168	0.01584975
	20	474/94.8%	0.09283605	
	100	484/96.8%	0.09970069	
	300	491/98.2%	0.10364770	

Table 5.3: OSF Interval Coverage and Length for $p = 1 - G(T) = 0.001$, $T = 25$, X_0 is sampled from Lead Intake, in all cases $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

Method	Fusion Sample X_1	Coverage	CI Length	MAE
OSF	Unif(0, 30)	312/62.4%	0.00325662	0.00103555
	Unif(0, 60)	343/68.6%	0.00434648	0.00117528
	Unif(1, 30)	432/86.4%	0.00608691	0.00140467
	Unif(1, 60)	420/84.0%	0.00502910	0.00142479
AC	-	500/100%	0.04602715	0.01849893
EP	-	52/10.4%	0.00306818	0.00183200
POT	-	443/88.6%	0.00713434	0.00180256
BM	-	496/99.2%	0.04644317	0.01605755

as suggested by AIC values. Therefore, the default gamma tilt function is adopted in this case.

The same problem persists for all methods. See Table 5.3 for a detailed comparison of various methods. AC and BM intervals provide perfect coverage with extremely wide confidence intervals. The confidence intervals given by these two methods are too wide to make any practical sense. The traditional EP method completely fails when no "success" is contained in the sample. In this difficult case, POT method still yields reasonable results. It gives point estimate with low MAE, and short confidence intervals with 88% coverage. The performance of the ROSF method is given in Table 5.4.

Table 5.4: ROSF Interval Coverage and Length for $p = 1 - G(T) = 0.001$, $T = 25$, X_0 is sampled from Lead Intake, in all cases $n_0 = n_1 = 100$, $h(x) = (x, \log x)$

Method & X_1	N	Coverage	CI Length	MAE
ROSF, $X_1 \sim$ Unif(0,30)	5	329/65.8%	0.00319330	0.0009782407
	20	398/79.6%	0.00477476	
	100	443/88.6%	0.00628520	
	300	465/93.0%	0.00716078	
ROSF, $X_1 \sim$ Unif(0,60)	5	356/71.2%	0.00444333	0.0011247530
	20	427/85.4%	0.00622525	
	100	467/93.4%	0.00754800	
	300	480/96.0%	0.00821010	
ROSF, $X_1 \sim$ Unif(1,30)	5	436/87.2%	0.00610342	0.001406795
	20	463/92.6%	0.00731402	
	100	479/95.8%	0.00836823	
	300	485/97.0%	0.00903573	
ROSF, $X_1 \sim$ Unif(1,60)	5	429/85.5%	0.00622608	0.001388078
	20	464/92.8%	0.00737810	
	100	481/96.2%	0.00830965	
	300	485/97.0%	0.00883218	

In all cases, the ROSF MAE is lower than POT MAE. It can be seen that useful information about the true magnitude of p can be obtained for $N = 20$. However, a larger N is needed to obtain adequate coverage. By choosing a larger N, the desired 95% coverage can be reached at the expense of slightly increased CI length.

5.1.5 Extreme Quantile Estimation

In this section, the ROSF method is applied to estimate the extreme quantiles for the mercury and lead data. Specifically, the 0.99 and the 0.999 quantiles obtained from 3000 observations are treated as the "true" quantiles. Then, the ROSF method and EVT methods are applied to samples of size 100. Our goal is to check capabilities of various methods in capturing the true quantiles from samples of limited number of observations.

For mercury exposures, the "true" (from the data) 0.99 and 0.999 quantiles are 15.75393 and 108.71895 respectively. Recall that the log-normal tilt $h(x) = (\log x, (\log x)^2)$ is more appropriate. In this example, fusing with two artificial uniform samples yields more reliable confidence intervals for the quantiles. For the 0.99 quantile, the first fusion samples X_1's are generated from Uniform(0,30) and the second fusion samples X_2's are generated from Uniform(0,200). For the 0.999 quantile, the first fusion samples X_1's are generated from Uniform(0,150), and the second fusion samples X_2's are generated from Uniform(0,500). For each confidence interval, F_Q is obtained based on $n_r = 2500$ and $N = 3000$. Detailed comparison of the performance of various methods is given in Table 5.5.

For lead exposures, the "true" 0.99 and 0.999 quantiles are 9.976886 and 21.264956 respectively. Recall that fitting both gamma and log-normal distribution to lead intake measurements gives almost identical AIC. No model is preferred as suggested by AIC values. We only present results from applying the log-normal tilt function. For the 0.99 quantile, fusion samples X_1's are generated from Uni-

Table 5.5: ROSF Interval Coverage and Length for 0.99 and 0.999 quantiles, X_0 is sampled from mercury intake, $n_0 = n_1 = n_2 = 100$, $h(x) = (\log x, (\log x)^2)$

		ROSF	BM_{Delta}	BM_{Boot}	$\text{POT}_{\text{Delta}}$	POT_{Boot}	EP
$q_{0.99}$	Coverage	96.6%	75.4%	83.6%	69.4%	76.2%	60.8%
	CI Length	23.50	29.25	36.82	23.41	25.37	52.21
	MAE	4.45	6.79	6.49	6.47	6.32	15.12
$q_{0.999}$	Coverage	95.7%	66.2%	85.6%	59.0%	73.4%	10.2%
	CI Length	224.36	283.26	841.99	196.11	320.27	51.60
	MAE	55.41	78.39	75.91	79.91	81.30	88.62

form(0,16); for the 0.999 quantile, the fusion samples X_1's are generated from Uniform(0,25). For each confidence interval, F_Q is obtained based on $n_r = 2000$ and $N = 3000$. Detailed comparison of the performance of various methods is given in Table 5.6.

Table 5.6: ROSF Interval Coverage and Length for 0.99 and 0.999 quantiles, X_0 is sampled from lead intake, $n_0 = n_1 = 100$, $h(x) = (\log x, (\log x)^2)$

		ROSF	BM_{Delta}	BM_{Boot}	$\text{POT}_{\text{Delta}}$	POT_{Boot}	EP
$q_{0.99}$	Coverage	95.6%	90.0%	92.0%	85.0%	84.4%	64.4%
	CI Length	6.35	21.66	22.89	15.35	15.10	11.34
	MAE	1.38	2.85	2.69	2.61	2.62	2.90
$q_{0.999}$	Coverage	97.4%	93.8%	95.4%	90.0%	91.4%	10.4%
	CI Length	8.42	184.19	405.48	104.72	150.31	10.69
	MAE	1.57	22.85	15.05	11.73	11.14	10.28

For both cases, the confidence intervals for quantiles produced by ROSF are in general much shorter. Yet, such short intervals have very high coverage for the true quantiles. Furthermore, the MAE of the ROSF estimates are much smaller the the MAE of other methods, meaning the ROSF estimates are very close to the true quantiles on average. From the above results, we can confidently conclude that for samples of limited number of observations, the ROSF outperforms the other methods in terms of accuracy and reliability.

5.2 Application in a Clinical Trial

In the drug development process, clinical research is a critical step to study how the drug will interact with the human body. Drug developers and researchers design and carry out clinical trials to evaluate drug efficacy and safety before marketing approval is granted by the US Food and Drug Administration (FDA). Typical clinical trials consist of three phases, from early stage small-scale, shorter duration Phase 1 studies to late stage, large scale, long duration Phase 3 studies. Phase 1 studies may involve 20 to 100 volunteers or people with the disease conditions and last for several months. Phase 3 studies typically involve 300 to 3,000 volunteers with the disease conditions and last for one up to four years.

Drug safety is a crucial part of clinical research where researchers focus on detection, assessment, and monitoring side effects, adverse effects and toxicity of pharmaceutical products. Due to the high cost of drug development, opportunity cost of investing in one compound rather than another and risk to patients, there is a strong desire for researchers and pharmaceutical companies to be able to detect potential toxic effects in early stages of the drug development process.

In this application, we will show how the ROSF method is applicable in clinical trials to evaluate drug toxicity related to liver health issues.

5.2.1 Data

According to the FDA guidance [18] and Common Terminology Ceriteria guidelines [27], liver toxicity is typically reflected by the elevation in values of certain clinical laboratory measurements. Alanine Aminotransferase (ALT) levels are commonly measured (in units per liter or U/L) clinically as biomarker for liver health. Elevation of ALT levels during the time period a drug is used suggests potential hepatotoxic effects of the drug. The probabilities that post-medication ALT level exceeds the upper limit of normal (ULN) are of interest.

The laboratory dataset of a drug developed by AstraZeneca is contained in the R package *texmex* (see Southworth and Heffernan 2015 [46]). The dataset consists of 606 observations on 9 variables related to liver health from a randomized, blind, parallel group clinical trial with four doses of the drug. The response variables include baseline (prior to treatment) and post-medication (on treatment) measurements of ALT (alanine aminostransferase), AST (aspartate aminotransferase), ALP (alkaline phosphatase), TBL (total bilirubin) and a dosage group indicator (a factor variable with levels A B C D).

Biologically, liver cells release ALT and AST as they die causing elevation in ALT and AST levels. If a sufficient amount of liver cells is destroyed, liver fails to function properly and ceases to clear bilirubin which leads to a rise in TBL level. Furthermore, ALP level may also go up as a consequence of blockage in the liver. There is a common understanding that ALT is a more sensitive biomarker for potential liver injury, thus the ALT level will be the focus of this study.

The doses are equally spaced on a log scale where dose D is twice dose C, dose C is twice dose B, and dose B is twice dose A. There are 152 ,148 148, and 158 subjects in each dose group respectively. The purpose of the study is to assess the capability of the proposed ROSF method in predicting drug toxicity in early stage of clinical trials. To accomplish this goal, samples of 60 patients are obtained from each dosage group. For each dose group, the probabilities of exceeding values of interest and the 100-patient return level are predicted based on samples of size 60. Specifically, tail probabilities P$(ALT > ULN)$, P$(ALT > 3 \times ULN)$, P$(ALT > 10 \times ULN)$, and 0.99 quantiles will be predicted by ROSF and a extreme value modeling approach for different dosage levels separately. ULN for ALT is defined differently in different studies, ranging from 30 to 48 U/L. All our analysis treat ULN as being 30 U/L.

In practice, the probabilities of exceeding specified multiples of ULN in early stage clinical trials are typically lower than their corresponding probabilities in late stage trials. There is more opportunity to observe extreme ALT elevations when more patients are included in clinical trials and when trials are of longer duration. To take this phenomenon into account, subsamples of size 60 are taken so that empirical tail probabilities and quantiles of the subsample are lower than their corresponding full-sample observed counterparts. In this setup, whether the nature and magnitude of the toxic effect of the drug on ALT levels could be adequately characterized by the EVT and ROSF method in early stage clinical trials can be examined. The histogram of the ALT levels from the full-sample and subsample are shown in figure 5.3.

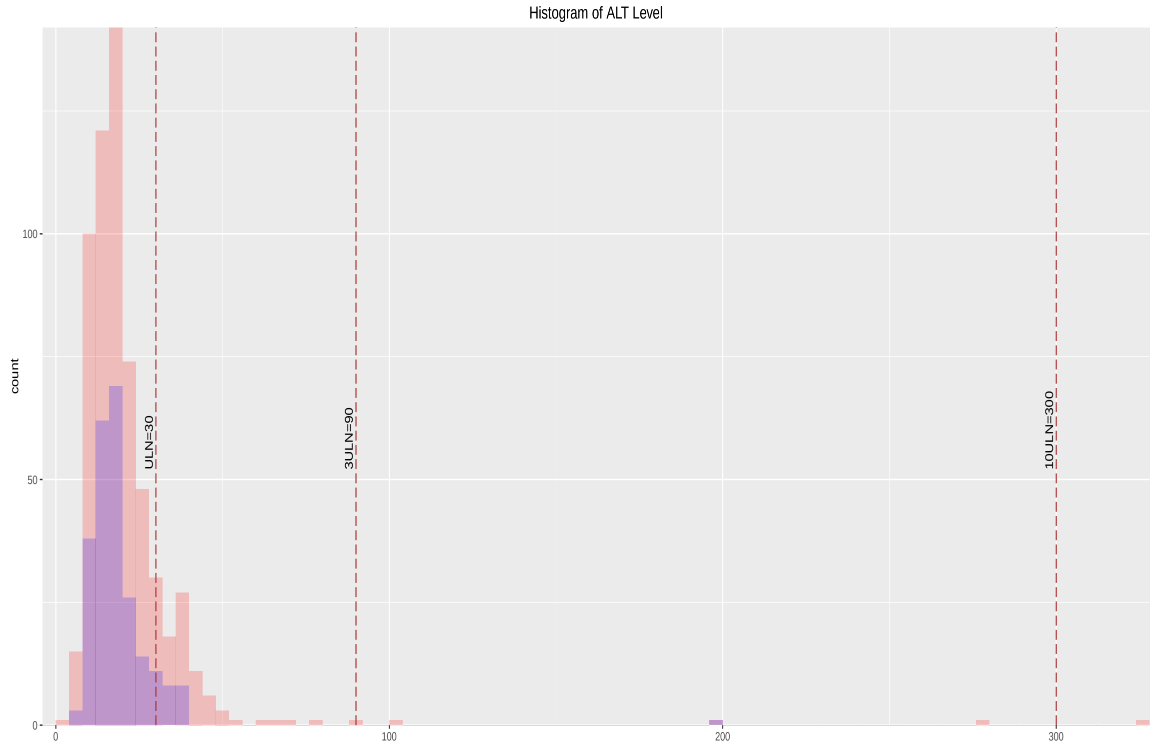

Figure 5.3: Histogram of ALT levels. The full sample consists of 606 observations are represented by red bars. The subsample consists of 240 observations are in blue.

5.2.2 Extreme Value Modeling

Southworth and Heffernan (2014) [45] proposed a two stage modeling procedure in predicting liver toxicity. In the first stage, a robust linear model is fitted to the data to take account the baseline effect. The peaks over threshold (POT) method is applied to the residuals from the robust regression in the second stage. In other words, the generalized Pareto Distribution (GPD) is fitted to the residuals above a predetermined threshold. The two stage modeling is adopted in this analysis

Table 5.7: Parameter Estimates from Robust Linear Model

	Value	SE	t-value
Intercept	0.406	0.121	3.350
log Baseline ALT	0.817	0.044	18.726
Dose	0.101	0.016	6.140

for comparison purposes.

As illustrated by the histogram of the ALT measurements,the data is non-normal and highly skewed. Thus the robust linear mode using MM-estimation [35] with Gaussian efficiency set to 85% and bisquare weight functions is preferred. The model

$$\log(ALT.M) \approx \log(ALT.B) + dose$$

is entertained. The log-transformed baseline and post-baseline ALT values are used so that the distribution of residuals are near symmetric and the assumption of the robust regression model holds approximately. The parameter estimates are given in Table 5.7.

The boxplots of the scaled residuals from the linear model is given in Figure 5.4. In the second stage, the POT method is applied to the residuals in the figure. Southworth and Heffernan (2014) [45] suggested that it is appropriate to fit GPD to residuals above the 70th quantile. The standard threshold selection methods, namely the mean residual life plot also indicates that the 70th quantile would be a reasonable threshold u.

The GPD model can be fitted with covariates in σ and/or ξ. Various models with covariates are fitted and model selection is performed based on AIC. The log-

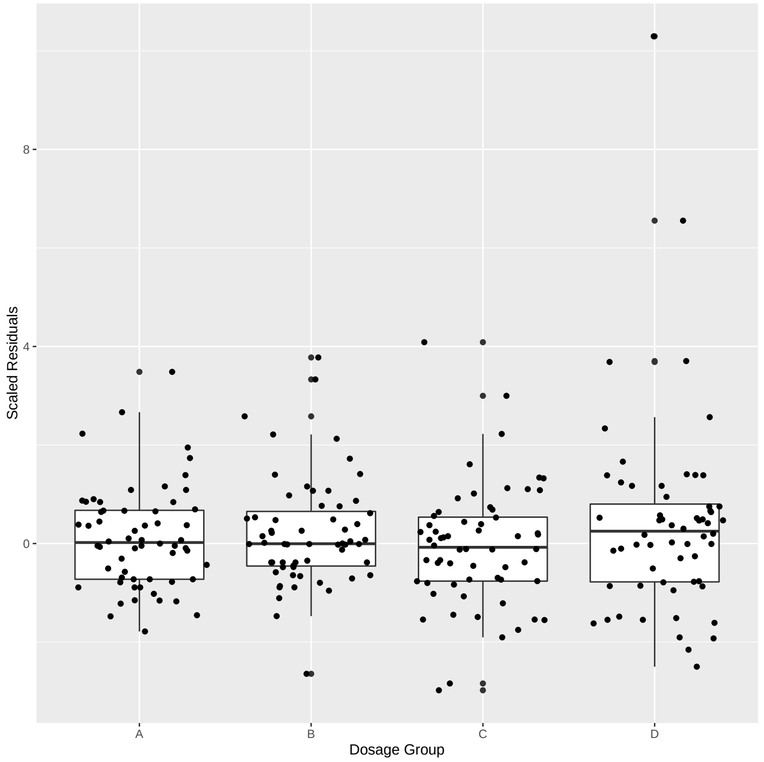

Figure 5.4: Boxplot of Scaled Residuals from Robust Linear Model

likelihood and the AIC for models considered are given in Table 5.8. According to AIC, the preferred model is the one with a term for dosage in the parameter ξ.

To take into account the uncertainty in the parameter estimates, Southworth (2014) suggested that the preferred model should be refitted by a Markov Chain Monte Carlo method. The diffusion Gaussian priors $\mathrm{N}(0, 10^4)$ were adopted for parameters so that predictions would depend on the data instead of the choice of the prior. The standard checks of the chains show that the algorithm has converged on the target distribution. Burn-in and thinning is performed to obtain a chain that is closer to independent. The predicted exceedance probabilities and 100-patient return levels are based on the simulated posterior distributions.

Table 5.8: GPD Models Considered, Number of Parameters, Log-likelihoods and AIC's

Model	No. Parameters	Log-likelihood	AIC
Null Model	2	19.64	-35.28
σ, ξ = f(factor(dose))	8	22.48	-29
σ, ξ = f(ndose)	4	21.48	-34.95
σ = f(ndose)	3	20.77	-35.55
ξ = f(ndose)	3	21.43	-36.86

5.2.3 ROSF

For each dosage group, the ROSF method is applied to predict the probabilities of exceeding multiples of ULN. In all cases, the gamma tilt function $h(x) = (x, \log x)$ is used. In each run, the reference sample of size 60 is fused with artificially generated uniform samples of size 60. Each predicted exceedance probability is based on 1000 repetitions of OSF run. The point estimate of the tail probability is the mean of the 1000 OSF point estimates. As described in Chapter 3, we obtain the empirical distribution for the upper bounds B_i (F_B) based on the sequence of 1000 B_i's. Then the upper bound of the ROSF interval is given by $F_B^{-1}(\alpha^{1/N})$. For all cases, we let $N = 5$ which approximately corresponds to the median of 1000 OSF upper bounds.

In this application, some tail probabilities of interest is not very small (eg. the observed P($ALT > ULN$) for group D patients is 0.2152). Therefore, it is no longer appropriate to use 0 as the lower bound. The ROSF lower bounds can be obtained in a very similar way as with the ROSF upper bounds B_i. Suppose $A_1, \ldots, A_N$ is a sequence of i.i.d. lower bounds from distribution F_A. Then considering the minimum of $A_1, \ldots, A_N$. The inequality $F_A^{-1}(1 - 0.05^{1/N}) \leq p$ holds with at least

95% confidence. Thus $F_A^{-1}(1 - 0.05^{1/N})$ can be used as the lower bound, and this decreases to 0 as N increases. For all cases, we let $N = 5$ which approximately corresponds to the median of 1000 OSF lower bounds.

The 100-patient return level or the 0.99 quantile is also predicted by the ROSF method as described in Chapter 4.

5.2.4 Results

The predicted exceedance probabilities from EVT based modeling and ROSF together with the sub-sample and the full-sample observed exceedance probabilities are given in in Figure 5.5 and Table 5.9.

The predicted 100-patient return level or the 0.99 quantile $\hat{q}_{0.99}$ from the two-stage EVT and ROSF are given in in Figure 5.6 and Table 5.10.

Despite that the empirical tail probabilities and quantiles from the subsample are lower than their corresponding observed counterparts obtained from the full sample, both ROSF and EVT demonstrated power in making extrapolation out of sample. There is only one case when the EVT confidence interval failed to captured the "true" tail probability $\mathrm{P}(ALT > 3ULN)$ for dose level D. In all other cases, the "true" tail probabilities and quantiles are contained in the confidence intervals produced by both ROSF and EVT. From the results, we may conclude with confident that the nature and magnitude of the toxic effect of the drug on ALT levels could be adequately characterized by the ROSF and EVT method in early stage clinical trials. Given early stage data with reasonable degree of characteristics of the late

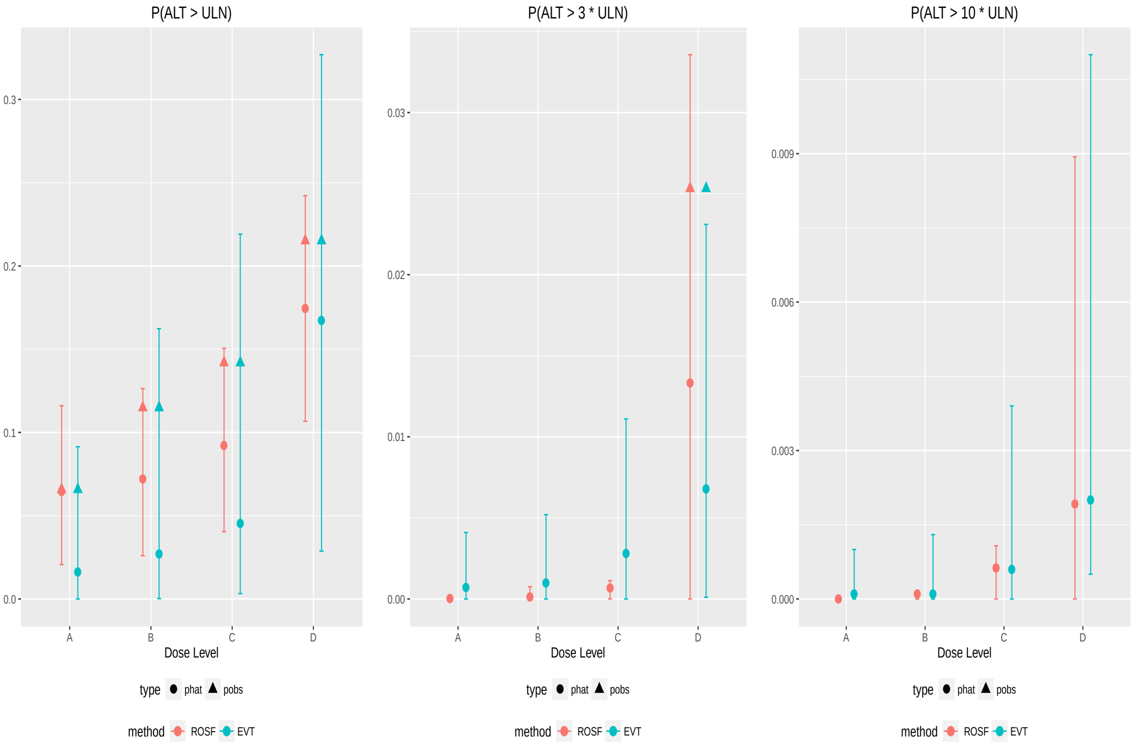

Figure 5.5: Predicted Probabilities of Exceeding 1, 3 and 10 Times the Upper Limit of Normal (ULN). Note that the Horizontal Scale Vary from Panel to Panel.

stage data, the predictions based on ROSF and EVT method would have suggested a potential liver toxicity. However, in many cases, the EVT confidence intervals are often too wide to be informative, especially for P$(ALT > ULN)$. ROSF provides an alternative way to obtain good predictions with short and reliable confidence intervals.

5.3 Discussion

In this section, we have described a new method (ROSF) in assessing risks in food safety and drug safety. The two quantities of interest in quantitative risk

Table 5.9: Predicted vs. Observed Probabilities of Exceeding Specified Multiples of ULN

Dose	Method	P($ALT > ULN$)	P($ALT > 3ULN$)	P($ALT > 10ULN$)
A	ROSF	0.06446 (0.02065,0.11602)	0.00002154 (0,0.00022631)	0.000002887 (0,0.00002923)
	Two Stage EVT	0.0162 (0,0.0914)	0.0007 (0,0.0041)	0.0001 (0,0.0010)
	Empirical	0	0	0
	Observed	0.06579	0	0
B	ROSF	0.07222 (0.02601,0.12635)	0.0001294 (0,0.0007516)	0.00010035 (0,0.00016794)
	Two Stage EVT	0.0271 (0.0002,0.1623)	0.0010 (0,0.0052)	0.0001 (0,0.0013)
	Empirical	0.08330	0	0
	Observed	0.11490	0	0
C	ROSF	0.09190 (0.04028,0.15016)	0.000689 (0,0.001127)	0.0006267 (0,0.0010731)
	Two Stage EVT	0.0453 (0.0032,0.2191)	0.0028 (0.0000, 0.0111)	0.0006 (0,0.0039)
	Empirical	0.1000	0	0
	Observed	0.1419	0	0
D	ROSF	0.1745 (0.1067,0.2422)	0.01333 (0,0.03357)	0.001917 (0,0.008936)
	Two Stage EVT	0.0678 (0.0073,0.3269)	0.0068 (0.0001,0.0231)	0.002 (0,0.011)
	Empirical	0.1167	0.01667	0
	Observed	0.2152	0.02532	0.006329

assessment include: threshold exceedance probabilities and extreme quantiles. We have shown that these two quantities can be precisely and reliably estimated through density ratio model by repeatedly fusing a given reference sample with computer generated uniform data. In real data applications, one could use different uniform fusion samples as done in the food safety application for validation purposes (See Table 5.3 and 5.4 for examples). The ROSF method is quite robust, similar results

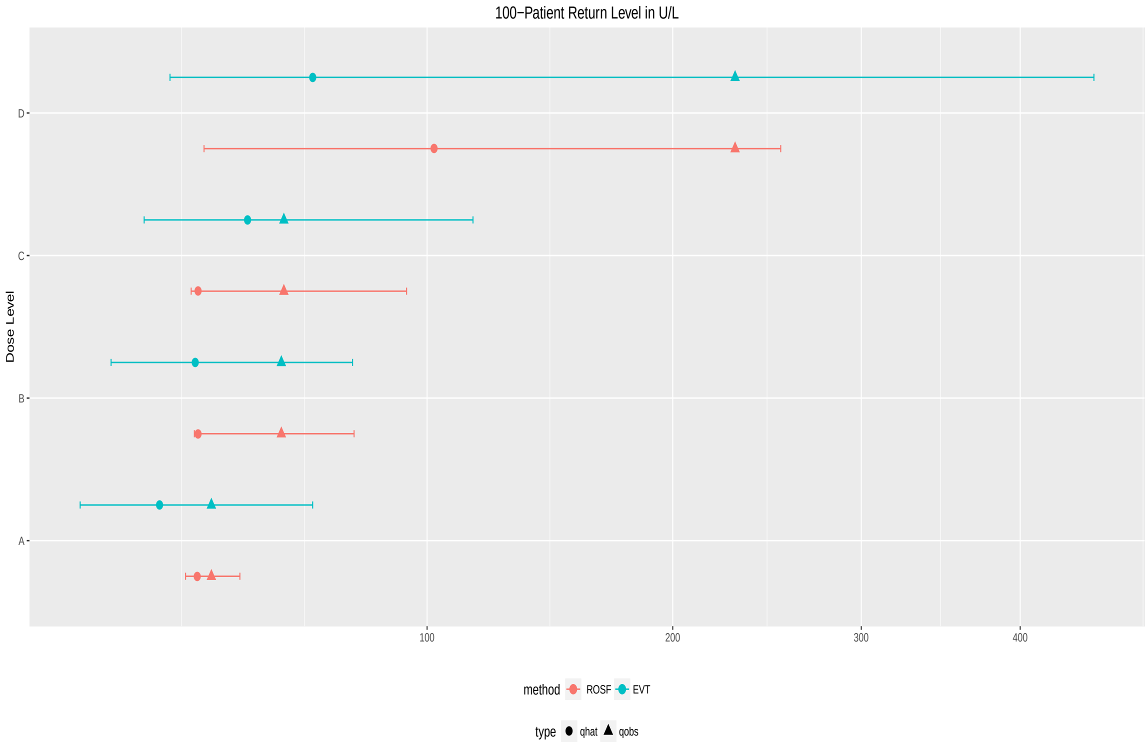

Figure 5.6: Predicted 100-patient return levels in U/L for Dosage A, B, C, D.

Table 5.10: Predicted vs. Observed 100-patient return level in U/L

Method	A	B	C	D
ROSF	37.54 (35.14,46.84)	37.71 (36.92,76.90)	37.71 (36.27,93.21)	102.40 (38.93, 254.80)
Two Stage EVT	30.14 (17.23,65.14)	37.11 (21.83,76.45)	48.63 (27.35,116.08)	65.20 (32.10, 451.27)
Empirical	26.64	37.23	37.41	103.37
max(x_0)	29.00	39.00	38.00	196.00
Observed	40.49	56.89	57.54	230.83

are obtained when different uniform fusion samples are used. The robustness of the method is supported by additional experiments not reported in this dissertation. Tools for quantifying the robustness of the models can be developed in future works.

In the dissertation, the gamma tilt function $h(x) = (x, \log x)$ and the log-normal tilt function $h(x) = (\log x \log^2 x)$ are considered. It is shown that the gamma tilt and the log-normal tilt together accommodate a wide range of skewed distributions in quantitative risk assessment. Natural extensions could consider more types of tilt functions. The effects and impacts of different types of tilt function needs to be studied in the future.

The construction of the ROSF confidence intervals involves choosing a tuning parameter N. The length of the confidence interval depends on the choice of N. As N becomes larger, the coverage of the ROSF confidence intervals for tail probabilities improve at the expense of slightly increased length of the intervals. From many simulation results,the choice $N = 100$ is prudent across most specified and misspecified cases. How to choose the optimal N is still an open problem. We plan to explore this problem in future works.

Furthermore, ROSF can be applied whenever it is required to estimate exceedance probabilities and/or large quantiles. An example in point is the estimation of the predictive distribution in time series, given that the time series is represented by a regression model (e.g. Kedem and Gagnon (2010 [30]). When the residuals follow approximately the normal distribution, it is sensible to fuse the residuals with artificial normal data. Then, the normal tilt function $h(x) = (x, x^2)$ can be adopted and the predictive distribution can be obtained through the density ratio model. In Kedem and Gagnon (2008) [29], instead of artificial fusion samples, residuals from several sources were fused. The ROSF method, on the other hand fuses the residuals with external samples. The process can be repeated as many times as desired for validation purposes.

Finally, given appropriate tilt functions, the density ratio model holds for both univariate and multivariate data. Hence, ROSF can be extended to multivariate data as well. In the multivariate setting, useful tilt functions can by suggested from the ratio of two multivariate distribution (e.g. the ratio of two multivariate normal distribution). Extending the ROSF method to multivariate case is another possible direction of future research.

Appendix A: Asymptotic Theory for p Quantile $\hat{q}_p$

The asymptotic theory for p Quantile $\hat{q}_p$ based on the semiparametric density ratio model is given in Chen and Liu (2013) [8]. The estimator $\hat{q}_p$ will be referred to as SP quantile for simplicity.

Theorem A.1. *Assume Density Ratio Model 2.4 holds, and the density function $g(x)$ is continuous and positive at $x = q_p$. Then the density ratio model based p quantile estimator $\hat{q}_p$ has Bahadur representation:*

$$\hat{q}_p = q_p - \frac{\hat{G}(q_p - p)}{g(q_p)} + O(n^{-3/4}(\log n)^{1/2}).$$

With the Bahadur representation and the multivariate asymptotic normality of the $\hat{G}$'s given in Appendix A, the multivariate asymptotic normality of the SP quantiles can be obtained. We define q_i be the population quantile of the ith population in the DRM at some level p_i, and similarly let q_j be the population quantile of the jth population in the DRM at some level p_j. We further denote an arbitrary term in the covariance matrix of the process $\sqrt{n}(\hat{G}(t) - G(t))$ in Theorem 2.2 by $v_{i,j}(x, y)$. Then the following theorem describes the asymptotic behavior of the SP quantiles.

Theorem A.2. *Assume Density Ratio Model 2.4 holds, the process*

$$\sqrt{n}(\hat{q}_i - q_i, \hat{q}_j - q_j)$$

is asymptotically bivariate normal with mean zero and covariance matrix

$$\Sigma = \begin{pmatrix} v_{ii}(q_i, q_i)/g_i^2(q_i) & v_{ij}(q_i, q_j)/\{g_i(q_i)g_j(q_j)\} \\ v_{ij}(q_i, q_j)/\{g_i(q_i)g_j(q_j)\} & v_{jj}(q_j, q_j)/g_j^2(q_j) \end{pmatrix}$$

where v_{ij} is the i, jth component in the covariance matrix of the process $\sqrt{n}(\hat{G}(t) - G(t))$ given in Theorem 2.2.

For the detailed derivation of theorem A.1 and A.2, the reader may refer to Chen and Liu (2013) [8].

Bibliography

[1] A. Agresti and B. A. Coull, *Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions*, American Statistician, **52**, 119-126, (1998).

[2] R. R. Bahadur *A Note on Quantiles in Large Samples*, The Annals of Mathematical Statistics **37**, No. 3, 577-580, (1966).

[3] J. Beirlant, Y. Goegebeur, J. Segers and J. Teugels, *Statistics of Extremes Theory and Application* (Wiley, 2004).

[4] G. Casella, R. L. Berger, *Statistical Inference*, (Cengage Learning, 2002).

[5] L. D. Brown, T. T. Cai and A. DasGupta, *Interval Estimation for a Binomial Proportion*, Statistical Science, **16**, 2, 101-133, (2001).

[6] L. D. Brown, T. T. Cai and A. DasGupta,, *Confidence Intervals for a Binomial Proportion and Asymptotic Expansions*, Annals of Statistics, **30**, 1, 160-201, (2002).

[7] Center for Disease Control and Prevention *Making Food Safer to Eat*, (http://www.cdc.gov/VitalSigns/foodsafety/, Access Date: Jul 1, 2015).

[8] J. Chen, Y. Liu. *Quantile and Quantile Function Estimations Under Density Ratio Model*, The Annals of Statistics **41**, No. 3, 1669-1692, (2013).

[9] D. Claisse, D. Cossa, G. Bretaudeau-Sanjuan, G. Touchard, B. Bombled *Methylmercury in Molluscs along the French Coast*, Marine Pollution Bulletin **42**, 329-332, (2001).

[10] S. Coles *An Introduction to Statistical Modeling of Extreme Values*, (Springer, 2001).

[11] A. Crepet, H. Harari-Kermadec, J. Tressou, *Using Empirical Likelihood to Combine Data: Application to Food Risk Assessment*, Biometrics, **65**, 257-266, (2009).

[12] A. C. Davison and R. L. Smith, *Models for Exceedances over High Thresholds*, Journal of the Royal Statistical Society, **B**, 53, 393-442, (1990).

[13] A. Dvoretzky, J. Kiefer and J. Wolfowitz, *Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator*, The Annals of Mathematical Statistics, **27**, 3, 642-669, (1956).

[14] P. Embrechts, C. Kluuppelberg, T. Mikosch *Modeling Extremal Events for Insurance and Finance* (Springer, 1997).

[15] Food Consumption and Exposure Assessment of Chemicals, *Report of a FAO/WHO Consultation*, 10-14 (1997).

[16] Evaluation of Certain Food Additives and Contaminants for Lead and Methylmecury, *Fifty Third Report of the Joint FAO/WHO Expert Committee on Food Additives, Technical Report Series 896*, (1999).

[17] Evaluation of Certain Food Additives and Contaminants for Cadmium, *Fifty Third Report of the Joint FAO/WHO Expert Committee on Food Additives, Technical Report Series 901*, (2000).

[18] US Food and Drugs Administration, *Guidance for Industry Drug-Induced Liver Injury: Premarketing Clinical Evaluation*, (2008).

[19] W. Fithian and S. Wager, *Semiparametric Expoential Families for Heavy-Tail Data*, Biometrika, **102** 486-493, (2015).

[20] K. Fokianos, B. Kedem, J. Qin, D. Short *A semiparametric approach to the one-way layout*, Technometrics, **43**, 56-65, (2001).

[21] K. Fokianos, I. Kaimi, *On the Effect of Misspecifying the Density Ratio Model*, Annals of the Institute for Statistical Mathematics **58**, 475-497, (2006).

[22] J. Gauchi, J. Leblanc, *Quantitative Assessment of Exposure to the Mycotoxin Ochratoxin A in Food*, Risk Analysis, **22**, 219-234, (2002).

[23] R. D. Gill, Y. Vardi, J. A. Wellner, *Large Sample Theory of Empirical Distribution in Biased Sampling Models*, Annals of Statistics, **16**, 3, 51-81, (1988).

[24] L. Haan and A. Ferreira, *Extreme Value Theory An Introduction*, (Springer, 2006).

[25] P. Hall and I. Weissman, *On the Estimation of Extreme Tail Probabilties*, Annals of Statistics, **25**, 3, 1311-1326, (1997).

[26] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, (Springer, 2009).

[27] US Department of Health and Human Services and National Institutes of Health and Cancer Institute *Common Terminology Criteria for Adverse Events Version 4*, (2009).

[28] M. Katzoff, W. Zhou, D. Khan, G. Lu, B. Kedem, *Out of Sample Fusion in Risk Prediction*, Journal of Statistical Theory and Practice, **8**, 3, 444-459, (2014).

[29] B. Kedem, G. Lu, and P. D. Williams, *Forecasting Mortality Rates Via Density Ratio Modeling*, Canadian Journal of Statistics, **36**, 193-206, (2008).

[30] B. Kedem, RE. Gagnon, *Semiparametric Distribution Forecasting*, Journal of Statistical Planning and Inference, **140**, 3734-3741, (2010).

[31] B. Kedem, L. Pan, W. Zhou, and C. Coelho, *Interval estimation of small tail probabilities applications in food safety*, Statistics in Medicine, DOI: 10.1002/sim.6921, (2016).

[32] J. Kiefer and J. Wolfowitz, *Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters*, The Annals of Mathematical Statistics, **27**, 4, 887-906, (1956).

[33] M. Leadbetter, G. Lindgren, H. Rootzen, *Extremes and Related Properties of Random Sequences and Processes* (Springer, 1983).

[34] G. Lu, *Asymptotic Theory for Multiple-Sample Semiparametric Density Ratio Models and its Application to Mortality Forecasting*, Ph.D. Dissertation, University of Maryland, College Park, (2007).

[35] R. Maronna, D. Martin, and V. Yohai, *Robust Statistics: Theory and Methods*, (Wiley, 2006)

[36] A. B. Owen, *Empirical Likelihood*, (Chapman and Hall/CRC, Boca Raton, 2001).

[37] G. P. Patil, and C. R. Rao, *Weighted Distributions and Size-Biased Sampling with Applications to Wildlife Populations and Human Families*, Biometrics, **34**, 2, 179-189, (1978).

[38] M. J. Paulo, H. van der Voet, J. C. Wood, G. R. Marion, J. D. van Klaveren, *Analysis of Multivariate Extreme Intakes of Food Chemicals*, Food and Chemical Toxicology, **44**, 994-1005, (2006).

[39] J. Qin, J. Lawless *Empirical Likelihood and General Estimating Equations*, The Annals of Statistics, **22**, No. 1 300, (1994).

[40] J. Qin, B. Zhang, *A Goodness of Fit Test for Logistic Regression Models Based on Case-control Data*, Biometrika, **84**, 609-618, (1997).

[41] J. Qin, *Inferences for Case-Control Data and Semiparametric Two-Sample Density Ratio Models*, Biometrika, **85**, 619-630, (1998).

[42] S. Resnick, *Extreme Values, Point Processes and Regular Variation* (Springer, 1987).

[43] R. L. Smith, *Maximum Likelihood Estimation in a Class of Nonregular Cases*, Biometrika **75**, 67-90, (1985).

[44] H. Southworth and E. Heffernan, *Extreme Value Modelling of Laboratory Safety Data from Clinical Studies*, Pharmaceutical Statistics, **11**, 5, (2012).

[45] H. Southworth, *Predicting Potential Liver Toxicity from Phase 2 Data: A Case Study with Ximelagatran*, Statistics in Medicine (2014).

[46] H. Southworth, and E. Heffernan, *Statistical Modeling of Extreme Values*, (https://cran.r-project.org/web/packages/texmex/texmex.pdf, Access Date: Mar 1, 2015).

[47] J. Tressou, A. Crepet, P. Bertail, M. H. Feinberg, J. Leblanc *Probabilistic Exposure Assessment to Food Chemicals Based on Extreme Value Theory. Application to Heavy Metals from Fish and Sea Products*, Food and Chemical Toxicology **42**, 1349-1358, (2004).

[48] Y. Vardi, *Nonparametric Estimation in the Presence of Length Bias*, The Annals of Statistics **10**, No. 2, 616, (1982).

[49] Y. Vardi, *Empirical Distributions in Selection Bias Models*, The Annals of Statistics **13**, No. 1, 178, (1985).

[50] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes with Applications to Statistics*, (Springer, 1996).

[51] A. Voulgaraki, B. Kedem, and B. I. Graubard, *Semiparametric Regression in Testicular Germ Cell Data*, Annals of Applied Statistics, **6**, 3 1185-1208, (2012).

[52] B. Zhang, *Quantile Estimation under a Two-Sample Semiparametric Model*, Bernoulli, **6**, 491-511, (2000).

[53] W. Zhou, *Out of Sample Fusion*. Ph.D. Dissertation, University of Maryland, College Park, (2013).