Lower Quantile Estimation of Wood Strength Data

by

Yang Liu

Bachelor of Science, Nankai University, 2010

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2012

Abstract

In wood engineering, lower quantile estimation is vital to the safety of the construction with wood materials. In this thesis, we will first study the censored Weibull maximum likelihood estimate (MLE) of the lower quantile as in the current industrial standard D5457 (ASTM, 2004a) from a statistical point of view. According to our simulations, the lower quantile estimated by the censored Weibull MLE with the 10th empirical percentile as the threshold has a smaller mean squared error (MSE) than the intuitive parametric or non-parametric quantile estimate. This advantage can be shown to be achieved by a good balance between the variance and bias with the help of subjective censorship.

However, the standard D5457 (ASTM, 2004a) only utilizes a small (10%) and ad-hoc proportion of the data in the lower quantile estimation, which stimulates us to further improve it. First, we can consider fitting a more complex model, such as the Weibull mixture, to a larger, (e.g., 70%) proportion of the data set with the subjective censorship, which leads to the censored Weibull mixture estimate of the lower quantile. Also, the bootstrap can be used to select a better censoring threshold for the censored Weibull MLE, which leads to the bootstrap censored Weibull MLE. According to our simulations, both proposals can yield a better lower quantile estimates than the standard D5457 and the bootstrap censored Weibull MLE is better than the censored Weibull mixture.

Table of Contents

List of Tables

List of Figures

Acknowledgments

Foremost I would like ot express my deep gratitude to my supervisors, Professor Ruben H. Zamar and Professor Matías Salibián-Barrera. Their inspiration, guidance and encouragement and great insight for my research helped me through the last two years.

Moreover, as a member of the Forest Products Stochastic Modeling Group, I am grateful to the group founders Professor James V. Zidek and Professor William J. Welch for the topic and funding for my research. I would also like to thank Mr. Conroy Lam, Mr. Isaec Chiu, Mr. Philp Eng from Forest Products Innovation for their help in the data collection and background knowledge. Also, I am grateful to Professor James V. Zidek for serving on my examining committee and for his valuable comments and suggestions.

Finally, I would like to give my heartful appreciation and gratitude to my parents for their support and encouragement on my study. This thesis is dedicated to them.

To my parents

Chapter 1

Introduction

In Canada, wood (lumber) products are very important in the building construction and they are used for roofs, walls and floors in many buildings Johnson et al. (2003). To ensure the safety and duration of those wood buildings, it is necessary to study the strength properties of those wood products. As introduced in Cheng (2010), wood strength properties are described by several strength measures, such as the Modulus of Rupture (MOR, strength of the lumber board under pressure vertical to the grain[1]), ultimate tensile strength (UTS, strength of the specimen under tension in the direction of the grain) and so on.

Among all these measures, the MOR is one of the most important, as it measures what pressure a board can bear when something is placed upon it and thus applying a force vertical to its grain, which is the typical pressure seen on a roof or floor of wood buildings. Therefore, MOR is the focus of our current study. In the laboratory, the MOR of a lumber board can be measured by gradually increasing the stress on the board and recording the stress level at which the board breaks. Obviously, this testing procedure is destructive—a board cannot be used for construction after we find out its MOR, and thus it is impossible to measure the MOR of every single board that will be used in the construction.

---

[1]This is an relatively unusual usage of this word, which denotes the direction that the tree grows

Moreover, when compared to other materials such as metal or concrete, wood products have more variable strength properties: Different trees, even of the same specie, have different MORs. Even two boards cut from the same tree, they can have different MORs as they are from different parts of the tree. In another words, the MOR of a board is totally at random.

For such stochastic properties of materials, people can only try to ensure that the probability of failure is smaller than a designed small fraction, for example, 5%, which also means that the load on boards should not be larger than the 5th percentile of the population of the MOR. However, as it is impossible to break all the boards to find the true population 5th percentile, the engineers need to estimate this 5th percentile with some statistical methods. This thesis is thus focused on the statistical methods of estimating the 5th percentile.

1.1　Data collection and example data sets

To estimate a quantile, engineers first randomly sample a few hundred (e.g., 300) boards from a population, for example, from a certain mill, and then destructively test them to obtain the MORs of this sample of lumber boards. A more detailed description of the test can be found in (Cheng, 2010).

Currently, our research group has two data sets of MORs, one with 98 observations and the other one with 282 observations. The MOR values have been rescaled for the convenience of numerical analysis. Their histograms are shown in Figure 1.1. These two data sets will be denoted by MOR1 or MOR2 respectively in our study. We will use models imitating these two data sets to evaluate different methods of lower quantile estimation. Due to the limitation of time and requirements of our collaborator–Forrest Product Innovations, we study only the 5th percentile estimate, but the methods discussed here can surely be applied to other lower quantiles, such as the 1st percentile.

Figure 1.1: Histogram of MOR1 and MOR2
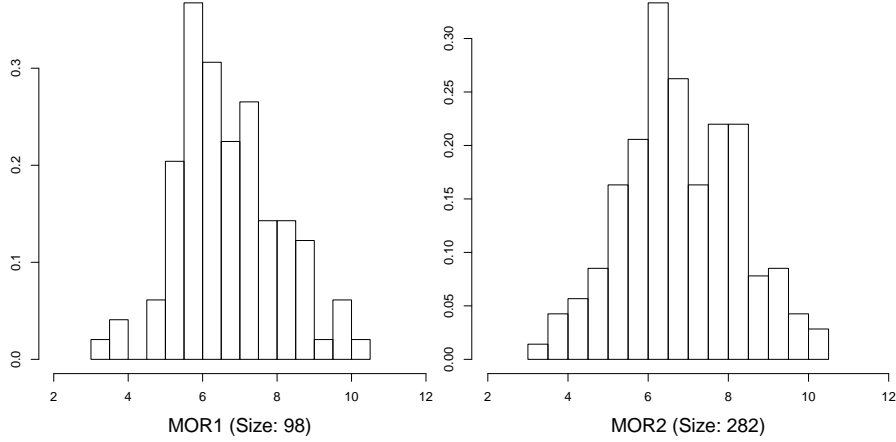
1.2 Quality of a quantile estimate

Before we introduce any quantile estimate, we will first discuss how to evaluate the quality of a quantile estimate. First, it has to be accurate (i.e. unbiased) as overestimation of the lower quantile would entail a higher risk of damage for the structure while underestimation would cause unnecessary waste of construction material and higher construction costs. Therefore, achieving accurate estimation of lower quantiles is an important statistical problem. It is also necessary to control the standard errors of quantile estimates to avoid having to collect a huge (and costly) data set to obtain such good estimates. In our work, the performance of quantile estimate is measured by the mean square error (MSE), which covers both accuracy and efficiency:

$$E_G(\tilde{q}_n - q)^2 = Var(\tilde{q}_n) + (E_G\tilde{q}_n - q)^2$$

where the G is the model of the population, $\tilde{q}_n$ is the quantile estimate from a sample of size n and q is the true quantile under model G. It is well known that the MSE is the sum of the variance of $\tilde{q}_n$ and its squared bias.

Meanwhile, for convenience in displaying and interpreting the results, we

3

will use the root mean squared error (RMSE), which, intuitively speaking, shows the error of the quantile estimate on the same scale as the quantile. For the most part in our study, the MSE or RMSE has to be estimated by Monte Carlo simulation, where a large number N, e.g., 10000 of replicates will be simulated from the model G and then each of them will yield a quantile estimate $\tilde{q}_n^i, i = 1, 2, \ldots, N$, where the index of the replicates is in the super-script to avoid confusion with the sample size n in the subscript. Let $d_i = (\tilde{q}_n^i - q)^2$. Then the MSE and RMSE of the quantile estimate $\tilde{q}$ under model G will be estimated by

$$
\begin{aligned}
\widehat{\text{MSE}} =& \bar{d} = \frac{1}{N} \sum_{i=1}^{N} (\tilde{q}_n^i - q)^2 \\
\widehat{\text{RMSE}} =& \sqrt{\widehat{\text{MSE}}} = \sqrt{\bar{d}}.
\end{aligned}
$$

Moreover, for a better understanding of the MSE or RMSE, we will also study the bias ($E_G \tilde{q}_n - q$, estimated by $\frac{1}{N} \sum_{i=1}^{N} \tilde{q}_n^i - q$ in the simulations) and standard deviation (equivalently the variance) of the quantile estimates, to show the accuracy and efficiency of the quantile estimate. But the MSE and RMSE are the most important measures of the goodness of the quantile estimates.

1.2.1 The precision of $\widehat{\text{RMSE}}$

As the RMSE is estimated by the Monte Carlo procedure, the precision of the $\widehat{\text{RMSE}}$ will depend on the number of replicates N. The larger N is, the better $\widehat{\text{RMSE}}$ estimates the RMSE. In order to account for the Monte Carlo error of our RMSE estimate, we will also consider the standard error of $\widehat{\text{RMSE}}$. Here is a brief derivation of how to calculate the Monte Carlo standard error of $\widehat{\text{RMSE}}$:

The variance of $\widehat{\text{MSE}}$ can be estimated by

$$
\hat{\sigma}^2_{\widehat{\text{MSE}}} = Var(\bar{d}) = \frac{1}{N(N-1)} \sum_{i=1}^{N} (d_i - \bar{d})^2.
$$

By central limit theorem (Casella and Berger, 2002), the $\widehat{\mathrm{MSE}} = \bar{d}$ will asymptotically follow a normal distribution. Then the variance of $\widehat{\mathrm{RMSE}}$ can be estimated by Delta method (Casella and Berger, 2002) as,

$$\hat{\sigma}^2_{\widehat{\mathrm{RMSE}}} = \left(\frac{1}{2\sqrt{\bar{d}}}\right)^2 Var(\bar{d}).$$

In this way, the Monte Carlo standard error of our RMSE estimate can be estimated by

$$\hat{\sigma}_{\widehat{\mathrm{RMSE}}} = \frac{1}{2\sqrt{\bar{d}}}\sqrt{Var(\bar{d})}.$$

1.3 Overview

For quantile estimation, people usually consider the non-parametric empirical quantile or some parametric quantile estimate. Those quantile estimates are easy to obtain but they have different drawbacks, which will be first reviewed in Chapter 2. To balance between the non-parametric and parametric quantile estimate, the current wood engineering industrial standard D5457 (ASTM, 2004a) applies a subjective censorship scheme to a complete data set and obtains the quantile estimate from the censored Weibull maximum likelihood estimate of the lower tail. In Chapter 3, we will first carry out a review of this approach for a statistical point of view, which has not been done despite the fact that the standard has been used for many years.

Based on our study, the censored Weibull MLE in the current industrial standard is superior to the non-parametric or parametric quantile estimates, but it only utilizes a small ad-hoc proportion of a complete data set, which inspires us to further improve it by using more data or choosing a better proportion (equivalent to selecting a better censoring threshold). In Chapter 4, the Weibull mixture and censored Weibull mixture are proposed to utilize more data for the lower quantile estimation. Chapter 5 describes a bootstrap procedure for choosing a better threshold for the censored Weibull MLE. The quantile estimate from the censored Weibull mixture and the bootstrap

censored Weibull MLE are shown to be better than the original censored Weibull MLE via our simulations.

Chapter 2

Non-parametric and Parametric Quantile Estimates

All kinds of quantile estimates can be viewed as an inverse of a distribution estimate. The distribution estimate can be obtained non-parametrically via the empirical distribution or kernel density estimate, which leads to the non-parametric quantile estimate. On the other hand, we can assume and fit a parametric distribution for the data and obtain a parametric quantile estimate. The parametric or non-parametric quantile estimates are easy to obtain but they both have some serious shortcomings.

2.1 Non-parametric quantile estimates

In this section, we will first introduce the empirical quantile and then the quantile obtained from the kernel density estimate.

2.1.1 Empirical distribution and empirical quantile

For an *i.i.d.* sample $X_1, X_2, \ldots, X_n$, the empirical distribution function $\hat{F}_n$ is defined as,

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\}, \qquad (2.1)$$

where the $\mathbf{1}(\cdot)$ is the indicator function. It is one when $X_i \leq x$ and it is zero when the event is false. According to the famous Glivenko–Cantelli theorem (Cantelli, 1933; Glivenko, 1933), the supremum $||\hat{F}_n - F||_\infty = \sup_x |\hat{F}_n(x) - F(x)|$ converges to 0 almost surely.

As this empirical distribution is discrete, several definitions of the empirical quantile are obtained by different definitions of the inverse or the linear interpolation of it. The function "quantile" in R provides nines ways to calculate the empirical quantile. For example, type I empirical quantile is based on the most commonly used definition of the inverse of the empirical distribution,

$$\hat{F}_n^{-1}(t) = \inf\{x : \hat{F}_n(x) \geq t\}. \qquad (2.2)$$

Under this definition, the pth quantile $\hat{q}$ for a sample of size n will be the $\lfloor np \rfloor$th order statistics $X_{(\lfloor np \rfloor)}$, where $\lfloor np \rfloor$ denotes the smallest integer no smaller than np. In some statistical software like SAS, the quantile is defined as the nearest even order statistic, which is Type III in R. Here the phrase "nearest even" is a little misleading. When we round np, the nearest integer to np is chosen. For example, for $np = 2.3$, $\hat{q} = X_{[np]} = X_{(2)}$ and for $np = 2.6$, $\hat{q} = X([np]) = X_{(3)}$. But when there is a tie we will take the even integer. For example, if $np = 1.5$, $\hat{q} = X_{(2)}$ but if $np = 2.5$, still $\hat{q} = X_{(2)}$. So the "nearest even" indeed means "nearest and then even".

The above definitions of empirical quantile based on the inverse of empirical distribution only involves one order statistic for the quantile estimate, which is often inefficient and biased according to our simulations (as shown later). People also consider linear interpolation between two order statis-

8

tics to achieve better empirical quantiles. Here let $p_i = Pr(X \leq X_{(i)})$, where $X_{(i)}$ is the i-th order statistic and we can then estimate p_i by as $\hat{p}_i = (i - a)/(n + b)$, where the a and b are some tuning parameters. The empirical quantile $\hat{q}(p)$ of probability p, where $p \in [\hat{p}_i, \hat{p}_{i+1}]$, interpolates between $X_{(i)}$ and $X_{(i+1)}$:

$$\frac{\hat{q}(p) - X_{(i)}}{p - \hat{p}_i} = \frac{X_{(i+1)} - X_{(i)}}{\hat{p}_{i+1} - \hat{p}_i}$$

$$\hat{q}(p) = X_{(i)} + (p - \hat{p}_i)\frac{X_{(i+1)} - X_{(i)}}{\hat{p}_{i+1} - \hat{p}_i}.$$

In this way, different choices of a, b in $\hat{p}_i = (i - a)/(n + b)$ will lead to different empirical quantiles. For example, $a = b = 0$ leads to interpolating the commonly defined empirical distribution (Type IV in R). It is also found that $a = b = 1/3$ is approximately unbiased for the median regardless of the true distribution (Type VIII in R). Moreover, $a = 3/8, b = 1/4$ is approximately unbiased (Type IX in R) for the quantiles of the normal distribution. Hyndman and Fan (1996) provide a detailed description of these nine definitions of empirical quantiles.

In order to choose the best definition of "empirical quantile" for our study, several simulations under varied models have been carried out to compare the nine types of empirical quantiles in R. Generally speaking, these nine types of empirical quantiles are very similar, but the types with interpolation between two order statistics are marginally better than those only based on only one order statistic. As an example, the box-whisker plot of the 5th empirical percentiles in a simulation under Weibull$(7, 7)$[1] with sample size of 300 and 2500 replicates is in Figure 2.1.

For expository, only the Type I, III, IV, VIII, IX introduced above are shown in Figure 2.1. It is easy to see that the last three types with linear interpolation (Type VIII, IX) are relatively less (median) biased. Based on the mean squared error of these different definitions of quantile estimate in

---

[1]This denotes a two-parameter Weibull distribution with shape parameter as 7 and scale parameter as 7. A detail introduction to the Weibull distribution will be provided in the next chapter.

Figure 2.1: Several definitions of empirical quantiles in data sets simulated from Weibull(7, 7)
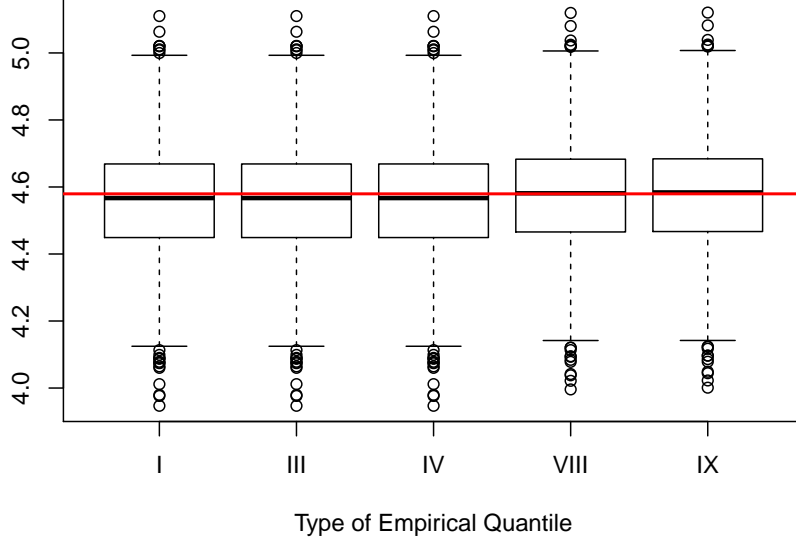
our simulations, we choose type IX as the default empirical quantile estimate for our study.

Moreover, regardless of the definition of empirical quantile in a finite sample, they are asymptotically equivalent. It is well known that the empirical p^{th} quantile $\hat{q}_n$ converges to the true quantile q almost surely. Moreover, according to Serfling (1980),

$$\sqrt{n}(\hat{q}_n - q) \to_d N(0, p(1-p)/f^2(q)), \tag{2.3}$$

where f is the PDF of the underlying model behind the data. For our study, we have few observations in the left tail, as shown in Figure 1.1 and thus $f(q)$ is very small. Thus, the variance of the empirical quantile will be very large.

2.1.2 Kernel density estimate and kernel quantile

In the above empirical quantiles, we only smooth between two order statistics, which only uses a small part of the information from the data. As illustrated later, the quantile estimate can be more efficient if we smooth over the entire sample. This leads us to apply some non-parametric density estimation techniques, such as, kernel density estimation, orthogonal series estimators, sieve estimators, etc. Please refer to Izenman (1991) for a throughout review on those techniques. In our study, we only consider the simplest and most popular kernel density estimate.

As in (2.1), the empirical distribution can be viewed as putting a mass of $1/n$ at each observation X_i. The kernel density will put a small "bump"—a kernel at each observation, which can be written as,

$$\tilde{k}_b(x) = \frac{1}{nb} \sum_{i=1}^{n} \kappa(\frac{x - X_i}{b}), \qquad (2.4)$$

where the $\kappa(\cdot)$ is the kernel function and b decides the kernel's width.

The kernel function $\kappa(\cdot)$ is a symmetric function that integrates to one. Some common kernels are the standard normal (Gaussian) density function, triangle, rectangular, uniform and Epanechnikov, etc. According to Izenman (1991), the Epanechnik kernel minimizes the mean integrated squared error (MISE, $E_F \int (\tilde{k}_b(x) - f(x))^2 dx$). However, we will only consider the normal kernel, which is the simplest to interpret for this study.

The kernel function decides the shape of the "bump", while the bandwidth b decides the width of this "bump", which is also referred to also the bandwidth and smoothing parameter. The larger b is, the smoother the density estimate is. According to the summary in Jones et al. (1996), there are various ways to select the bandwidth, such as rule of thumb, cross-validation and using pilot estimation of derivatives. Based on the recommendation of Jones et al. (1996), we will adopt the"Solve-the-Equation" approach proposed in Sheather and Jones (1991). This approach is well implemented in R, but its mathematical details are complicated and not stated here for brevity.

11

This kernel density estimate has been intensively studied by varied authors, which is well summarized in Silverman (1986). It is well known that kernel density estimate is a strongly consistent estimate of the true density ($\tilde{k}_b(x) \to_{a.s.} f(x)$ for all x) if the bandwidth $b \to 0$ and $nb \to \infty$ when $n \to \infty$ (Izenman, 1991; Silverman, 1986). With the kernel density estimate, the kernel estimated CDF is $\tilde{K}_b(x) = \int_{-\infty}^{x} \tilde{k}_b(u) du$ and we can numerically invert $\tilde{K}_b(x)$ to obtain the quantile estimate $\bar{q}$. According to Azzalini (1981), $\bar{q}$ is also a consistent estimate of the true quantile and the asymptotic variance of $\sqrt{n}\bar{q}$ is $\frac{p(1-p)}{f^2(q)} - O(n^{-1/3})$ under certain regularity conditions. Recall that the asymptotic variance of $\sqrt{n}\hat{q}$ is $\frac{p(1-p)}{f^2(q)}$ as in (2.3), which indicates the empirical quantile and the quantile obtained from the kernel density estimate $\bar{q}$ have the same asymptotic variance. But by the $O(n^{-1/3})$ term in the asymptotic variance of $\bar{q}$, the variance of $\bar{q}$ can be smaller than the variance of $\hat{q}$ in a finite sample, which can be seen in our simulations in Chapter 3.

However, it is necessary to point out that this numerical inverse of the kernel density estimate $\bar{q}$ is not the one usually referred to as "kernel quantile estimate", which is developed with similar ideas but from a different perspective (Parzen, 1979). According to Lio and Padgett (1991) and Jones (1992), the kernel quantile estimate and the one inverted from kernel density estimate $\bar{q}$ are asymptotically equivalent. Thus, only $\bar{q}$ is compared to the other quantile estimates in our study.

2.2 Parametric quantile estimation

The quantile obtained from the kernel density estimate is more efficient than the empirical quantile in a finite sample, but this improvement is quite limited in practice. Due to the large variance of non-parametric lower quantile estimates, people sometimes prefer to find a parametric approximation to the true distribution and then obtain a more efficient parametric quantile estimate (Silverman, 1986).

However, as will be shown in Chapter 3, the parametric quantile estimate can be seriously biased when the model does not approximate the true pop-

ulation distribution. It is thus critical to choose a good parametric model for the data. However, there can be several models fitting the data equally well in practice. Taken the MOR2 data set as an example, we can fit the Weibull, log-normal, gamma and the minimum type Gumbel distribution to it, as illustrated in Figure 2.2.

Figure 2.2: Different parametric models for MOR2 data set
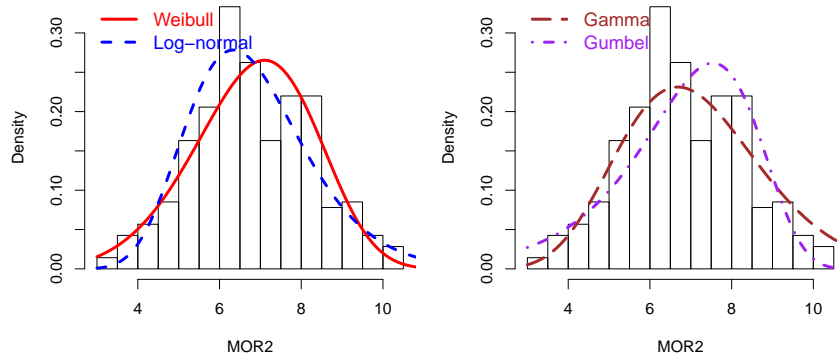
These four distributions are commonly used in modelling the strength of materials. The Weibull distribution will be introduced in Chapter 3 and a short introduction of the other three are provided in the appendix. For a more detailed introduction, please refer to Lawless (1982). The parameters of these distribution are fitted via MLE approach, which have been summarized in Table 2.1.

From Figure 2.2, it is easy to find that these four models are very different but it is hard to visually tell which one fits the MOR2 best. Therefore, we will compare the goodness-of-fit achieved by these four distributions by the Kolmogorov–Smirnov statistic:

$$D(\tilde{F}_n) = \sup_x \{|\tilde{F}_n(x) - \hat{F}_n(x)|\},$$

where the $\tilde{F}_n$ is the parametric density estimate and $\hat{F}_n(x)$ is the empirical distribution. It measures the maximum absolute discrepancy between the fitted distribution and the empirical distribution. The $D(\tilde{F}_n)$ of these four

models are also included in Table 2.1.

Table 2.1: Parametric models for MOR2, Goodness-of-fit and corresponding quantile estimates with

Model	Parameters		$D(\tilde{F}_n)$	$\tilde{q}(0.05)$
Weibull	$\alpha = 5.23$	$\eta = 7.39$	0.07	4.19
Log-normal	$\mu = 1.89$	$\sigma = 0.22$	0.05	4.63
Gamma	$k = 21.46$	$s = 0.32$	0.05	4.59
Minimum Gumbel	$a = 7.54$	$b = 1.41$	0.10	3.36

From Table 2.1, it is easy to see that the goodness-of-fit achieved by these four models are very close and the goodness-of-fit cannot help us decide the best model for our data. To make the situation worse, the quantile estimates are very different, as we learn from the last column of Table 2.1. Clearly, the task of choosing the parametric model will be challenging and limits the application of parametric quantile estimate in practice. Moreover, if we specify the wrong model, the parametric quantile estimate can be very biased, which will be shown in the next chapter.

As a summary of this chapter, the non-parametric quantile is unbiased but very inefficient. Although the parametric quantile estimate can be fully efficient, we cannot take the risk of model misspecification and using a seriously biased quantile. In this way, we need to find a quantile estimate balanced between the non-parametric and parametric quantile.

Chapter 3

Censored Weibull MLE, A Semi-parametric Approach

For the lower quantile estimation, the wood engineering industry does not apply the non-parametric empirical quantile estimates nor the parametric quantile estimate due to their different shortcomings, as discussed in the previous chapter. They find it unnecessary to have an accurate density estimate for the entire distribution and the lower quantiles can be sufficiently determined by the density below the true quantile. Therefore, it is sufficient to only specify a parametric distribution for the lower tail, which leads to a semi-parametric model:

$$g(x; \boldsymbol{\theta}, X_0) = \begin{cases} f(x; \boldsymbol{\theta}), x \leq X_0 \\ h(x), x > X_0 \end{cases}. \tag{3.1}$$

Here $f(x; \boldsymbol{\theta})$ is the density for the parametric lower tail, for example, a Weibull distribution with $\boldsymbol{\theta} = (\alpha, \eta)$. X_0 is the change point where the parametric density ends the non-parametric part starts. The non-parametric part $h(x)$ is an unspecified density except for the following two technical requirements, which ensures that $g(x)$ is a well defined density:

1. $h(x) \geq 0, x > X_0$

2. $\int_{X_0}^{\infty} h(x)dx = 1 - G(X_0) = 1 - F(X_0, \boldsymbol{\theta})$.

Under this semi-parametirc model, the quantile estimate can be obtained only from the parameter estimates of the left tail when the lower quantile q of interest is smaller than the change point X_0. Thus, it is not necessary to estimate X_0 nor the non-parametric part of this semi-parametric model if we are only interested in estimating the lower quantile.

In order to focus on the estimation the parametric left tail, we will apply subjective censoring in our parameter estimate, which will be introduced in Section 3.1. In Section 3.2, we will support our choice of the Weibull distribution as the parametric left tail from both engineering and statistical point of view. The computation of the censored Weibull MLE will be introduced in Section 3.3. The simulation comparison of the censored Weibull MLE and the non-parametric or parametric quantile estimate will be presented in Section 3.4. To further understand the advantage of the censored Weibull MLE, we will compare it to the ordinary MLE in Section 3.5 and discuss the goodness-of-fit in the left tail in Section 3.6.

3.1 Subjective censoring

Before the subjective censoring and its consistency are discussed, we will begin with a short introduction of censoring.

3.1.1 Introduction to censoring

In statistics, engineering and other fields, censoring denotes that the measurement of an observation is not complete known, but this partially observed observation will still be kept in the analysis. For example, a wooden board survives under a certain stress level will not be further stressed to broken but the information that it has survived under this stress level can still be used in a statistical analysis.

According to Lawless (1982, 2003) and Rinne (2009), censoring can be classified into informative and non-informative based on whether the censoring depends on the experimental outcome. Generally speaking, a non-informative censoring has a pre-determined censoring rule. For example, it can be decided before the experiment that after a specimen has survived

under a 1000 pound stress load, it will not be further stressed until broken. Based on this censoring rule, non-informative censoring can be further classified into Type I and II censoring: In Type I censoring, the board's strength is only known exactly when it is smaller than a predetermined strength value, as in the previous example; In Type II censoring, only the strength of r (a predetermined number) weakest lumber boards in a sample of n boards are completely known.

It is better to explain Type II censoring with the example in our lumber strength testing. We start to increase the stress from zero on all n lumber boards in our experiment simultaneously with the same rate of stress increase. Instead of increasing the stress until all the boards are broken, we will stop the experiment when r out n boards broke. Also, for the convenience in recording the data, the strength (the stress level at which it breaks) of the first broken board will be labelled X_1, the second broken one X_2, and so on. In this way, the strengths of r broken boards are completely known while the strengths of the $n - r$ unbroken boards are known only to be larger than X_r. It is necessary to point out that this $X_1, X_2, \ldots, X_r$ are actually the smallest r order statistics when the strength of the n boards are fully observed.

3.1.2 Definition and properties of subjective censoring

However, the censoring we will apply does not fall into those categories directly. The strength of every board in our sample is known and thus a complete data set without any censoring is available. But we want to focus on the information in just the left tail by censoring the rest of the data in the estimation of the parametric left tail in (3.1). Observations larger than a censoring threshold C will be censored and the information they contribute to the estimation process will be only that $X_i > C$. In the current industrial standard (ASTM, 2004a), the censoring threshold C is chosen to be the 10th empirical percentile (Type III definition, nearest even order statistics) in the estimation of the 5th percentile. In this way, 90% of the available data are censored to focus on the left tail. Similar ideas are used in Hill (1975) where

the left part of the data is subjectively censored to conduct extreme value inference.

In this way, this subjective censoring can be either viewed as a Type I censoring with the threshold chosen to be the sample 10th percentile or a Type II censoring when only 10% of the specimens fail in the experiment. If the subjective censoring is treated as a Type I censoring, its likelihood is:

$$L^I(\mathbf{X}; \boldsymbol{\theta}) = \prod_{i=1}^{n} [f(X_i; \boldsymbol{\theta})]^{\delta_i} [1 - F(C; \boldsymbol{\theta})]^{1-\delta_i}, \qquad (3.2)$$

where $\delta_i = \mathbf{1}\{X_i \leq C\}$ indicates whether the ith unit is censored or not. Here we also assume $C < X_0$ and thus the change point X_0 and the non-parametric part of model (3.1) will be subjectively censored.

If the subjective censoring is treated as a type II censoring, according to (Lawless, 1982), the likelihood is

$$L^{II}(\mathbf{X}; \boldsymbol{\theta}) = \frac{n!}{(n-r)!} \prod_{i=1}^{r} f(X_{(i)}; \boldsymbol{\theta})[1 - F(X_{(r)}; \boldsymbol{\theta})]^{n-r}$$

where $X_{(i)}, i = 1, 2, \ldots, n$ are the order statistics of sample $X_1, X_2, \ldots, X_n$.

When $C = X_{(r)}$, it is easily to verify the likelihoods $L^I(\mathbf{X}; \boldsymbol{\theta})$ and $L^{II}(\mathbf{X}; \boldsymbol{\theta})$ only differs in a constant $\frac{n!}{(n-r)!}$, which indicates the $\tilde{\boldsymbol{\theta}}$ maximizing $L^I(\mathbf{X}; \boldsymbol{\theta})$ will also be the one maximizing $L^{II}(\mathbf{X}; \boldsymbol{\theta})$. In this way, whether we treat the subjective censoring in the industrial standard D5457 (ASTM, 2004a) as a type I or type II does not affect the parameter estimates and thus does not affect the quantile estimate. For ease of tracking asymptotic properties of the censored MLE, we will work with the likelihood of type I censoring in (3.2).

Meanwhile, in order to distinguish the subjective censoring from the ordinary Type I censoring, we will name the above approach of maximizing the subjectively censored likelihood as the censored MLE with p^{th} empirical percentile as the threshold. It is necessary to point out that the empirical quantile used as the threshold is defined as the nearest even order statistic (Type III definition), as in the industrial standard D5457 (ASTM, 2004a).

18

On the other hand, the empirical quantile we treat as a quantile estimate is the Type IX definition, which is better than the Type III definition according to our discussions in Chapter 2. We will distinguish these two types of empirical quantile when they are in use.

For simplicity, the censored Weibull MLE with 10th empirical percentile as the threshold will be denoted by CMLE. Also, in order to distinguish the MLE without censoring from the censored MLE, we will call the MLE without censoring the ordinary MLE (OMLE for short).

Clearly, maximizing the above censored likelihood is not the only way to estimate the parameter in the left tail. Other methods include least squares and conditional tail modelling. But based on Durrans et al. (1998), the censored MLE is the best one for lower quantile estimation and thus will be the only one considered in our study.

3.2   Choice of parametric lower tail

Besides the subjective censoring, another important ingredient of (3.1) is the choice of the parametric tail. The current industrial standard (ASTM, 2004a) chooses the two-parameter Weibull distribution (Weibull, 1951), whose PDF and CDF is:

$$F(x) = 1 - \exp\left(-\left(\frac{x}{\eta}\right)^{\alpha}\right), \qquad\qquad x \geq 0 \qquad (3.3)$$

$$f(x) = \frac{\alpha}{\eta}\left(\frac{x}{\eta}\right)^{\alpha-1}\exp\left(-\left(\frac{x}{\eta}\right)^{\alpha}\right), \qquad\qquad x \geq 0. \qquad (3.4)$$

As introduced in Lawless (1982) and Rinne (2009), the Weibull distribution has been a popular parametric model in survival analysis and varied engineering fields for more than half a century. The history and a very detailed introduction of the Weibull distribution can be found in Rinne (2009).

In wood engineering, the application of the Weibull distribution first appeared in Warren (1973) and a review of the history of the Weibull distribution in wood engineering can be found in Johnson et al. (2003). Specially for the lower quantile estimation of wood strength data, Durrans et al.

(1998) compared the censored MLE, tail modelling and least squares and confirmed that the censored Weibull MLE is the best method among them. In the 1990's, the American Society for Testing and Materials (now called ASTMInc. or ASTM for short) started to apply the Weibull distribution in their standards for testing materials on varied properties. Such standards include D5457 (ASTM, 2004a), D5055 (ASTM, 2004b), etc. For the lower quantile estimation, the latest relevant standard is D5457, which describes and standardizes the censored Weibull MLE with 10th empirical percentile as the threshold.

Besides the popularity of the Weibull distribution in wood engineering, some empirical evidence can also be found from our two real data sets to support the choice of the Weibull.

3.2.1 Empirical evidence from the real data sets

There are several popular distributions other than the Weibull in survival analysis, such as log-normal, gamma and minimum Gumbel. All these four models can be fitted to the lower tail of the two real data sets with the subjective censorship. Here the censoring threshold is the 10th empirical percentile (Type III definition) of the two data sets. The curves from the censored MLE of these four models are shown in Figure 3.1.

Figure 3.1: Censored parametric models for MOR1 and MOR2
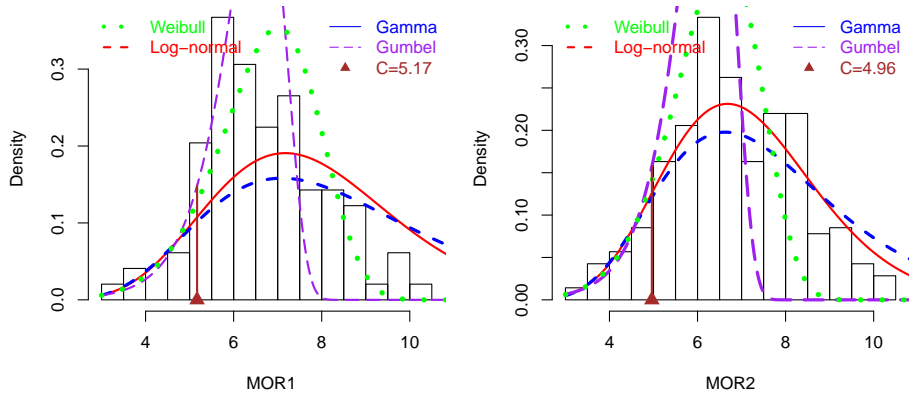
Although the curves of these four models are very different in the right tail, they actually are very similar in the left tail. To study the goodness of these parametric models in approximating the left tail of the real data set, we can use the following truncated Kolmogorov–Smirnov statistic, which measures the maximum absolute discrepancy between the fitted parametric tail $\tilde{F}_n$ and the empirical distribution $\hat{F}_n(x)$ of the real data set up the censoring threshold C:

$$D^{\ddagger}(\tilde{F}_n, \hat{F}_n(x)) = \sup_{x \leq C}\{|\tilde{F}_n(x) - \hat{F}_n(x)|\}. \tag{3.5}$$

The $D^{\ddagger}(\tilde{F}_n, \hat{F}_n)$ of the four models on the two real data sets are summarized in Table 3.1. It is easy to see that the $D^{\ddagger}(\tilde{F}_n, \hat{F}_n)$ of different parametric models on the same data set are very similar. They are all around 0.02 for MOR1 and 0.01 for MOR2, which indicates that these four models approximate the lower tail of the real data sets similarly well.

However, if we must choose a parametric model based on $D^{\ddagger}(\tilde{F}_n, \hat{F}_n)$, the Weibull can be the best one balanced between the two data sets: the minimum Gumbel achieves the best fit in the left tail of MOR1while it is the worst fit of MOR2 left tail; Log-normal and gamma are the worst fits in MOR1 but the best fits in MOR2; The fits of the Weibull are neither the best nor the worst, but it is close to the best fit both in MOR1 and MOR2, which supports our choice of the Weibull.

Table 3.1: Tail goodness-of-fit of the parametric models in the left tail of MOR1 and MOR2 and corresponding 5th percentile estimates

	$D^{\ddagger}(\tilde{F}_n, \hat{F}_n)$		Quantile Estimate [SD]	
Model	MOR1	MOR2	MOR1	MOR2
Weibull	0.021	0.011	4.51 [0.255]	4.64 [0.141]
Log-normal	0.023	0.009	4.47 [0.257]	4.57 [0.138]
Gamm	0.022	0.009	4.48 [0.231]	4.59 [0.138]
Minimum Gumbel	0.018	0.013	4.53 [0.242]	4.69 [0.140]

Moreover, based on the curves of these censored parametric models, it is not surprising to find that the difference of those quantile estimates from the

four different censored parametric models only lies in the third digit. This difference is not significant if we consider the standard deviation (SD) of the quantile estimate [1] and compare them with a Wald test (Casella and Berger, 2002). Therefore, unlike the parametric quantile estimate in Table 2.1, the choice of parametric tail will not remarkably influence quantile estimate under subjective censorship.

In this way, we will only consider the Weibull distribution as the parametric tail density for the semi-parametric model (3.1) in the sequel.

3.3 Computation of the censored Weibull MLE

As discussed above, we maximize (3.2) to achieve the parameter estimate of the lower tail and then obtain the quantile estimate. However, the intractability of (3.2) makes the direct maximization difficult. In particular, differentiating $L(\mathbf{X}; \alpha, \eta)$ with respect to the parameters is difficult. It is simpler to maximize the logarithm of it:

$$
\begin{aligned}
l(\mathbf{X}; \alpha, \zeta, C) = & m \ln(\alpha) + m \ln(\zeta) - \zeta \sum_{i=1}^{m} X_i^\alpha \\
& + (\alpha - 1) \sum_{i=1}^{m} \ln(X_i) - (n - m)\zeta C^\alpha,
\end{aligned} \tag{3.6}
$$

where we assume for $X_1, X_2, \ldots, X_n$ are $i.i.d.$ from a model G and $X_1, X_2, \ldots, X_m \leq C$ and $X_{m+1}, X_{m+2}, \ldots, X_n \geq C$. Also $\zeta = \eta^{-\alpha}$ is introduced to save computational effort. It is easy to see that if $C > \max\{X_i\}$ for all i (and then $m = n$), the (3.6) is the log-likelihood of the ordinary Weibull MLE.

Differentiating (3.6) with respect to α and ζ, we get,

$$
\begin{cases}
\dfrac{\partial l}{\partial \alpha} = \dfrac{m}{\alpha} - \zeta \sum_{i=1}^{m} X_i^\alpha \ln(X_i) + \sum_{i=1}^{m} \ln(X_i) - (n - m)\zeta C^\alpha \ln(C) \\
\dfrac{\partial l}{\partial \zeta} = \dfrac{m}{\zeta} - \sum_{i=1}^{m} X_i^\alpha - (n - m)C^\alpha
\end{cases} \tag{3.7}
$$

---

[1]The standard deviation here is estimated by bootstrap with 5000 replicates, whose detail will be introduced in Chapter 5.

According to (Rinne, 2009), the solution $\tilde{\alpha}, \tilde{\eta}$ to the above system of equations are the maximum likelihood estimates. Some basic algebra yields:

$$\frac{1}{\alpha} = \frac{\zeta}{m}[\sum_{i=1}^{m} X_i^\alpha \ln(X_i) + (n-m)C^\alpha \ln(C)] - \frac{\sum_{i=1}^{m} \ln(X_i)}{m} \qquad (3.8)$$

$$\frac{1}{\zeta} = \frac{\sum_{i=1}^{m} X_i^\alpha + (n-m)C^\alpha}{m}. \qquad (3.9)$$

The Newton-Raphson method can be applied here to solve (3.8) and (3.9) simultaneously and quickly, but it is prone to numeric instability. Here we can substitute ζ in (3.8) by (3.9):

$$\frac{1}{\alpha} = \frac{\sum_{i=1}^{m} X_i^\alpha \ln(X_i) + (n-m)C^\alpha \ln(C)}{\sum_{i=1}^{m} X_i^\alpha + (n-m)C^\alpha} - \frac{\sum_{i=1}^{m} \ln(X_i)}{m}. \qquad (3.10)$$

This iteration algorithm is quite insensitive to the initial value of α and guaranteed to converge under mild conditions (Rinne, 2009). Then, after solving $\tilde{\alpha}$, it is straightforward to obtain $\tilde{\zeta}$ based on (3.9)and the original scale parameter $\tilde{\eta} = \tilde{\zeta}^{-1/\alpha}$. With the $\tilde{\alpha}, \tilde{\eta}$ obtained from this censored Weibull MLE, the quantile estimate $\tilde{q}$ can be calculated as,

$$\tilde{q} = \tilde{\eta} \left[-\log(1-p) \right]^{1/\tilde{\alpha}}. \qquad (3.11)$$

3.4 Simulation comparison of the censored Weibull MLE and the parametric or non-parametric quantile estimates

In this section, we will present the simulations to compare the 5th percentile estimate obtained by the ordinary Weibull MLE (OMLE), censored Weibull MLE (CMLE), empirical quantile (EMP, Type IX definition) and the quantile estimate obtained from kernel density estimate (KDE) under varied models. Following the industrial standard, the censoring threshold C in the CMLE will be the 10th empirical percentile of each replicate.

Generally, we have two sets of simulations to imitate MOR1 and MOR2 data set respectively, which will be first introduced in this section. Under

these models, 10000 replicates of size 300 are simulated to evaluate these four quantile estimates. The RMSE of these four quantile estimate will be first compared and then we will further analyse the variance and bias of the CMLE to understand its advantages and limitations.

3.4.1 Model settings

Based on the complexity of the models from which we simulated the data, we will classify them into three categories, parametric models, mixture models and nonparametric models and introduce them separately.

Parametric models

The parametric models from which we simulated data are identical to those we used to obtain the censored MLE of the quantile estimates in Section 3.2 and Figure 3.1. Here their parameters are summarized in Table 3.2.

Table 3.2: Parameters for the censored parametric models to imitate the left tail of MOR1 and MOR2

Model	MOR1		MOR2	
Weibull	$\alpha = 6.822$	$\eta = 7.173$	$\alpha = 7.378$	$\eta = 6.738$
Log-normal	$\mu = 2.072$	$\sigma = 0.336$	$\mu = 1.976$	$\sigma = 0.2916$
Gamma	$k = 12.93$	$s = 0.601$	$k = 16.16$	$s = 0.4407$
Minimum Gumbel	$a = 6.620$	$b = 0.650$	$a = 6.315$	$b = 0.5997$

Mixture models

From the histogram of the MOR1 and MOR2 in Figure 1.1, there are two modes in each data set. Then, it is natural to consider our data as generated by a mixture of two uni-modal distributions with PDF:

$$f(x) = p_1 f_1(x) + (1-p)f_2(x), \tag{3.12}$$

where the f_1, f_2 are the density functions of the two sub-populations. They can be either from the same distribution families or from different distribu-

tion families. Here p indicates the proportion from the first sub-population, which also means that a random variable X from (3.12) has probability p of being from the first sub-population and probability $1 - p$ from the second sub-population. Of course, we can also consider mixture models with more than two sub-populations.

Our simulations suggest that we can consider mixture models with only two sub-populations and the sub-populations are from the same distribution family. The distributions considered here are normal, log-normal and Weibull. The parameters in this mixture model can be estimated by the EM algorithm, which will be introduced in the next chapter. The estimated parameters of the three mixture models are summarized in Table 3.3. Here we does not apply censoring in the EM algorithm for the mixture models, as they have more parameters and thus more flexibility to approximate the tail even without censoring.

Table 3.3: Mixture models to imitate MOR1 and MOR2

	Model	p	Majority Population		Minority Population	
	Normal	0.5629	$\mu_1 = 5.953$	$\sigma_1 = 0.970$	$\mu_2 = 7.676$	$\sigma_2 = 1.215$
MOR1	Log-normal	0.9758	$\mu_1 = 1.897$	$\sigma_1 = 0.189$	$\mu_2 = 1.245$	$\sigma_2 = 0.102$
	Weibull	0.7448	$\alpha_1 = 5.494$	$\eta_1 = 7.599$	$\alpha_2 = 15.81$	$\eta_2 = 5.983$
	Normal	0.5406	$\mu_1 = 5.924$	$\sigma_1 = 1.042$	$\mu_2 = 7.859$	$\sigma_2 = 1.095$
MOR2	Log-normal	0.6649	$\mu_1 = 1.976$	$\sigma_1 = 0.167$	$\mu_2 = 1.736$	$\sigma_2 = 0.226$
	Weibull	0.7932	$\alpha_1 = 5.427$	$\eta_1 = 7.642$	$\alpha_2 = 12.01$	$\eta_2 = 6.186$

Non-parametric models

As we have discussed above, we can never be certain about the validity of our model assumptions for the data set. So the model we use to simulate data might not represent the real data set (even if we consider only the tail). On the other hand, we can use the non-parametric kernel density estimate with Gaussian kernel in (2.4) to simulate data sets similar to our real data set, which can be viewed as a kind of smoothed bootstrap (Shao, 1993).

To simulate data from a kernel density estimate with Gaussian kernel and bandwidth b of a real data set, we will first sample with replacement from

Figure 3.2: Mixture models for MOR1 and MOR2
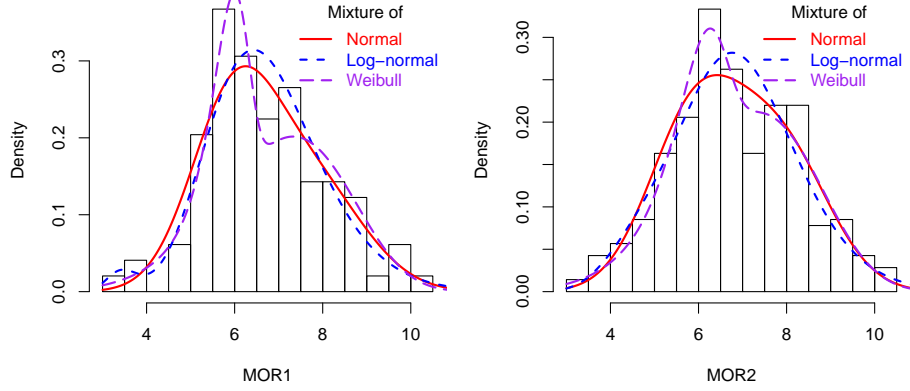
the original data set. Then *i.i.d.* random noise from the normal distribution with mean 0 and standard deviation b will be added to each observation in the re-sample of the real data set. In this way, we will obtain a simulated data set re-sampled from the original data set plus some noise.

As discussed in Section 3.1, we will use the "Solve-the-Equation" approach (Sheather and Jones, 1991) to select the bandwidth for the kernel density estimate. The bandwidth selected for MOR1 is 0.0419 while the bandwidth for MOR2 is 0.04056. The curves of these two kernel density estimates are shown in Figure 3.3.

3.4.2 RMSE of the quantile estimates

Although our list of models imitating the real data sets could be expanded, we believe the above 16 models are sufficient to evaluate these quantile estimates. Following the settings above (10000 replicates with size 300), these 16 simulations have been carried out and the RMSE of these four quantile estimates are summarized in Table 3.4 and 3.5 for the models imitating MOR1 and MOR2, respectively.

The Monte Carlo standard error of the RMSE estimates are smaller than 0.001 in these simulations and thus omitted in the tables. For readability and the comparison of results, we have highlighted the background of each

Figure 3.3: Kernel density models for MOR1 and MOR2
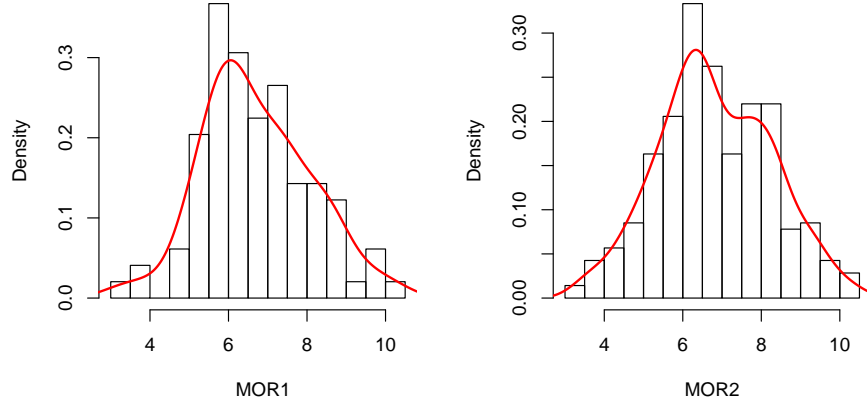
cell with different degree of gray. The lighter the background is, the smaller the root MSE is.

Table 3.4: RMSE of the quantile estimates from OMLE, CMLE, KDE, and EMP under the models imitating MOR1

Model	OMLE	CMLE	KDE	EMP
Weibull	0.112	0.150	0.171	0.173
Log-normal	1.009	0.169	0.300	0.184
Gamma	0.671	0.160	0.258	0.181
Minimum Gumbel	0.183	0.168	0.157	0.168
Normal Mixture	0.533	0.113	0.177	0.129
Log-normal Mixture	0.548	0.158	0.225	0.194
Weibull Mixture	0.506	0.162	0.178	0.185
MOR1 (KDE)	0.522	0.164	0.218	0.188

From these two tables, it is easy to see that the CMLE has the smallest RMSE among all four approaches in most models considered in our simulations. The exceptions are the Weibull and Gumbel model, where the OMLE or KDE does the best. But the disadvantage of CMLE under the Weibull or Gumbel model is negligible when it advantage under the other models are considered. For example, the RMSE of CMLE is around 1.4 times larger

Table 3.5: RMSE of the quantile estimates from OMLE, CMLE, KDE, and EMP under the models imitating MOR2

Model	OMLE	CMLE	KDE	EMP
Weibull	0.099	0.135	0.152	0.157
Log-normal	0.871	0.142	0.243	0.157
Gamma	0.609	0.139	0.217	0.157
Minimum Gumbel	0.165	0.153	0.145	0.155
Normal Mixture	0.370	0.125	0.191	0.143
Log-normal Mixture	0.375	0.139	0.172	0.152
Weibull Mixture	0.350	0.167	0.198	0.190
MOR2 (KDE)	0.284	0.149	0.189	0.177

than the RMSE of OMLE under the Weibull model, but OMLE has an RMSE 5 to 6 times larger than the RMSE of the CMLE under the Gamma or log-normal models.

3.4.3 Bias and standard deviation of the quantile estimates

To further understand the advantages and the limitations of CMLE, it is better to study the bias and standard deviation of these four quantile estimates, as shown in Table 3.6. Here a similar table can be obtained for the models imitating MOR1 but it is omitted as it shows the same trend as Table 3.6.

From Table 3.6, it can be learned that the CMLE does not have an extraordinarily large bias relative to the OMLE in the non-Weibull models nor a large standard deviation relative to the empirical quantile. Generally speaking, if these four quantile estimates are ranked by the absolute bias, the order is EMP < CMLE < KDE < OMLE. If they are ranked according to the standard deviation, the order is OMLE < CMLE < KDE < EMP. The censored Weibull MLE is neither the most accurate as the empirical quantile nor the most efficient as the OMLE, but it is the best one balanced between the accuracy and efficiency.

Moreover, the reasons that the CMLE has a larger RMSE than OMLE or KDE under the Weibull or Gumbel models can also be found in Table 3.6.

Table 3.6: Bias and [standard deviation] of the quantile estimates (×100) from OMLE, CMLE, KDE, and EMP in the models imitating MOR2

Model	OMLE	CMLE	KDE	EMP
Weibull	0.62 [9.90]	0.56 [13.5]	6.39 [13.8]	0.19 [15.7]
Log-normal	86.05 [13.7]	4.75 [13.4]	19.57 [14.4]	0.09 [15.7]
Gamma	59.48 [13.1]	3.61 [13.4]	16.39 [14.1]	0.04 [15.7]
Minimum Gumbel	13.88 [8.91]	4.16 [14.7]	3.47 [14.1]	0.64 [15.5]
Normal Mixture	35.21 [11.3]	2.42 [12.3]	14.16 [12.9]	0.07 [14.3]
Log-normal Mixture	35.41 [12.3]	5.45 [12.8]	10.82 [13.4]	0.42 [15.2]
Weibull Mixture	33.10 [11.5]	2.05 [16.5]	10.40 [16.9]	0.68 [19.0]
MOR2 (KDE)	25.97 [11.4]	4.59 [14.2]	10.92 [15.5]	0.26 [17.7]

Under the Weibull model, both the OMLE and CMLE are almost unbiased, but the OMLE is more efficient than the censored Weibull MLE as it uses all the data, which will be further discussed in the next section. The kernel density estimate usually approximates the left tail of the true distribution worse than the censored Weibull MLE and thus the quantile estimate from KDE is more biased than the CMLE in most models we considered. However, the KDE approximates the Gumbel model amazingly well, which will be illustrated in Section 3.6 and thus the CMLE loses its advantage in the accuracy under the Gumbel model.

3.5 Comparison of the censored Weibull MLE and ordinary Weibull MLE

The above simulation only serves as empirical evidence of the advantage of CMLE over the OMLE, but we need some theoretical work to understand how subjective censoring works. Here in this section, we will compare the score function of the CMLE and OMLE and then their asymptotic variance.

3.5.1 Score function

As both OMLE and CMLE solve a system of equations as in (3.7), we can further study their difference from their corresponding score functions

(Huber, 1981; Serfling, 1980). Briefly speaking, the system of equations as in (3.7) is the sum of the score functions evaluated at the observations X_i, $i = 1, 2, \ldots, n$. In this way, different X_i's will have different "weights" in this system of equations and thus different "weights" on the parameter estimates.

The score function for $\boldsymbol{\theta} = (\alpha, \eta)$ for the ordinary Weibull MLE is

$$\psi_\alpha^O(x) = -\left(\frac{x}{\eta}\right)^\alpha \log\left(\frac{x}{\eta}\right) + \log\left(\frac{x}{\eta}\right) + \frac{1}{\alpha} \tag{3.13}$$

$$\psi_\eta^O(x) = -\frac{\alpha}{\eta} + \frac{\alpha x^\alpha}{\eta^{\alpha+1}}; \tag{3.14}$$

The upper-script O indicates the ordinary MLE. Similarly, the score function for $\boldsymbol{\theta} = (\alpha, \eta)$ of the censored Weibull MLE is:

$$\psi_\alpha^C(x) = \begin{cases} -\left(\frac{x}{\eta}\right)^\alpha \log\left(\frac{x}{\eta}\right) + \log\left(\frac{x}{\eta}\right) + \frac{1}{\alpha}, & x \leq C \\ -\left(\frac{C}{\eta}\right)^\alpha \log\left(\frac{C}{\eta}\right), & x > C \end{cases}; \tag{3.15}$$

$$\psi_\eta^C(x) = \begin{cases} -\frac{\alpha}{\eta} + \frac{\alpha x^\alpha}{\eta^{\alpha+1}}, & x \leq C \\ \frac{\alpha C^\alpha}{\eta^{\alpha+1}}, & x > C. \end{cases} \tag{3.16}$$

It is easy to see that (3.15) and (3.16) are truncated versions of (3.13) and (3.14) respectively. The curves of $-\psi_\alpha^C(x)$, $-\psi_\alpha^O(x)$, $\psi_\eta^C(x)$ and $\psi_\eta^O(x)$ under Weibull(7, 7) are provided in Figure 3.4, where we plot the negative score functions for α to make it easier to read. The censoring threshold C for the censored Weibull MLE is chosen as the 10% theoretical quantile of Weibull(7, 7).

If we compare one observation in the left tail (e.g., $X_i < 4$) and an observation in the right tail (e.g., $X_{i'} > 9$), it is easy to see that $X_{i'}$ has a larger absolute score function than the X_i for the OMLE. As the OMLE of the Weibull parameters is the solution of the system of equations as in (3.7) (with $n = m$), the $X_{i'}$ will have a larger influence on the parameter estimate,

Figure 3.4: Score function $-\psi_\alpha^C(x)$, $-\psi_\alpha^O(x)$, $\psi_\eta^C(x)$ and $\psi_\eta^O(x)$ under Weibull(7, 7)
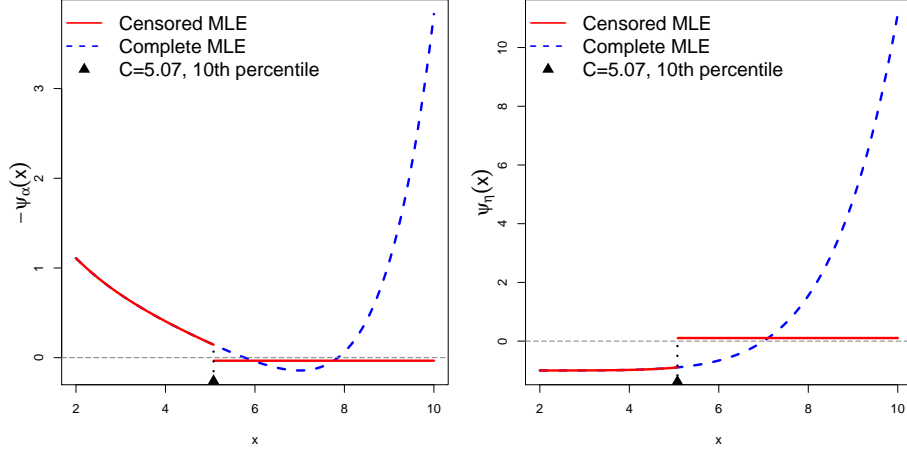
and thus the quantile estimate, than X_i. This contradicts our intuition that the lower quantile should be more influenced by the observations in the left tail rather than the right tail. This is one of the reasons that the ordinary Weibull MLE is not suitable for the lower quantile estimation.

For the CMLE, the score function is a constant when x is larger than censoring threshold and this constant (around -0.02 for α and 0.05 for η) is close to zero when compared to the score function value of the uncensored part. Therefore, the $X_{i'}$ in the right tail will have a smaller influence on the parameter estimate and equivalently, the lower quantile estimate than the X_i in the left tail. In this way, the lower quantile estimate will be mostly determined by the observations in the left tail.

3.5.2 Asymptotic Variance of censored Weibull MLE

Although censoring helps to avoid the undesired influence of some large observations on the lower quantile estimate, it sacrifices some of the efficiency of the ordinary MLE. As shown in Table 3.6, the CMLE always has a larger standard deviation than the OMLE. It is interesting to find this loss of efficiency theoretically via the asymptotic variance of the quantile estimates.

According to Lawless (1982) and Rinne (2009), the usual large-sample maximum likelihood properties are valid for the censored MLE when the shape parameter exceeds 2 and the number of uncensored observations approaches infinity at rate n as $n \to \infty$. For our data, the estimate of the shape parameter, censored or not, is larger than 4. So, it is reasonable to assume that the first condition above is satisfied. As censoring is subjective, the second condition can be also be satisfied. In this way, information matrix

$$I(\boldsymbol{\theta}) = -E_G\left(\frac{\partial \psi}{\partial \boldsymbol{\theta}}\right)$$

can be calculated by taking the expectation of the partial derivatives of the score function in (3.15) and (3.16). Here $\boldsymbol{\theta} = (\alpha, \eta)$ denotes the parameters of the Weibull tail. As the information matrix here cannot be written in a close form and the formulas are long, its detailed calculation is provided in the appendix.

Suppose we have derived the information matrix. Based on maximum likelihood theories,

$$\sqrt{n}\left[\begin{pmatrix} \tilde{\alpha}_n \\ \tilde{\eta}_n \end{pmatrix} - \begin{pmatrix} \alpha \\ \eta \end{pmatrix}\right] \to_d N\left(\mathbf{0}, I^{-1}(\boldsymbol{\theta})\right),$$

where

$$I(\boldsymbol{\theta}) = \begin{bmatrix} -E_G \dfrac{\partial \psi_\alpha(x)}{\partial \alpha} & -E_G \dfrac{\partial \psi_\alpha(x)}{\partial \eta} \\ -E_G \dfrac{\partial \psi_\alpha(x)}{\partial \eta} & -E_G \dfrac{\partial \psi_\eta(x)}{\partial \eta} \end{bmatrix}.$$

As in (3.11), the quantile estimate $\tilde{q}_n$ is a function of the parameter estimates and the true quantile is a function of the true parameters. The asymptotic variance of $\sqrt{n}(\tilde{q}_n - q)$ can be derived via Delta method (Casella and Berger, 2002). The derivative of the quantile q with respect to the

parameters $\boldsymbol{\theta} = (\alpha, \eta)$ is

$$\nabla_q(\boldsymbol{\theta}) = \begin{pmatrix} \dfrac{\partial q}{\partial \alpha} \\ \dfrac{\partial q}{\partial \eta} \end{pmatrix} = \begin{pmatrix} -\dfrac{\eta}{\alpha^2} \left[-\log(1-p) \right]^{\frac{1}{\alpha}} \log\left(-\log(1-p) \right) \\ \left[-\log(1-p) \right]^{\frac{1}{\alpha}} \end{pmatrix}.$$

Then, with the Delta method,

$$\sqrt{n}(\tilde{q}_n(\tilde{\boldsymbol{\theta}}) - q(\boldsymbol{\theta})) \to_d N\left(0, \nabla_q(\boldsymbol{\theta})^T I^{-1}(\boldsymbol{\theta}) \nabla_q(\boldsymbol{\theta}) \right). \qquad (3.17)$$

Moreover, Lawless (1982) derived the asymptotic variance of Type I censored log-Weibull MLE, where a log-Weibull random variable is the logarithm of a Weibull random variable. The formulas there are relatively easier to read but they still require numerical integration to calculate. We have verified that our results above are the same as the one converted from the results in Lawless (1982).

Figure 3.5: Asymptotic variance of $\sqrt{n}(\tilde{q}-q)$ and CDF under different threshold
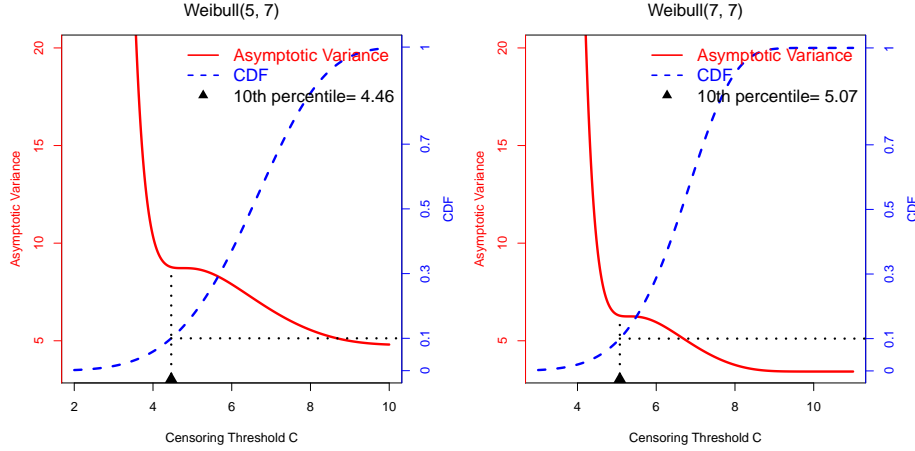
To illustrate the efficiency sacrificed in the censoring Weibull MLE, we plot the curve of the asymptotic variance of the 5th percentile estimate $\tilde{q}$ (the variance of $\sqrt{n}(\tilde{q} - q)$) at different censoring thresholds C together

33

with corresponding cumulative probability up to C in Figure 3.5. The true models are chosen to be Weibull$(5, 7)$ and Weibull$(7, 7)$. Similar patterns can be observed under Weibull distribution of different parameters. It is easy to see that the asymptotic variance decreases with the increase of censoring threshold, as more data points and information are taken into consideration.

Meanwhile, it is interesting that the asymptotic variance after the 10th percentile is relatively flat when compared to the curve before the 10th percentile. The asymptotic variance of the quantile estimate when we censor at 10th empirical percentile is around 2 times larger than that of the ordinary MLE. In this way, we only pay a small cost of efficiency to achieve an accurate estimate of the lower tail.

3.6 Goodness-of-fit in the left tail

In the above two sections, we have verified that the CMLE is almost unbiased for the 5th percentile estimate and the CMLE can avoid the influence of the observations in the upper tail with a small cost of efficiency, but the above conclusions are limited to the 5th percentile estimate. It is interesting to find out how the CMLE performs in the estimation of other lower quantiles, such as the 1st or 3rd percentile.

The efficiency of the quantile estimates is determined by the censoring threshold and we have confirmed that the censoring with 10th empirical percentile only causes a small efficiency loss. What remains to be verified is that the CMLE can still provide almost unbiased quantile estimates for the 1st or 3rd percentile as the 5th percentile estimate. Instead of carrying out intensive simulations to evaluate other quantile estimates, we will study the goodness-of-fit in the left tail achieved by the CMLE. If the CMLE fits the lower tail of other models well, the quantile estimate will not be very biased. In this section, we will first introduce the statistic we use to evaluate the tail goodness-of-fit and then discuss the corresponding simulation results.

34

3.6.1 Statistic to study the Goodness-of-fit

As in Section 3.2, we will measure the goodness-of-fit in the lower tail with the following truncated Kolmogorov–Smirnov statistic:

$$D^\dagger(\tilde{F}_n) = \sup_{x \leq C}\{|\tilde{F}_n(x) - G(x)|\}, \qquad (3.18)$$

where $\tilde{F}_n(\cdot)$ is a distribution estimate, e.g., the one from the censored Weibull MLE, and $G(\cdot)$ is the true population CDF (it does not have to be the Weibull). Here we choose to truncate the statistic at threshold C in order to focus on the goodness-of-fit in the left tail.

The distribution of $D^\dagger(\tilde{F}_n)$ under a model G can be estimated by Monte Carlo methods, but the distribution of $D^\dagger(\tilde{F}_n)$ is not informative: For example, if it is found that the distribution is nested in the interval of $(0.1, 0.3)$ when the sample size is 300, we cannot know if this 0.1 or 0.3 is large or small for a reasonably good approximation.

On the other hand, it is possible to find a reference non-parametric distribution estimate, such as the empirical distribution in (2.1) and the kernel density estimate in (2.4). Then we can compare the truncated Kolmogorov-Smirnov statistic $D^\dagger$ in the Weibull left tail against the empirical distribution or the kernel density estimate. Let $\hat{F}_n$ denote the empirical CDF and $\tilde{K}_b(x) = \int_{-\infty}^x \tilde{k}(u)du$ be the distribution estimate obtained from the kernel density estimate. For one data set, we will have the following statistics:

$$d(\tilde{F}_n, \hat{F}_n) = D^\dagger(\tilde{F}_n) - D^\dagger(\hat{F}_n) \qquad (3.19)$$

$$d(\tilde{F}_n, \tilde{K}_n) = D^\dagger(\tilde{F}_n) - D^\dagger(\tilde{K}_b). \qquad (3.20)$$

We can use the Monte Carlo method to estimate the distribution of $d(\tilde{F}_n, \hat{F}_n)$ and $d(\tilde{F}_n, \tilde{K}_n)$ and then compare $D^\dagger(\tilde{F}_n)$ to $D^\dagger(\hat{F}_n)$ or $D^\dagger(\tilde{K}_b)$. As $d(\tilde{F}_n, \hat{F}_n)$ or $d(\tilde{F}_n, \tilde{K}_n)$ is a random variable, we are mainly interested in $Pr\{d(\tilde{F}_n, \hat{F}_n) < 0\}$ and $Pr\{d(\tilde{F}_n, \tilde{K}_n) < 0\}$.

35

3.6.2 Simulation results

The simulations to obtain the Monte Carlo estimate of the distribution of $d(\tilde{F}_n, \hat{F}_n)$ and $d(\tilde{F}_n, \tilde{K}_n)$ are carried out in the same setting as in Section 3.4. The results for the models imitating MOR1 and MOR2 are very similar, so here we will only display the results from the models imitating MOR2. The box-whisker plots of $d(\tilde{F}_n, \hat{F}_n)$ and $d(\tilde{F}_n, \tilde{K}_n)$ are included in Figure 3.6.

Figure 3.6: Box-whisker plot of $d(\tilde{F}_n, \hat{F}_n)$ and $d(\tilde{F}_n, \tilde{K}_n)$ in the models imitating MOR2
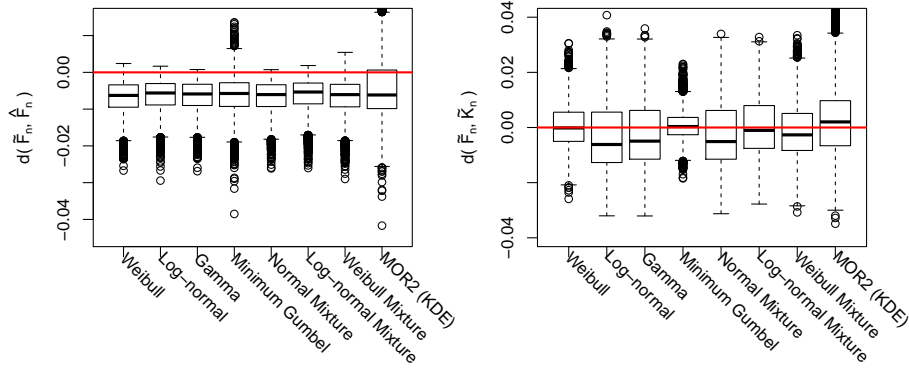
As shown in Figure 3.6, the medians of $d(\tilde{F}_n, \hat{F}_n)$ and $d(\tilde{F}_n, \tilde{K}_n)$ are smaller than zero in most of the models considered in our simulations. Table 3.7 summarizes the Monte Carlo estimate of $Pr\{d(\tilde{F}_n, \hat{F}_n) < 0\}$ $(Pr\{d(\tilde{F}_n, \tilde{K}_n) < 0\})$. As the number of replicates is 10000, the Monte Carlo standard deviation of these estimate are only in the second digits (maximum 0.05) and thus omitted.

From Table 3.7, it is easy to see that $Pr\{d(\tilde{F}_n, \hat{F}_n) < 0\} > 70\%$ in all the models considered in our simulations and except for the kernel density model imitating MOR2, $Pr\{d(\tilde{F}_n, \hat{F}_n) < 0\} > 90\%$. Therefore, we may conclude that the left tail achieved by the CMLE approximates the true distribution better than the empirical distribution most of the time for a data set similar to the MOR2.

The advantage of the CMLE over the kernel density estimate is not as remarkable as the advantage of the CMLE over the empirical distribution,

Table 3.7: $Pr\{d(\tilde{F}_n, \hat{F}_n) < 0\}$ ($Pr\{d(\tilde{F}_n, \tilde{K}_n) < 0\}$ under the models imitating MOR2

	$Pr\{d(\tilde{F}_n, \hat{F}_n) < 0\}$	$Pr\{d(\tilde{F}_n, \tilde{K}_n) < 0\}$
Weibull	0.98	0.51
Log-normal	0.95	0.66
Gamma	0.96	0.64
Minimum Gumbel	0.95	0.47
Normal Mixture	0.97	0.64
Log-normal Mixture	0.94	0.54
Weibull Mixture	0.98	0.60
MOR2 (KDE)	0.74	0.44

but $Pr\{d(\tilde{F}_n, \tilde{K}_n) < 0\}$ is still larger than 50% in most of the models considered in our simulations, which indicates that the censored Weibull left tail is better than the kernel density estimate in the sense that it is more likely to be closer to the true distribution in the left tail. The two exceptions are the minimum Gumbel and kernel density model. It is not surprising that the CMLE cannot beat the kernel density model when the data is simulated from a kernel density model. For the minimum Gumbel model, the goodness-of-fit achieved by the kernel density estimate helps to explain why the quantile estimate from KDE is less biased than the quantile estimate from CMLE as in Table 3.6.

Moreover, besides $Pr\{d(\tilde{F}_n, \hat{F}_n) < 0\}$ or $Pr\{d(\tilde{F}_n, \tilde{K}_n) < 0\}$, we can also study the mean of $d(\tilde{F}_n, \hat{F}_n)$ or $d(\tilde{F}_n, \tilde{K}_n)$, which is smaller than zero in most of our models, which indicates that the left tail estimate by CMLE is better than the empirical distribution and KDE on average. For the simplicity of writing, those results are omitted here.

In this way, we have verified that the left tail estimate by the CMLE can be as good as the non-parametric distribution estimate such as the empirical distribution or the kernel density estimate. Thus the CMLE can provide almost unbiased quantile estimate besides the 5th percentile.

As a summary of this chapter, we have studied the censored Weibull MLE from a statistical point of view and confirmed that it is a good lower

quantile estimate from varied aspects. However, the censored Weibull MLE only uses an ad-hoc and small proportion of the data, which inspires us to further improve it.

Chapter 4

Weibull Mixture and Censored Weibull Mixture

As discussed in the previous chapter, the censored Weibull MLE approach censors some information in the data and thus is not fully efficient. In order to fully utilize the data set and keep the accuracy of the quantile estimate, we can consider fitting a more complex model, for example, a mixture model, to the data set. In this chapter, we mainly consider a mixture of two Weibull distributions, which will be estimated by the EM algorithm. The definition of the mixture model and the EM algorithm are introduced in Section 4.1. To compare the Weibull mixture to a single Weibull distribution, the bootstrap homogeneity test is introduced and discussed in Section 4.2. Moreover, to ensure the accuracy of quantile estimates obtained from the Weibull mixture, subjective censoring can still be applied to the Weibull mixture and this is discussed in Section 4.3. Finally, a simulation comparison of the mixture approaches against the censored Weibull MLE and empirical quantile is in Section 4.4.

4.1 Definition and estimation of a Weibull mixture

The general definition of mixture models has been introduced in Section 3.4. As both MOR1 and MOR2 have two modes, it is reasonable to assume our data are from a mixture of two uni-modal distributions. Following the fairly standard practice of using the Weibull distribution, we will only consider a mixture of two two-parameter Weibull distributions with the following PDF:

$$f(x; p, \alpha_1, \eta_1, \alpha_2, \eta_2) = pf(x; \alpha_1, \eta_1) + (1-p)f(x; \alpha_2, \eta_2), \tag{4.1}$$

$$= p\frac{\alpha_1}{\eta_1}\left(\frac{x}{\eta_1}\right)^{\alpha_1 - 1} e^{-\left(\frac{x}{\eta_1}\right)^{\alpha_1}} + (1-p)\frac{\alpha_2}{\eta_2}\left(\frac{x}{\eta_2}\right)^{\alpha_2 - 1} e^{-\left(\frac{x}{\eta_2}\right)^{\alpha_2}}$$

for $x > 0$. In this way, the random variable from this distribution has a probability p to be from a Weibull (α_1, η_1) and probability $1 - p$ from Weibull (α_2, η_2), but the label of the sub-population is not observed.

4.1.1 EM estimation of the Weibull mixture

To estimate the parameters $\boldsymbol{\theta} = (p, \alpha_1, \eta_1, \alpha_2, \eta_2)$ in (4.1), we can still try to maximize the following log-likelihood:

$$l^I(\boldsymbol{X}; \boldsymbol{\theta}) = \sum_{i=1}^n \log\left(pf(X_i; \alpha_1, \eta_1) + (1-p)f(X_i; \alpha_2, \eta_2)\right). \tag{4.2}$$

However, the direct maximization of the above log-likelihood is numerically difficult and unstable. For example, the partial derivative of (4.2) against α_1 will involve α_2, which can be complicated to deal with. The problem would be much easier to solve if we knew which sub-population X_i is from, for example, if $Z_i = \mathbf{1}\{X_i \text{ is from Weibull}(\alpha_1, \eta_1)\}$ is also observed. In this case, the PDF can be written as

$$f(x, z; \boldsymbol{\theta}) = [pf(x; \alpha_1, \eta_1)]^z[(1-p)f(x; \alpha_2, \eta_2)]^{1-z}.$$

Correspondingly, the $\boldsymbol{\theta}$ can be estimated by maximizing the following log-likelihood:

$$l^C(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) = \sum_{i=1}^{n} Z_i[\log(pf(X_i; \alpha_1, \eta_1))] +$$

$$(1 - Z_i)[\log((1 - p) + f(X_i; \alpha_2, \eta_2))]. \qquad (4.3)$$

To distinguish the log-likelihood in (4.3) from that in (4.2), (4.3) is called the "complete" log-likelihood ("C" in the upper-script) while (4.2) is the incomplete log-likelihood ("I" in the upper-script).

The EM algorithm (Dempster et al., 1977) is designed to carry out the maximum likelihood estimation in (4.3). The basic idea of EM algorithm is the following: Although Z_i is not observed in practice, we can still calculate its expectation (the "E" step) conditioning on the current parameter estimates and then maximizing (the "M" step) the parameters conditioning on the expectation of Z_i; the E step and M step will be alternated until the parameter estimates and likelihood converge. In this way, the complex maximization problem can be accomplished by a series of smaller and easier maximization problem. Under certain regularity conditions, the EM algorithm provides consistent parameter estimates (Wu, 1983). For a more detailed description of the EM algorithm, refer to Dempster et al. (1977).

For our mixture of Weibulls, the expectation of Z_i conditioned on the current parameter $\boldsymbol{\theta}^{(s)}$ can be easily calculated as

$$\tilde{Z}_i^{(s)} = E(Z_i | X_i, \boldsymbol{\theta}^{(s)}) = P(Z_i = 1 | X_i, \boldsymbol{\theta}^{(s)})$$

$$= \frac{p^{(s)} f_1(X_i; \alpha_1^{(s)}, \eta_1^{(s)})}{p^{(s)} f_1(X_i; \alpha_1^{(s)}, \eta_1^{(s)}) + (1 - p^{(s)}) f_2(X_i; \alpha_2^{(s)}, \eta_2^{(s)})}. \qquad (4.4)$$

Plugging $\tilde{Z}_i^{(s)}$ into (4.3), we have:

$$\tilde{l}^C(\mathbf{X}, \tilde{\mathbf{Z}}^{(s)}; \boldsymbol{\theta}) = \sum_{i=1}^{n} \tilde{Z}_i^{(s)}[\log(p) + \log(f(X_i; \alpha_1, \eta_1))]$$

$$+ (1 - \tilde{Z}_i^{(s)})[\log(1 - p) \log(f(X_i; \alpha_2, \eta_2))]. \qquad (4.5)$$

After the above "E" step, we will carry on to the "M" step and find the parameters maximizing (4.5). It is easy to find that the $\tilde{p}^{(s+1)}$ maximizing (4.5) is

$$\tilde{p}^{(s+1)} = \frac{1}{n} \sum_{i=1}^{n} \tilde{Z}_i^{(s)},$$

by differentiating (4.5) against p.

Similarly, by differentiating (4.5) with respect to other parameters, we can obtain the equations of $\alpha_1, \eta_1, \alpha_2, \eta_2$ whose solution will maximize (4.5). With the help of $\tilde{Z}_i$, the equation to solve for $\tilde{\alpha}_1^{(s+1)}, \tilde{\eta}_1^{(s+1)}$ will not involve α_2, η_2, that is, we can solve the following iteration function to obtain $\tilde{\alpha}_1^{(s+1)}$:

$$\frac{1}{\alpha} = \frac{\sum_{i=1}^{n} \tilde{Z}_i^{(s)} X_i^{\alpha} \ln(X_i)}{\sum_{i=1}^{n} \tilde{Z}_i^{(s)} X_i^{\alpha}} - \frac{\sum_{i=1}^{n} \tilde{Z}_i^{(s)} \ln(X_i)}{\sum_{i=1}^{n} \tilde{Z}_i^{(s)}}, \tag{4.6}$$

and then obtain $\tilde{\eta}_1^{(s+1)}$ by

$$\eta_1^{(s+1)} = \left(\frac{\sum_{i=1}^{n} \tilde{Z}_i^{(s)} X_i^{\tilde{\alpha}_1^{(s+1)}}}{\sum_{i=1}^{n} \tilde{Z}_i^{(s)}} \right)^{1/\tilde{\alpha}_1^{(s+1)}}. \tag{4.7}$$

Next, we can replace $\tilde{Z}_i^{(s)}$ by $1 - \tilde{Z}_i^{(s)}$ in (4.6) and (4.7) to achieve the formulas for $\tilde{\alpha}_2^{(s+1)}, \tilde{\eta}_2^{(s+1)}$. These "E" and "M" steps will be alternated until convergence of the parameter estimates or the likelihood.

After obtaining $(\tilde{p}, \tilde{\alpha}_1, \tilde{\eta}_1, \tilde{\alpha}_2, \tilde{\eta}_2)$, we can get the quantile estimate by equating the CDF of Weibull mixture

$$
\begin{aligned}
F(x; p, \alpha_1, \eta_1, \alpha_2, \eta_2) &= pF(x; \alpha_1, \eta_1) + (1-p)F(x; \alpha_2, \eta_2) \tag{4.8}\\
&= 1 - pe^{-\left(\frac{x}{\eta_1}\right)^{\alpha_1}} - (1-p)e^{-\left(\frac{x}{\eta_2}\right)^{\alpha_2}},
\end{aligned}
$$

and the percentage of the quantile of interest, e.g. 5th percentile, and then solve the resulting equation numerically with bi-section method.

4.1.2 The issue of non-finite likelihood of Weibull mixture

The above EM algorithm is aimed at finding the parameters that maximize the log-likelihood (4.2). However, (4.2) diverges to infinity, e.g., when $\eta_1 = X_1$ and $\alpha_1 \to \infty$. So, the real maximum of the likelihood in (4.2) is not well defined. This is also the main issue in the failure of convergence of the EM algorithm as sometimes the estimates of α_1 or α_2 diverge to infinity very quickly.

This is also a problem for the mixture of normal distributions with both means and standard deviations to be estimated for both sub-populations. Several solutions for this issue have been proposed: Hathaway (1985) put additional constraints on the parameter space, e.g., limiting the space of σ in the normal mixture case; Chen et al. (2008) added a penalty function on σ. For our Weibull mixture, we will simply constrain the space of α_1, α_2 to $[0, 30]$ and terminate the EM algorithm when α_1 or α_2 violates this constrain.

4.1.3 Weibull mixture estimates for real data sets

For our two real data sets, the EM algorithm successfully converges and the parameter estimates have been summarized in Table 3.3 and Table 4.1. The curves of the densities of the mixture models together with the densities of the sub-populations are summarized in Figure 4.1.

Figure 4.1: Weibull mixture model for MOR1 and MOR2

43
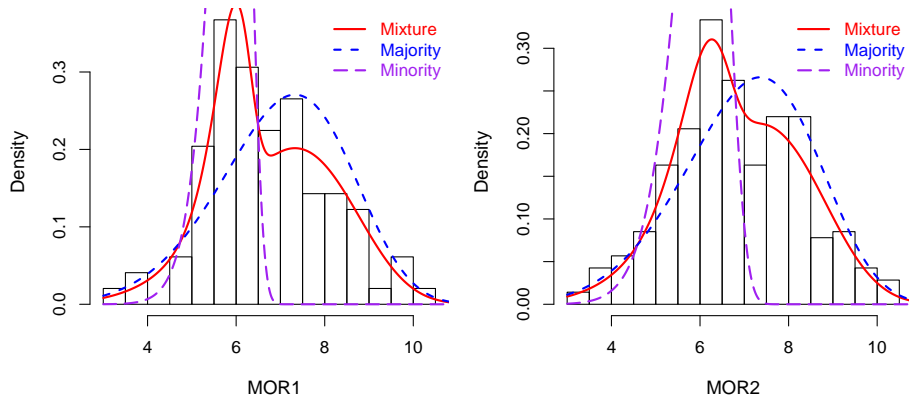
Table 4.1: Weibull mixture models for MOR1 and MOR2

Data	p	Majority Population		Minority Population	
MOR1	0.7448	$\alpha_1 = 5.494$	$\eta_1 = 7.599$	$\alpha_2 = 15.81$	$\eta_2 = 5.983$
MOR2	0.7932	$\alpha_1 = 5.427$	$\eta_1 = 7.642$	$\alpha_2 = 12.01$	$\eta_2 = 6.186$

The Weibull mixtures in MOR1 and MOR2 are very similar: the major population contributes 80% of the data and it has almost the same range as the data set. The minor population is nested in the central part of the major population with a relatively larger shape parameter. Also, the major population fits the right mode of the real data set while the minor population fits the left mode.

The above mixture models could be better supported if they were explained by the MSRC variables ("most strength reducing characters") measured on the tested boards, as introduced in Cheng (2010). We have tried to fit the posterior probability (the value of $\tilde{Z}_i^{(s)}$ in the last round of EM algorithm) with varied regression models on those covariates, but unfortunately no significant relationship was found. Other methods like classification tree and forests have also been tried, but they also failed to explain our mixture models either. Perhaps one possible reason is that the minor population is nested in the major population and it is extremely difficult to distinguish between them.

4.2   Bootstrap homogeneity test

Although the Weibull mixture currently cannot be explained by natural clusters (covariates) in the data, it is still possible to support it by testing whether the Weibull mixture fits the data significantly better than a single Weibull distribution. In this way, the following homogeneity hypothesis is tested:

H_0 : The sample $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ is from a single Weibull distribution v.s.

H_1 : $\mathbf{X}$ is from a mixture of two Weibull distributions.

Here the word "homogeneity" means that the data come from only one population and its opposite word "heterogeneity" means there is more than one population in our data set.

As the Weibull model is nested in the Weibull mixture model, the most natural test uses the likelihood ratio test. However as discussed above, the likelihood of the Weibull mixture is not finite, which violates one of the regularity conditions that ensures the common asymptotic properties of the likelihood ratio test. Together with some other issues similar to those for the homogeneity test of a normal mixture (Chen and Li, 2009), the twice the negative of log likelihood ratio between the single Weibull and Weibull mixture will not have asymptotically a χ^2 distribution with degrees of freedom equal the difference between number of parameters of the two models.

Varied approaches have been designed to perform the homogeneity test for the normal mixture, such as resampling (Wolfe, 1971), the bootstrap (McLachlan, 1987), the modified likelihood ratio (Chen et al., 2001), the L^2 distance (Charnigo and Sun, 2004), and the EM based modified likelihood ratio (Chen and Li, 2009; Li et al., 2009). The EM based modified likelihood ratio is the most powerful of them, but it requires additional work to design the penalty functions for the mixture proportion p and the shape parameter α. As our main purpose is to estimate the quantile rather than perform homogeneity test, we will adapt the simpler but less powerful bootstrap approach of McLachlan (1987).

4.2.1 Definition

First proposed by Efron (1979), the bootstrap method estimates the distribution of a statistic by re-sampling from a density estimate of the data, such as the empirical distribution or some parametric density estimate. A more detailed introduction will be provided in the next chapter. To conduct the homogeneity test, the log likelihood test statistic $-2\log \lambda$ will be bootstrapped, where λ is the ratio of the likelihood under H_0 (the single Weibull model) against the likelihood under H_1 (Weibull mixture).

According to McLachlan (1987), the bootstrap homogeneity test is carried out as follows: $(\tilde{\alpha}, \tilde{\eta})$ and $\tilde{\boldsymbol{\theta}} = (\tilde{p}, \tilde{\alpha}_1, \tilde{\eta}_1, \tilde{\alpha}_2, \tilde{\eta}_2)$ are the MLE of the parameters of the single Weibull and Weibull mixture models on the real data set and $-2 \log \lambda$ is the log likelihood ratio statistic between these two models; Next, a so-called bootstrap sample is generated from Weibull$(\tilde{\alpha}, \tilde{\eta})$, the value of $-2 \log \lambda^*$ is computed after fitting the single Weibull and the Weibull mixture for this bootstrap sample; This sampling and computation of $-2 \log \lambda^*$ will be repeated for $B = 500$ times to achieve an estimate of the distribution of $-2 \log \lambda$; The p-value is of this test is $\frac{1}{B} \sum_1^B \mathbf{1}\{-2 \log \lambda^* > -2 \log \lambda\}$. Here the choice of $B = 500$ is a moderate bootstrap replicates size, but it is already much larger than the recommendations in McLachlan (1987).

4.2.2 Evaluation of the p-value

Before we apply the bootstrap homogeneity test on our data sets, it is necessary to evaluate whether this test procedure yields the correct p-value. If so, the p-value should be uniformly distributed on $(0, 1)$ when the data are from the null hypothesis (the single Weibull for our case).

Here we simulated $N = 1000$ replicates from Weibull$(7, 7)$ with sample size 100 and ran the above test procedure on each of them. The empirical CDF of the p-values are presented in Figure 4.2, from which it is easy to tell that the empirical CDF of those p-values are quite close to the CDF of Uniform$(0, 1)$. The Komologov-Smirnov statistic between them is only 0.0256 with a p-value of 0.5461 for the goodness-of-fit test (The detail of this test can be found in D'Agostino and Stephens (1986)). Similar simulations have been done under different parameters of a Weibull model and with different sample sizes, which all yield similar conclusions. This indicates that the p-value is almost uniformly distributed between $(0, 1)$ and the p-value from this bootstrap homogeneity test is reliable.

4.2.3 Results and discussion

Following this procedure, the p-value of this homogeneity test on MOR1 is 0.032 and 0.016 for MOR2. Both the p-values are smaller than 0.05, which

Figure 4.2: Empirical CDF of the p-values and the CDF of Uniform$(0, 1)$
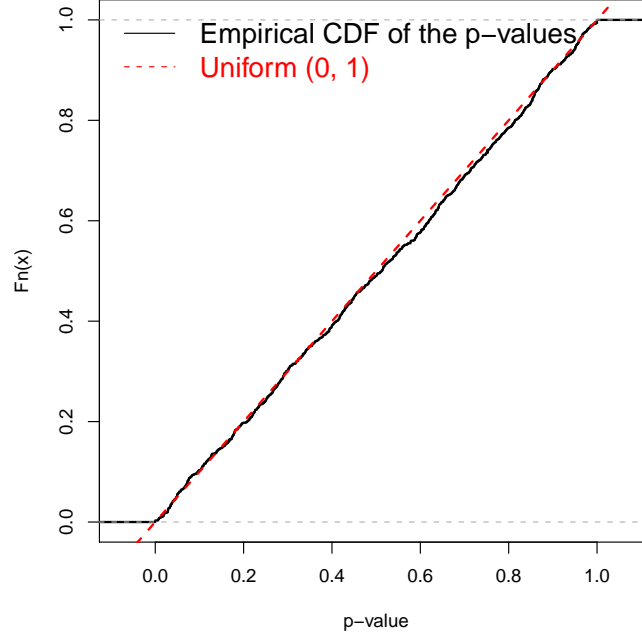
indicates that it is hardly possible that MOR1 and MOR2 are from a single Weibull distribution when compared to the Weibull mixture. Although we cannot conclude that they are from a Weibull mixture with this test, they partially support our choice of the Weibull mixture.

4.3 Censored Weibull mixture approach

As will be shown in the simulations, the Weibull mixture is not a very flexible approximation for some non-Weibull models. To ensure the accuracy of the quantile estimates, we can try to apply the idea of subjective censoring to the Weibull mixture. As the Weibull mixture is more flexible than a single Weibull distribution, it seems possible to achieve the accuracy as the censored Weibull MLE with less of censoring in the mixture model and the quantile estimate can be more efficient as more information are used.

Similarly as in the censored Weibull MLE case, the censoring threshold is denoted by C. The corresponding likelihood of the censored Weibull mixture with parameters $\boldsymbol{\theta} = (p, \alpha_1, \alpha_2, \eta_1, \eta_2)$ is:

$$L(\mathbf{X}; \boldsymbol{\theta}) = \prod_{i=1}^{n} [pf(X_i; \alpha_1, \eta_1) + (1-p)f(X_i; \alpha_2, \eta_2)]^{\delta_i} *$$

$$[p(1 - F(C; \alpha_1, \eta_1)) + (1-p)(1 - F(C; \alpha_2, \eta_2))]^{1-\delta_i}, \quad (4.9)$$

where $\delta_i = \mathbf{1}\{X_i \leq C\}$ indicates whether X_i is censored or not and $f(\cdot)$ and $F(\cdot)$ denote the PDF and CDF of the Weibull distribution respectively.

The maximum likelihood estimate of the $\boldsymbol{\theta}$ can be computed by the EM algorithm as in the case of the uncensored Weibull mixture. Here we will only provide the formulas to compute a few important quantities in the EM algorithm:

- E-step: The expectation of the sub-population label Z_i conditioning on the current $\boldsymbol{\theta}^{(s)}$ as in (4.4):

$$\tilde{Z}_i^{(s)} = E(Z_i | X_i, \boldsymbol{\theta}^{(s)}) = P(Z_i = 1 | X_i, \boldsymbol{\theta}^{(s)}) =$$

$$\begin{cases} \dfrac{p^{(s)} f(X_i; \alpha_1^{(s)}, \eta_1^{(s)})}{p^{(s)} f(X_i; \alpha_1^{(s)}, \eta_1^{(s)}) + (1 - p^{(s)}) f(X_i; \alpha_2^{(s)}, \eta_2^{(s)})} & , X_i \leq C \\ \dfrac{p^{(s)} (1 - F(C; \alpha_1^{(s)}, \eta_1^{(s)}))}{p^{(s)} (1 - F(C; \alpha_1^{(s)}, \eta_1^{(s)})) + (1 - p^{(s)})(1 - F(X_i; \alpha_2^{(s)}, \eta_2^{(s)}))} & , X_i > C \end{cases}.$$

- M-step: After plugging in the $\tilde{Z}_i^{(s)}$ into the incomplete likelihood, the updated $\tilde{p}^{(s+1)}$ is still $\frac{1}{n} \sum_{i=1}^{n} \tilde{Z}_i^{(s)}$. The iteration equation to solve for $\tilde{\alpha}_1^{(s+1)}$ is a re-weighted version of the iteration equation to be solved in the censored Weibull MLE case as in (3.10):

$$\frac{1}{\alpha} = \frac{\sum_{i=1}^{m} \tilde{Z}_i^{(s)} X_i^{\alpha} \ln(X_i) + (n-m)\tilde{Z}_n^{(s)} C^{\alpha} \ln(C)}{\sum_{i=1}^{m} \tilde{Z}_i^{(s)} X_i^{\alpha} + (n-m)\tilde{Z}_n^{(s)} C^{\alpha}} - \frac{\sum_{i=1}^{m} \tilde{Z}_i^{(s)} \ln(X_i)}{\sum_{i=1}^{m} \tilde{Z}_i^{(s)}},$$

where n is the sample size and we assume $X_1, X_2, \ldots, X_m \leq C$ while $X_{m+1}, X_{m+2}, \ldots, X_n \geq C$. Here it is easy to verify that $\tilde{Z}_{m+1}^{(s)} =$

$\tilde{Z}_{m+2}^{(s)} = \ldots, \tilde{Z}_n^{(s)}$ based on their definitions and they are all denoted by $\tilde{Z}_n^{(s)}$. Then we can obtain $\tilde{\eta}_1^{(s+1)}$ by

$$\tilde{\eta}^{(s+1)} = \left(\frac{\sum_{i=1}^m \tilde{Z}_i^{(s)} X_i^{\alpha^{(s+1)}} + (n-m)\tilde{Z}_n^{(s)} C^{\alpha^{(s+1)}}}{\sum_{i=1}^m \tilde{Z}_i^{(s)}} \right)^{1/\alpha^{(s+1)}}.$$

Then, equations to solve for the $\tilde{\alpha}_1^{(s+1)}$ and $\tilde{\eta}_1^{(s+1)}$ can be found by replacing $\tilde{Z}_n^{(s)}$ by $1 - \tilde{Z}_n^{(s)}$ in the above equations.

4.3.1 Choice of censoring threshold

As in the case of the censored Weibull MLE, a key issue in the application of censored mixture is the choice of the censoring threshold C. This issue has never been addressed in literature and we do not have an industrial standard to follow as in the case of censored Weibull MLE. Initially as a first step in our investigation of the censored Weibull mixture, we consider only the cases of C being the median and 70th empirical percentile of the sample.

Here we choose the censoring threshold no smaller than the median as it is found in our simulations that when C is smaller than the median, the EM algorithm for the censored mixture fails to converge in more than 10% of the replicates when the sample size is around 100 (similar as the size of MOR). This 10% failure ratio stops the censored mixture from becoming a reliable method in practice. (The other methods, such as the censored Weibull MLE and Weibull mixture, do not fail to converge in more than 1% of the replicates.) On the other hand, if the C is larger than the third quartile, the parameters in the censored mixture are very close to the uncensored mixture. However, our choice of the threshold is still far from scientifically validated and more work needs to be done on threshold selection for the censored mixture.

4.3.2 Censored Weibull mixture for real data

When the censoring threshold C is chosen to be the median, the fitted mixture models on the two data sets are plotted in Figure 4.3 and the parameter

estimates are in Table 4.2. For MOR1, the censored Weibull mixture is still very similar to the uncensored Weibull mixture as in Table 4.1. For MOR2, the major population and the minor population in the censored Weibull mixture have similar shape parameters but different scale parameters, which allows them to have different positions for the mode.

Figure 4.3: Censored Weibull mixture models with the median as the censoring threshold
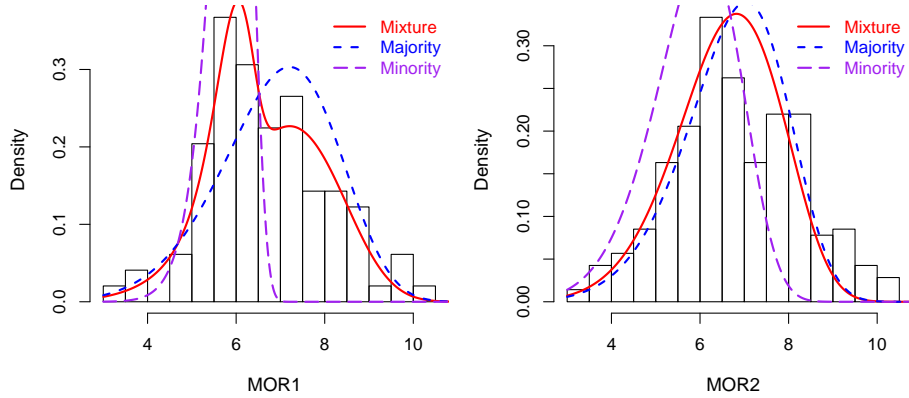
Table 4.2: Censored Weibull mixture models for MOR1 and MOR2 with median as the censoring threshold

Data	p	Majority Population		Minority Population	
MOR1	0.7478	$\alpha_1 = 6.034$	$\eta_1 = 7.428$	$\alpha_2 = 14.87$	$\eta_2 = 6.013$
MOR2	0.8397	$\alpha_1 = 6.806$	$\eta_1 = 7.208$	$\alpha_2 = 6.945$	$\eta_2 = 6.250$

When the censoring threshold is increased to the 70th empirical percentile, the censored mixture models are summarized in Table 4.3 and Figure 4.4. These censored mixture models are very similar to the uncensored Weibull mixture for both MOR1 and MOR2.

Although the parameters in the censored Weibull mixture might be very similar to that of the uncensored Weibull mixture models, the censored Weibull mixture can approximate the left tail of the real data sets better than the uncensored mixture, as shown by the truncated Kolmogorov-Smirnov

Figure 4.4: Censored Weibull mixture models with the 70th empirical percentile as the censoring threshold
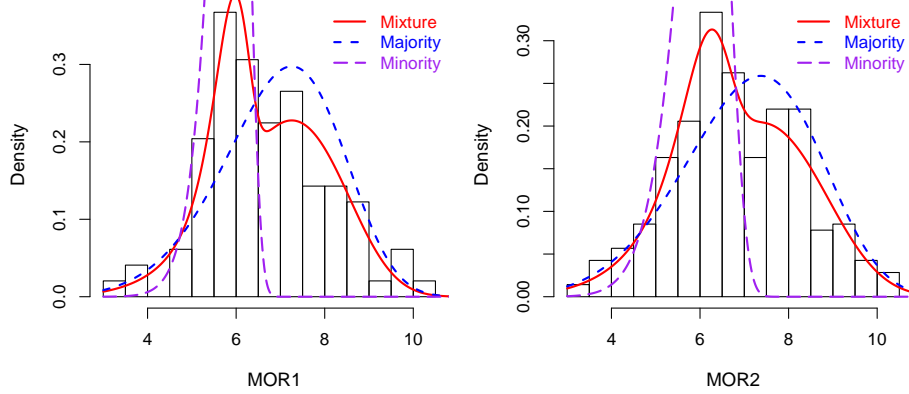
Table 4.3: Censored Weibull mixture models for MOR1 and MOR2 with the 70th empirical percentile as the censoring threshold

Data	p	Majority Population		Minority Population	
MOR1	0.7653	$\alpha_1 = 5.960$	$\eta_1 = 7.481$	$\alpha_2 = 16.42$	$\eta_2 = 5.951$
MOR2	0.7877	$\alpha_1 = 5.294$	$\eta_1 = 7.671$	$\alpha_2 = 12.22$	$\eta_2 = 6.205$

$D^{\ddagger}(\tilde{F}_n)$ in Table 4.4.

Table 4.4: Tail Goodness-of-Fit up to 10th empirical percentile $(D^{\ddagger}(\tilde{F}_n))$ of different Weibull mixture models on MOR1 and MOR2

	Uncensored	Censored Mixture with C		Censored
Data	Mixture	Median	70th Percentile	Weibull MLE
MOR1	0.0245	0.0202	0.0183	0.0207
MOR2	0.0146	0.0109	0.0132	0.0108

Here in Table 4.4, the definition of the $D^{\ddagger}(\tilde{F}_n)$ is the same as in (3.5) with 10th empirical percentile as the threshold and the $\tilde{F}_n$ will be the distribution estimate achieved by the censored or uncensored Weibull mixture. Also, the $D^{\ddagger}(\tilde{F}_n)$ achieved by the censored Weibull MLE is also included as a reference.

From Table 4.4, it is easy to see that the censored Weibull mixture

approximates the left tail of the real data sets better than the uncensored Weibull mixture. Moreover, although the censored Weibull mixture censors much less data (50% and 30%) than the censored Weibull MLE (90%), they achieve very similar tail goodness-of-fits. In the case of MOR, the censored mixture is even better than the censored Weibull MLE.

This indicates that the censored Weibull mixture can still approximate the tail reasonably well while it utilizes more information from the data, which probably makes it more efficient. However, our intuition that the more we censor the right part of the data, the better we approximate the left tail does not seem to be true for the censored Weibull mixture, as in MOR1 the one with the 70th empirical percentile as the threshold approximates the left tail better then the one with the median as the threshold.

4.4   Simulation comparison

In the previous chapter, we concluded that the censored Weibull MLE is the best quantile estimation approach under most of the models considered in our simulations. Here in this chapter, we will compare the censored Weibull MLE with the threshold as the 10th empirical percentile (CMLE), to the quantile obtained from the Weibull mixture (MIX), censored Weibull mixture with censoring threshold C being the median (50% censoring, CMIX5 for short) and the 70th empirical percentile (30% censoring, CMIX7 for short). The ordinary Weibull MLE, quantile obtained from the kernel density estimate, and the empirical quantile will not be included here as they are shown to be inferior to the CMLE.

The quantile of interest is still the 5th percentile. As the conclusions from the parametric models imitating MOR1 and MOR2 are very similar, we will only show the simulation results from the parametric models imitating MOR2 with sample size of 300 in this chapter. The number of replicates is still 10000. Similarly as in tables of RMSE in the previous chapter, the darker the background is, the larger the RMSE is.

4.4.1 RMSE of the quantile estimates

The RMSE of these five quantile estimates in the nine models are summarized in Table 4.5. Here the Monte Carlo standard deviation is controlled to be smaller than 0.001 and thus is omitted. From the background color in Table 4.5, it is hard to find a winner in most of the models considered here. The Weibull mixture has the smallest RMSE under the Weibull, Weibull mixture and the kernel density model imitating MOR2 while the CMIX7 is the best under the remaining parametric models: Gamma, log-normal, Minimum Gumbel, normal mixture and log-normal mixture. Although at least one of the Weibull mixture models (censored and uncensored) does better than the censored Weibull MLE in terms of the RMSE in all the models we considered in these simulations, none of them does uniformly better than the censored Weibull MLE in all the models considered here.

Table 4.5: RMSE of the quantile estimates from CMLE, MIX, CMIX5, and CMIX7, under the models imitating MOR2

Model	CMLE	MIX	CMIX5	CMIX7
Weibull	0.134	0.129	0.135	0.133
Log-normal	0.142	0.283	0.140	0.134
Gamma	0.139	0.203	0.140	0.136
Minimum Gumbel	0.155	0.140	0.140	0.139
Normal Mixture	0.125	0.125	0.124	0.121
Log-normal Mixture	0.139	0.146	0.138	0.138
Weibull Mixture	0.166	0.158	0.164	0.167
MOR2 (KDE)	0.148	0.147	0.151	0.155

4.4.2 Bias and standard deviation of the quantile estimates

To further compare these methods, we will decompose the RMSE into the bias and standard deviation as shown in Table 4.6. It is easy to see that the uncensored Weibull mixture has an extraordinarily large bias under the log-normal, gamma and normal mixture models, which indicates that it fails to approximate the left tail of those models.

On the other hand, it is interesting to find that the CMIX7 is gener-

ally less biased than all the other three quantile estimates. Meanwhile, the quantile estimates from the CMIX7 and CMLE have very similar standard deviations. In the sense of similar RMSEs but less bias, the CMIX7 offers better quantile estimates than the censored Weibull MLE.

Table 4.6: Bias and [Standard deviation] of the quantile estimates (×100) of the quantile estimates from CMLE, MIX, CMIX5, and CMIX7 under the models imitating MOR2

Model	CMLE	MIX	CMIX5	CMIX7
Weibull	0.46 [13.4]	1.55 [12.8]	1.00 [13.5]	1.08 [13.3]
Log-normal	4.75 [13.4]	-23.56 [15.6]	3.59 [13.6]	0.87 [13.4]
Gamma	3.46 [13.5]	-14.20 [14.5]	2.59 [13.7]	1.21 [13.6]
Minimum Gumbel	-4.27 [14.9]	0.45 [14.0]	-0.24 [14.0]	-0.16 [13.9]
Normal Mixture	2.16 [12.3]	-4.35 [11.7]	2.10 [12.3]	0.81 [12.1]
Log-normal Mixture	5.38 [12.8]	-5.37 [13.6]	4.37 [13.1]	3.74 [13.3]
Weibull Mixture	-2.15 [16.4]	-0.69 [15.8]	0.18 [16.4]	-0.43 [16.7]
MOR2 (KDE)	4.19 [14.2]	2.35 [14.5]	4.40 [14.5]	3.84 [15.1]

Table 4.6 also reveals several trends which contradict our intuition mentioned in the beginning of this chapter. It is expected that the more data we use, the more efficient quantile estimate we can achieve. Nonetheless the standard deviation of the quantile estimates from Weibull mixture is larger than the standard deviation of the CMIX5 under the log-normal, gamma, Gumbel and log-normal mixture models. Also, the standard deviation of CMIX7 is larger than that of the CMIX5 under the kernel density model imitating MOR2.

One possible explanation for this unexpected result is that when the Weibull mixture fails to approximate the true underlying model, e.g., log-normal and gamma, more information (no censoring) will not help to improve the efficiency. Also, the Weibull mixture is too complicated to be tracked theoretically during my master study and thus a more convincing explanation remains to be explored.

Perhaps we will find a threshold that makes the censored Weibull mixture uniformly better than the censored Weibull MLE, but the Weibull mixture is too complex to be a good starting point of the study on how to choose

the threshold. In this way, we will study how to select a better censoring threshold for the censored Weibull MLE in the next chapter.

Chapter 5

Bootstrap Threshold Selection in the Censored Weibull MLE

As discussed in Chapter 3, the choice of censoring threshold C to be the 10th empirical percentile is an ad-hoc choice in the current industrial standard (ASTM, 2004a). In this chapter, we will first study the relationship between the censoring threshold and the mean squared error of the quantile estimate in the censored Weibull MLE approach. In order to find the optimal threshold for a data set, we propose to first estimate the MSE of quantile estimates of the censored Weibull MLE with different thresholds by the bootstrap and then choose the optimal threshold with the smallest estimated MSE. According to our simulations, this censored Weibull MLE with bootstrap threshold selection method is better than the original CMLE.

5.1 Relationship between MSE and the censoring threshold

In Chapter 3, the censoring threshold C is fixed to be the 10th empirical percentile in the censored Weibull MLE approach, which is an ad-hoc choice in industrial standard (ASTM, 2004a). If we increase C, more data will

be included in the analysis, so the quantile estimate can be more efficient. But at the same time, the increase of C might reduce the goodness of the approximation of the non-Weibull distributions achieved by the Weibull left tail and thus the quantile estimate might be become more biased. As the mean squared error is the sum of variance and bias, it is not clear that how the MSE will change with the censoring threshold, but this relationship can be easily estimated by Monte Carlo methods if the model is known.

Here Figure 5.1 shows the relationship between MSE and censoring threshold (as an empirical percentile). These theoretical MSEs are evaluated by 20000 replicates under all the 8 models imitating MOR2 with sample size of 300. The candidates for censoring threshold range from the 10th empirical percentile to 100th empirical percentile (no censoring) in steps of 5%. Similar curves can be found under the models imitating MOR1 and are thus omitted.

Figure 5.1: Relationship between the MSE and censoring threshold in the models imitating MOR2
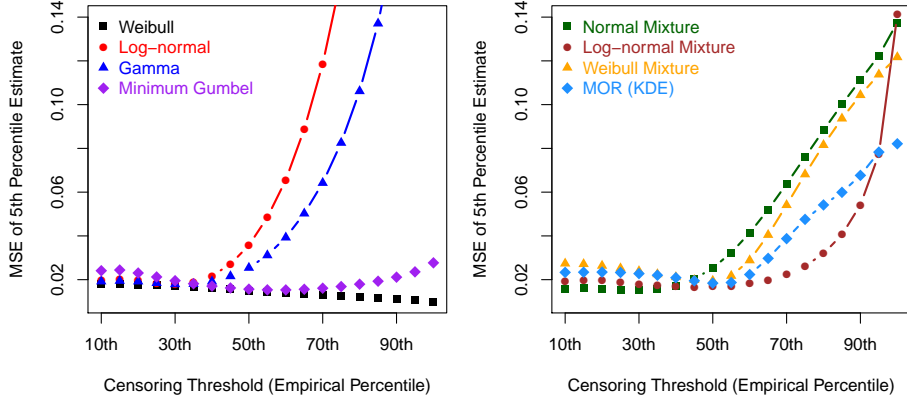
Under the Weibull model, the MSE continuously decreases with the threshold, as the censored Weibull MLE is almost unbiased under any threshold, but the variance of the quantile estimate decreases with the increase of threshold. For all the other models, the MSEs all decrease first and then increase dramatically for large thresholds. As for small censoring thresholds, the bias does not increase much while the variance decreases, resulting in a

decreasing MSE. On the other hand, when the censoring threshold is large, for example, larger than the median, the bias of the quantile estimate increases dramatically while the benefits in the variance shrinks, which leads to the increase of MSE. The optimal censoring threshold (with the smallest MSE) of most of those non-Weibull models are around 30th to 50th percentile.

5.2 Bootstrap estimate of MSE

In the above simulations, we can easily find out the relationship between the MSE and the censoring threshold as we know the true model. But in practice, the true model is never known. Meanwhile, the empirical distribution of the real data set is a consistent estimate of the true distribution. It is possible to estimate the MSE under a certain censoring threshold by bootstrapping from the real data set and then choosing the threshold with the smallest estimated MSE.

The bootstrap, first proposed by Efron (1979), is a re-sampling technique to approximate and estimate the sampling distribution of a statistic and its properties, such as the bias and variance. For our problem, the bootstrap estimate of MSE

$$E_{G_n} \left(\tilde{q}_n^*(C) - \hat{q}_n \right)^2,$$

can be an estimate of the true MSE under the model G

$$E_G \left(\tilde{q}_n(C) - q \right)^2.$$

Here the G_n is the empirical distribution of a sample from model G and $\hat{q}_n$ is the empirical quantile (Type IX definition in R). $\tilde{q}_n(C)$ is the quantile estimate obtained from the censored Weibull MLE with threshold C and $\tilde{q}_n^*(C)$ is the censored Weibull MLE under the bootstrap sample. Notice here the G does not necessarily have to be a Weibull model or a model with a Weibull tail. It can be any reasonable parametric model for our data.

If the $E_{G_n} \left(\tilde{q}_n^*(C) - \hat{q}_n \right)^2$ is a consistent estimator of $E_G \left(\tilde{q}_n(C) - q \right)^2$,

namely,

$$E_{G_n}(\tilde{q}_n^*(C) - \hat{q}_n)^2 - E_G\left(\tilde{q}_n(C) - q\right)^2 \to_p 0, n \to \infty, \qquad (5.1)$$

we will have,

$$\lim_{n \to \infty} Pr\left\{E_{G_n}(\tilde{q}_n^*(C_1) - \hat{q}_n)^2 > E_{G_n}(\tilde{q}_n^*(C_2) - \hat{q}_n)^2\right\} = 1, \qquad (5.2)$$

when

$$E_G(\tilde{q}_n(C_1) - \hat{q}_n)^2 > E_G(\tilde{q}_n(C_2) - \hat{q}_n)^2, \qquad (5.3)$$

for any given C_1 and C_2.[1]

In this way, we can choose the best threshold corresponding to the smallest MSE of the quantile estimate from a finite set of threshold candidates asymptotically. In a finite sample with fairly large sample size, the probability that we choose the best threshold can also be larger than the probability we fail to choose the best threshold.

In this section, the computation of the bootstrap estimate of MSE will first be described. Then, we will discuss a promising plan to prove the consistency of bootstrap estimate of MSE. Later, a few simulations on the consistency of bootstrap MSE will be presented.

5.2.1   Computation of the bootstrap MSE

Before we start to discuss the simulations designed to verify the consistency of the bootstrap estimate MSE, it is first necessary to clarify how to compute $E_{F_n}(\tilde{q}_n^*(C) - \hat{q}_n)^2$: For an *i.i.d.* sample $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$, we will first calculate the Type IX empirical quantile introduced in Chapter 3 as $\hat{q}_n$. Then, we will re-sample $\mathbf{X}$ with replacement and get the bootstrap sample $\mathbf{X}^* = \{X_1^*, X_2^*, \ldots, X_n^*\}$. Next, we will apply a Type II subjective censoring on this bootstrap sample, namely, calculate the censored Weibull MLE based

---

[1]Here we may need a few more conditions to ensure we can achieve (5.2) from (5.3) and (5.1). As the proof of (5.1) is still incomplete at the time of writing, we will not discuss those conditions here.

on the smallest r observations in the bootstrap sample, where r is the number of the observations no larger than the censoring threshold C in the original sample $\mathbf{X}$. Based on the censored Weibull MLE of the parameters, we can obtain the quantile estimate on this bootstrap sample $\tilde{q}^*_{\mathbf{X}^*}$. The re-sampling and calculation of $\tilde{q}^*_{\mathbf{X}^{*i}}$ will be repeated for B (e.g., $B = 5000$) times and the upper-script will be used to indicate the i-th bootstrap sample; Finally, $E_{G_n}(\tilde{q}^*_n(C) - \hat{q}_n)^2$ is calculated as

$$E_{G_n}(\tilde{q}^*_n(C) - \hat{q}_n)^2 = \frac{1}{B}\sum_{i=1}^{B}(\tilde{q}^*_{\mathbf{X}^{*i}} - \hat{q}_n)^2.$$

Given a set of threshold candidates C for the sample $\mathbf{X}$, the threshold with the smallest bootstrap MSE will be selected as the best threshold and the corresponding quantile estimate under this threshold will be an quantile estimate of this sample, which is named as the bootstrap censored Weibull MLE (BMLE for short).

As the wood engineering field does not want to use less than 10% of the data nor more than half of the data, we now consider only selecting the thresholds from $10^{th}, 20^{th}, \ldots, 50^{th}$ empirical percentiles for the 5th percentile estimate. The step of 10% is chosen for the computational simplicity.

Here are some issues to be clarified for the bootstrap procedure.

Discussion on the types of censoring in bootstrap

As we have discussed in Chapter 3, whether the subjectively censoring in the CMLE is treated as Type I or Type II censoring will not affect the quantile estimate. But whether we use Type I or Type II censoring in the bootstrap will affect the quantile estimate and thus the MSE estimate. We can also apply Type I censoring in the bootstrap sample, that is, censoring the bootstrap sample with the threshold C, which is a sample quantile of the original sample (an order statistic decided from Type III definition of the empirical quantile), such as the 10th empirical percentile. The two different types of censoring in the bootstrap will result in slightly different MSE estimates.

If we compare the RMSE of BMLE quantile estimates with type I or type II censoring, type II censoring is slightly better than type I censoring in our simulations. Those results are available upon request. In this way, we will use type II censoring for our study.

Number of bootstrap replicates

In the bootstrap procedure, the number of bootstrap replicates B will affect the precision of the MSE estimate. Clearly, the larger B is, the more precise the MSE estimate is and the more computation is required. For our study, the B is usually chosen to be 5000, which is chosen to guarantee that the MSE estimates in our two real data sets given the threshold candidates $10\%, 20\%, \ldots, 50\%$ will be significantly different when the Monte Carlo randomness in the bootstrap is considered. In this way, the threshold selected will not be affected by Monte Carlo randomness and be sure to have the smallest RMSE one under the empirical distribution of the real data sets.

However, in our simulations, each replicate is different and sometimes the bootstrap estimated MSE under different thresholds are almost equal. It is thus almost impossible to ensure the threshold selected in the simulated data sets will not be affected by the bootstrap Monte Carlo randomness when $B = 5000$. Although we want more bootstrap replicates, $B = 5000$ is already the computational limit we can afford in the simulations.

5.2.2 Consistency of the bootstrap estimate of MSE

To ensure our bootstrap censored Weibull MLE works asymptotically, it is necessary to show the consistency of the bootstrap estimate of MSE as in (5.1). The population MSE and bootstrap MSE can be decomposed into the squared bias and variance two parts:

$$
\begin{aligned}
E_{G_n}(\tilde{q}_n^* - \hat{q}_n)^2 =& Var(\tilde{q}_n^*) + (E_{G_n}\tilde{q}_n^* - \hat{q}_n)^2 \\
E_G(\tilde{q}_n - q)^2 =& Var(\tilde{q}_n) + (E_G\tilde{q}_n - q)^2.
\end{aligned}
$$

Based on the Slutsky's theorem (as introduced in Casella and Berger

(2002)), we can establish the consistency of the bootstrap estimate of MSE (5.1) by showing that

$$(E_{G_n}\tilde{q}_n^* - \hat{q}_n)^2 - (E_G\tilde{q}_n - q)^2 \to_p 0, n \to \infty \tag{5.4}$$

$$Var(\tilde{q}_n^*) - Var(\tilde{q}_n) \to_p 0, n \to \infty. \tag{5.5}$$

Simplification of (5.4)

The left hand of (5.4) can be simplified as

$$(E_{G_n}\tilde{q}_n^* - \hat{q}_n)^2 - (E_G\tilde{q}_n - q)^2 =$$
$$[(E_{G_n}\tilde{q}_n^* - \hat{q}_n) + (E_G\tilde{q}_n - q)] \, [(E_{G_n}\tilde{q}_n^* + \hat{q}_n) - (E_G\tilde{q}_n - q)]$$

The empirical quantile $\hat{q}_n$ and the true quantile q are finite and the estimated quantile is obtained by $\tilde{q} = \tilde{\eta} \left[-\log(1 - p) \right]^{1/\tilde{\alpha}}$. The proportion of interest is smaller than 0.1, which indicates that $-\log(1 - p) < -\log(0.9) < 0.2$. The observations of MOR are always finite in our study and the parameter estimates η and α for them are bounded according to Rinne (2009). So, the quantile estimate is guaranteed to be bounded, which further indicates that $[(E_{G_n}\tilde{q}_n^* - \hat{q}_n) + (E_G\tilde{q}_n - q)]$ will be bounded. In this way, (5.4) is equivalent to

$$[(E_{G_n}\tilde{q}_n^* + \hat{q}_n) - (E_G\tilde{q}_n - q)] \to_p 0, n \to \infty.$$

Again, with Slutsky's theorm, (5.4) can be proved by showing

$$E_{G_n}\tilde{q}_n^* - E_G\tilde{q}_n \to_p 0, \qquad\qquad n \to \infty \tag{5.6}$$

$$\hat{q}_n - q \to_p 0, \qquad\qquad n \to \infty \tag{5.7}$$

The equation (5.7) is a well-known result, for example in Serfling (1980). What remains to be proved is (5.6).

Strengthening (5.5)

As discussed in Chapter 3, the asymptotic variance of $\sqrt{n}\tilde{q}_n$ is a finite number that depends on the parameters and censoring threshold under the Weibull model, which indicates that $Var(\tilde{q}_n) \to_p 0$ when $n \to \infty$. Therefore, (5.5) can be proved if $Var(\tilde{q}_n^*) \to_p 0$ when $n \to \infty$. However, without any bootstrap, $Var(\tilde{q}_n^*) \equiv 0$ is also a consistent estimate of $Var(\tilde{q}_n)$. Although it might be true that the MSE equals to the squared bias asymptotically, we still need to estimate the variance together with the bias part for the usage in a finite sample.

In this way, we will strengthen (5.5) to:

$$Var(\tilde{q}_n^*)/Var(\tilde{q}_n) \to_p 1, n \to \infty. \tag{5.8}$$

If (5.8) is true,

$$1 - Var(\tilde{q}_n^*)/Var(\tilde{q}_n)) \to_p 0, n \to \infty$$

and thus

$$Var(\tilde{q}_n^*) - Var(\tilde{q}_n) \to_p 0, n \to \infty.$$

when $Var(\tilde{q}_n)$ is finite.

A tentative plan to prove the consistency

As discussed above, we can establish the consistency of bootstrap MSE estimate by showing (5.6) and (5.8), which are kind of convergence of in the first two moments (Shao and Tu, 1995). More detailed introduction of this stochastic convergence in moments can be found in Bickel and Freedman (1981) and Shao and Tu (1995).

Most of the techniques for proving the consistency of bootstrap have been summarized in Shao and Tu (1995). These include the Mallows' distance, imitation and linearization, etc. Shao and Tu (1995) developed most of their theorems on the consistency of bootstrapping M-estimators with

the help of the differentiability of the estimators. However, it is difficult to verify the differentiability of our censored Weibull MLE due to the lack of corresponding theoretical results. According to our literature review, some sufficient conditions to ensure an M-estimator is differentiable are studied in Clarke (1983) and Clarke (1986), which requires the estimator to be continuous in x (the domain of the random variable). As shown in Section 3.5, our censored Weibull MLE is not continuous in x. Although the continuity is only a sufficient condition for differentiability, it will be very time consuming to verify the differentiability from definition.

On the other hand, it is possible to find the Bahadur representation (a kind of linearization of the estimator, as discussed in (Serfling, 1980)) of the censored Weibull MLE with the similar techniques as discussed in Jureckova and Sen (1987) and He and Shao (1996). With the Bahadur representation the estimate, it is possible to demonstrate the stochastic convergence of the distribution of the bootstrapped censored Weibull MLE with similar techniques as discussed in Section 3.1.5 in Shao and Tu (1995). Then, with some conditions including uniform integrability, the stochastic convergence in moment might be established.

The above is only a tentative plan to prove the consistency of our bootstrap MSE estimate, but the substantial amount of work leads us to defer this to future work. However, we will partially demonstrate the consistency with simulations.

5.2.3   Simulation evaluation of the consistency

If the bootstrap MSE is a consistent estimate of the true population MSE, we should observe in the simulations that the distribution of the difference between the bootstrap MSE and the true population MSE degenerates to zero with the increase of sample size, as the convergence in probability implies converges in distribution.

For these simulations, we will simulate $N = 10000$ replicates from a known model imitating our real data sets with sample size as (300, 500, 1000, 2000, 3000) respectively. For each replicate of each sample size, we will

first calculate the quantile estimate $\tilde{q}_n^i$ from censored Weibull MLE using the 10th empirical percentile as the threshold, where the superscript indicates the i-th replicate with sample size n (subscript). This quantile estimates will be used to calculate the Monte Carlo estimate of the population MSE under sample size n, donated as M_n:

$$M_n = \frac{1}{N} \sum_{i=1}^{N} (\tilde{q}_n^i - q)^2,$$

where the q is the true quantile under the model we simulate data from.

Moreover, for each replicate, we will also apply the bootstrap procedure introduced above to obtain the MSE estimate $\widehat{M}_n^i$ with $B = 5000$ bootstrap replicates and the $\widehat{M}_n^i - M_n$ gives a Monte Carlo estimate of the distribution of the difference of the bootstrap MSE and population MSE.

This set of simulations has been performed on all the models imitating MOR1 and MOR2 as introduced in Section 3.4. The trend of the distribution of $\widehat{M}_n^i - M_n$ with respect to n is very similar among all those models. For simplicity, we will only present the results on the Weibull and Weibull mixture models imitating MOR2. The box-plots of $\widehat{M}_n^i - M_n$ under different sample sizes are in Figure 5.2.

Figure 5.2: The box-whisker plot of $\widehat{M}_n^i - M_n$ under different sample sizes
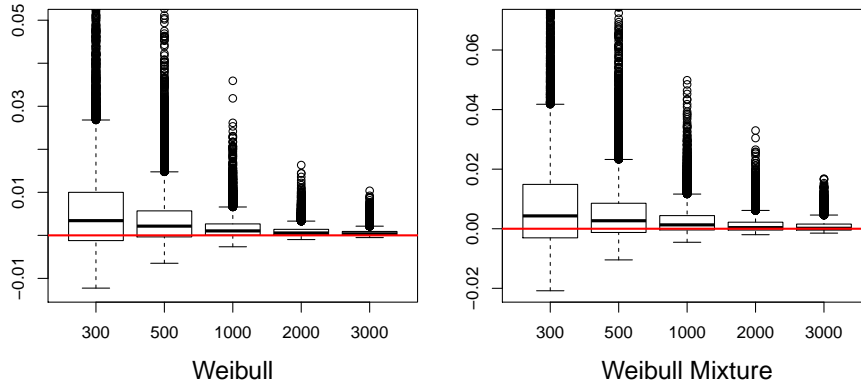
From Figure 5.2, it is easy to find that the "box" (distribution) of $\widehat{M}_n^i - M_n$ concentrates at zero and the right tail becomes shorter with an increase of the sample size. To evaluate this convergence, the $\sqrt{n}$ times the mean and standard deviation of $\widehat{M}_n^i - M_n$ are summarized in Table 5.1 and Table 5.2. It is easy to see that both of them strictly decrease with the increase of sample size, which probably indicates that the mean and standard deviation of $\widehat{M}_n - M_n$ decrease at a rate of $1/\sqrt{n}$.

Table 5.1: $\sqrt{n}$ times the mean of $\widehat{M}_n^i - M_n$ under different sample sizes

Model	300	500	1000	2000	3000
Weibull	0.104	0.083	0.059	0.045	0.036
Weibull Mixture	0.152	0.122	0.091	0.062	0.051

Table 5.2: $\sqrt{n}$ times the standard deviation of $\widehat{M}_n^i - M_n$ under different sample sizes

Model	300	500	1000	2000	3000
Weibull	0.193	0.145	0.094	0.065	0.051
Weibull Mixture	0.316	0.239	0.166	0.133	0.117

The above trends in these simulations are just some necessary conditions for the consistency of bootstrap MSE. So far, we just have not found any evidence to disprove the consistency. It is still necessary to find a rigorous mathematical proof to prove its consistency. We need to make sure that this bootstrap censored Weibull MLE is a better method than the original industrial standard (CMLE) and thus worthy proving the consistency of bootstrap MSE.

5.3   Simulation comparison

In the following simulations, we will compare the 5th percentile estimate achieved by the bootstrap censored Weibull MLE (BMLE) to the other quantile estimates introduced in the previous chapters, which includes the cen-

Table 5.3: RMSE of the quantile estimates from CMLE, BMLE, MIX, CMIX7, and EMP under the models imitating MOR1

Model	CMLE	BMLE	MIX	CMIX7	EMP
Weibull	0.150	0.147	0.149	0.152	0.175
Log-normal	0.167	0.174	0.396	0.159	0.186
Gamma	0.162	0.166	0.269	0.159	0.183
Minimum Gumbel	0.169	0.143	0.153	0.147	0.165
Normal Mixture	0.115	0.120	0.139	0.110	0.130
Log-normal Mixture	0.158	0.156	0.210	0.155	0.196
Weibull Mixture	0.163	0.149	0.152	0.163	0.187
MOR1 (KDE)	0.163	0.159	0.169	0.156	0.188

sored Weibull MLE with 10th empirical percentile as the threshold (CMLE), the Weibull mixture (MIX), the censored Weibull mixture with 70th empirical percentile as the threshold (CMIX7). These three quantile estimates are "winners" among the quantile estimates proposed in Chapters 3 and 4. Moreover, the empirical quantile will also be considered as reference.

The models from which we simulate data are the same as in Chapter 3: eight models imitating MOR1 and eight models imitating MOR2. The sample size here are all chosen to be 300. Under each model, we will have $N = 10000$ replicates and each replicate will be bootstrapped $B = 5000$ times for the BMLE. The RMSE of those quantile estimates are summarized in Table 5.3 and 5.4. As in the previous chapters, a darker background corresponds to a larger RMSE. Here the Monte Carlo estimation standard error of those RMSE is controlled to be around 0.001 and thus the at least the first two digits are significant. For ease of displaying the tables, we will not include the Monte Carlo standard error here, but they are available upon request.

In the rest of this section, we will analyse the simulation results in Table 5.3 and 5.4 to compare the five quantile estimates. To under the advantage and limitation of the BMLE, we will compare the bias and standard deviation of the BMLE, CMLE in Section 5.3.2. Later, we will study the threshold chosen by the BMLE procedure to understand the advantages and

Table 5.4: RMSE of the quantile estimates from CMLE, BMLE, MIX, CMIX7, and EMP under the models imitating MOR2

Model	CMLE	BMLE	MIX	CMIX7	EMP
Weibull	0.134	0.131	0.130	0.134	0.155
Log-normal	0.142	0.148	0.283	0.134	0.158
Gamma	0.138	0.142	0.202	0.135	0.156
Minimum Gumbel	0.155	0.134	0.140	0.137	0.154
Normal Mixture	0.127	0.132	0.127	0.123	0.144
Log-normal Mixture	0.139	0.134	0.146	0.139	0.153
Weibull Mixture	0.167	0.156	0.159	0.168	0.192
MOR2 (KDE)	0.150	0.147	0.148	0.158	0.179

limitations of BMLE.

5.3.1 Comparison of the RMSE of the quantile estimates

From the background shades of Table 5.3 and 5.4, it is easy to see that either BMLE or CMIX7 is the best quantile estimates in most of the models we considered. Among the models imitating MOR1, CMIX7 has the smallest RMSE in 5 models while BMLE has the smallest RMSE in 3 models; Among the models imitating MOR2, BMLE is the best in 4 models while CMIX7 is the best in 3 models.

Although it seems that CMIX7 achieves the largest number of "first places" in the 16 models considered in our simulations, it also has the second largest RMSE in several models, such as the Weibull or Weibull mixture models, which means that it is only better than the empirical quantile under those models. On the other hand, only the normal mixture model imitating MOR2, BMLE has the second largest RMSE.

In order to summarize the performance of these five quantile estimates, we will compare their average rank in the models from which we simulate data. For each model, we will first calculate the rank of these five quantile estimates. For example, in the Weibull model imitating MOR, the ranks are: CMLE (3), BMLE (1), MIX (2), CMIX7 (4) and EMP (5). Then the ranks of each quantile estimate will be averaged across the 16 models. The

results are:

	CMLE	BMLE	MIX	CMIX7	EMP
Averaged rank	2.9375	2.1875	2.9375	2.2500	4.6875

It is easy to find that the BMLE has the smallest averaged rank among these five quantile estimates and the CMIX7 is in the second place. Furthermore, as the lumber strength data is commonly modelled by a Weibull distribution, the models closely related to the Weibull model, such as Gumbel, Weibull mixture, and the KDE models of real data sets should be assigned more weight in the averaged rank. Those models are also the models where BMLE generally does better than CMIX7. In this way, the average rank obtained in our simulations might "underestimates" the advantage of BMLE over CMIX7 in practice.

The BMLE is uniformly better than the empirical quantile and the BMLE has a smaller RMSE than the Weibull mixture (MIX) in all except one models considered in our simulations. However, when we compare the BMLE to the fixed threshold CMLE, the BMLE has larger RMSE than the CMLE under the Gamma, log-normal and normal mixture models, which stimulates us to first compare the bias and standard deviation of the BMLE and CMLE.

5.3.2 Comparison of the bias and standard deviation of BMLE and CMLE

The bias and standard deviation (SD) of the quantile estimates from BMLE and CMLE in the models imitating MOR2 are summarized in Table 5.5. The results from the models imitating MOR1 are very similar as in Table 5.5 and are thus omitted. Similarly as in the previous tables, a darker background corresponds to a larger absolute bias (standard deviation).

For Table 5.5, it is easy to find that the BMLE reduces the bias of the quantile estimate in most of the models here. However, except the Weibull, Gumbel, Weibull mixture models, the quantile estimate from BMLE has a larger SD than the CMLE. The increase of SD is especially large in the

Table 5.5: Bias and standard deviation (SD) of the quantile estimates ($\times 100$) from CMLE and BMLE under the models imitating MOR2

	Bias		SD	
Model	CMLE	BMLE	CMLE	BMLE
Weibull	0.87	0.95	13.4	13.1
Log-normal	4.64	-0.43	13.4	14.8
Gamma	3.66	-0.50	13.3	14.2
Minimum Gumbel	-4.19	-0.24	14.9	13.4
Normal Mixture	2.29	-1.19	12.4	13.2
Log-normal Mixture	5.31	1.94	12.8	13.3
Weibull Mixture	-1.87	-0.97	16.6	15.6
MOR2 (KDE)	4.26	2.12	14.4	14.5

Gamma, log-normal and normal mixture models, where the BMLE has a larger RMSE than the CMLE.

Taken the log-normal model as an example, standard deviation of the $100\times$ quantile estimates from the censored Weibull MLE with fixed censoring threshold at $10^{th}, 20^{th}, \ldots, 50^{th}$ empirical percentile are 13.5, 13.4, 13.6, 13.7, 13.7 respectively, which are all smaller than the $100\times$ standard deviation of the BMLE (14.8).

This is because the selection procedure will actually bring in additional variance to the quantile estimates, as the BMLE is a data-driven procedure and it can select different thresholds in different samples. To understand the randomness introduced by the selection procedure, we will study the threshold selected in our simulations.

5.3.3 Discussion on the threshold selection of BMLE

Here Table 5.6 has summarized the proportion of each threshold been selected by the bootstrap procedure in the $N = 10000$ replicates in the models imitating MOR2. Similar trend is observed in the models imitating MOR1 and thus we have omitted the corresponding table. As a reference, the theoretical RMSE of the quantile estimate from censored Weibull MLE with

fixed thresholds as $10\%, 20\%, \ldots, 50\%$ empirical quantile under those models are also included.

In Table 5.6, the first row under each model is the theoretical RMSE (estimated by Monte Carlo in our simulations) and the second row is the proportion of the threshold selected by the bootstrap procedure. Similarly as in the previous tables, a darker background corresponds to a larger RMSE. Ideally, the larger the RMSE is, the less frequent it will be selected by the bootstrap. So for the proportion of selected thresholds, a darker background will indicate a smaller proportion, which ideally will make the background colour of the two rows under a model agree.

Table 5.6: RMSE of the quantile estimates from censored Weibull MLE with different fixed censoring thresholds and the proportion of these thresholds selected by the BMLE under the models imitating MOR2

Model	10%	20%	30%	40%	50%
Weibull	0.134	0.134	0.131	0.127	0.122
	21.0%	12.7%	12.7%	15.5%	38.1%
Log-normal	0.142	0.141	0.135	0.148	0.190
	22.9%	19.2%	21.5%	19.2%	17.3%
Gamma	0.138	0.138	0.133	0.138	0.160
	22.9%	17.1%	18.5%	18.0%	23.5%
Minimum Gumbel	0.155	0.151	0.139	0.130	0.124
	15.9%	10.0%	9.8%	12.5%	51.1%
Normal Mixture	0.127	0.126	0.123	0.131	0.159
	23.0%	17.8%	19.6%	18.7%	21.0%
Log-normal Mixture	0.139	0.139	0.134	0.130	0.130
	23.0%	14.3%	15.1%	16.2%	31.4%
Weibull Mixture	0.167	0.164	0.157	0.148	0.141
	10.7%	6.5%	9.8%	21.3%	51.6%
MOR2 (KDE)	0.150	0.150	0.148	0.143	0.133
	21.4%	11.4%	10.6%	13.7%	42.9%

Obviously, the background shades of every two rows in Table 5.6 barely agree, which indicates that the BMLE procedure does not exactly follow our expectation to choose the threshold with a smaller RMSE more often

than the threshold with a larger RMSE. It seems that the BMLE procedure chooses the 10th and 50th as the best threshold more often than the other thresholds, especially in the log-normal, gamma and normal mixture models. This is the main reason why the BMLE is worse than the CMLE under those models.

However, as the empirical distribution is only an approximation of the true population distribution, we should not expect to be able tos choose the optimal threshold of the true population based on a sample of 300. Moreover, it might be difficult for the Weibull left tail to approximate the left tail of those models. On the other hand, in the models like Weibull, Gumbel, Weibull mixture and the KDE model imitating MOR2, which are relatively easier for a Weibull model to approximate, the BMLE chooses the theoretically best threshold more frequently than any other thresholds.

5.3.4 Advantages and limitations of BMLE

To further study the advantage and limitation of BMLE, we will first get the following two differences from our simulation results: The first is the difference between the RMSE of the quantile estimates from CMLE and BMLE, which will be denoted by

$$R_{\text{CMLE}} - R_{\text{BMLE}}.$$

The other difference of the difference between the RMSE of the quantile estimates from CMLE (censored Weibull MLE with 10th empirical percentile as threshold) and minimum RMSE of the quantile estimate from the censored Weibull MLE with fixed thresholds in our threshold candidates set, which will be denoted by

$$R_{\text{CMLE}} - \min\{R\}.$$

Take the Weibull model imitating MOR2 as an example, the minimum RMSE is achieved by the threshold of 50th empirical percentile. The $R_{\text{CMLE}} - \min\{R\}$ equals $0.134 - 0.122$, as shown in Table 5.6.

Here we will plot the $R_{\text{CMLE}} - R_{\text{BMLE}}$ against $R_{\text{CMLE}} - \min\{R\}$ in Figure 5.3. It is easy to find that the $R_{\text{CMLE}} - \min\{R\}$ are all smaller than 0.009 in the models where the BMLE has a larger RMSE than the CMLE. This indicates that when the difference in the RMSE between the original CMLE and the best censored Weibull MLE with a fixed threshold (given in our candidates set) is very small, it is very difficult for the bootstrap procedure to identify this difference and choose the best threshold. This limitation of BMLE might help to explain why the bootstrap procedure selects the original 10th percentile as threshold more often than the best threshold in the log-normal, gamma and normal mixture models.

Figure 5.3: Relationship between $R_{\text{BMLE}} - R_{\text{CMLE}}$ and $R_{\text{CMLE}} - \min\{R\}$ in the mdoels imitating MOR1 and MOR2
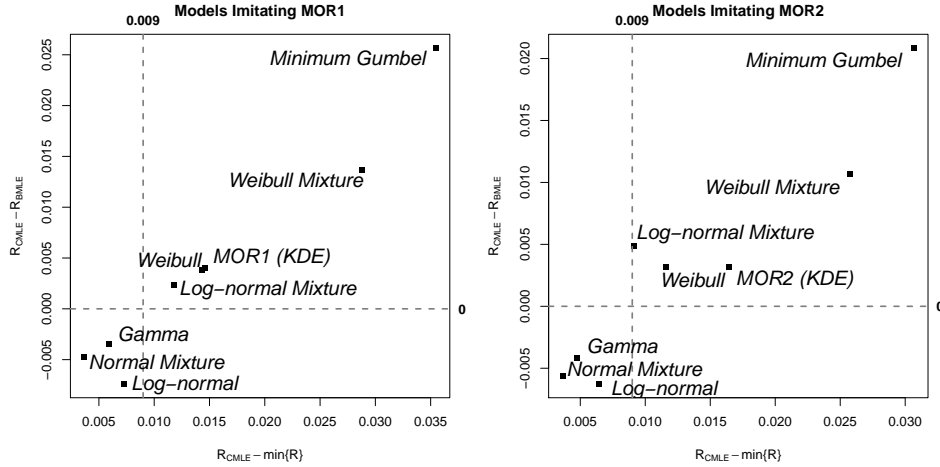
Moreover, from Figure 5.3, another interesting trend is that $R_{\text{BMLE}} - R_{\text{CMLE}}$ generally increases with $R_{\text{CMLE}} - \min\{R\}$; that is, the BMLE will have more advantage over the CMLE when the difference between the RMSE of CMLE and the censored Weibull MLE with the best threshold becomes larger and thus easier for the bootstrap procedure to identify.

In order to further improve the BMLE, it is necessary to find a better candidate threshold sets for our data, which remains to be explored. After all, the bootstrap censored MLE is a better quantile estimate than the

73

CMLE as the BMLE reduces the RMSE of the quantile estimates in the models relatively closer to our real data sets and also reduces the bias of quantile estimate in most of the models we consider.

Chapter 6

Conclusions and Future Work

In this thesis, we mainly focus on the statistical methods of lower quantile estimation of wood strength data, which is a critical problem for the safety of construction with lumber materials. The intuitive non-parametric quantile estimates, such as the empirical quantile, are unbiased but very inefficient, while the parametric quantile estimates are fully efficient but seriously biased under mis-specified models. In order to overcome the shortcomings of the non-parametric or parametric quantile estimates, the wood engineering field applies the censored Weibull MLE approach (ASTM, 2004a), which can be viewed as a semi-parametric approach.

The basic idea of this censored Weibull approach is that the quantile can be sufficiently determined by the density below it. Thus we can focus on fitting a parametric model to the left tail and subjectively censor the rest observations. Our simulations and theoretical study yield that this censored Weibull MLE strikes a good balance between the variance and bias, with a smaller mean squared error (MSE) than the parametric or non-parametric quantile estimates.

When compared to the empirical quantile, the censored Weibull MLE is slightly more biased but remarkably more efficient. This serves as an example that we can improve the empirical methods by finding a proper

parametric approximation to the model. When compared to the parametric quantile estimate, the censored Weibull MLE is less efficient under the correct, namely, the Weibull model, but much less biased under the misspecified models. A small part of the efficiency of the parametric model is sacrificed to achieve the accuracy in this censored Weibull MLE approach.

However, the censored Weibull MLE is not perfect and still has room to be improved. In this thesis, we first consider fitting a more complex data to the model, which enables us to utilize more information from the data. In this way, the Weibull mixture and censored Weibull mixture are proposed. According to our simulations, the censored Weibull mixture reduces the bias of the censored Weibull MLE, but we still have to find a way to select a better censoring threshold for the censored Weibull mixture.

The other way to improve the censored Weibull MLE is to select a better censoring threshold based on the data set, which leads us to bootstrap the real data set and achieve an MSE estimate based on the bootstrap samples. For a given finite set of threshold candidates, we can choose the threshold with the smallest bootstrap MSE estimate, which leads us to the censored Weibull MLE with bootstrap threshold selection (BMLE). Although, the bootstrap threshold selection procedure introduces additional variance to the quantile estimate, which is a common issue of a data-driven procedure, the BMLE generally performs better than all the other methods mentioned in this thesis.

As for future research, we will first prove the consistency of bootstrap estimate of the MSE and the consistency of the threshold selection based on the bootstrap MSE estimate, which may provide deeper insight into the performance of the BMLE. Also, we plan to explore the censored Weibull mixture with bootstrap threshold selection to further improve the quantile estimate.

Bibliography

ASTM (2004a). *American Society for Testing Materials: Standard specification for computing reference resistance of wood-based materials and structural connections for load and resistance factor design D5457.* ASTM, Philadephia, Pa.

ASTM (2004b). *American Society for Testing Materials: Standard specification for establishing and monitoring structural capacities of prefabricated wood I-Joists D5505.* ASTM, Philadephia, Pa.

Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, 68(1):326–328.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9(6):1196–1217.

Cantelli, F. P. (1933). Sulla determinazione empirica delle leggi di probabilita. *Giorn. Ist. Ital. Attuari*, 4:421–424.

Casella, G. and Berger, R. L. (2002). *Statistical inference.* Thomson Learning.

Charnigo, R. and Sun, J. (2004). Testing homogeneity in a mixture distribution via the l distance between competing models. *Journal of the American Statistical Association*, 99(466):488–498.

Chen, H., Chen, J., and Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(1):19–29.

Chen, J. and Li, P. (2009). Hypothesis test for normal mixture models: The em approach. *The Annals of Statistics*, 37:25232542.

Chen, J., Tan, X., and Zhang, R. (2008). Inference for normal mixtures in mean and variance. *Statist. Sinica.*, 18:443465.

Cheng, Y. (2010). Wood property relationships and survival models in reliability. Master's thesis, The University of British Columbia.

Clarke, B. R. (1983). Uniqueness and frechet differentiability of functional solutions to maximum likelihood type equations. *The Annals of Statistics*, 11(4):1196–1205.

Clarke, B. R. (1986). Nonsmooth analysis and frehet differentiability of m-functionals. *Probability Theory and Related Fields*, 73:197–209. 10.1007/BF00339936.

D'Agostino, R. B. and Stephens, M. A. (1986). *Goodness-of-fit techniques.* New York : M. Dekker.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38.

Durrans, S., Triche, M. H., and Suddarth, S. (1998). Estimation of lower tail quantiles of weibull probability distributions for lumber strength. *Forest Products Journal*, 48(1):96–101.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):pp. 1–26.

Glivenko, V. (1933). Sulla determinazione empirica della legge di probabilita. *Giorn. Ist. Ital. Attuari*, 4:92–99.

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, 13(2):795–800.

He, X. and Shao, Q.-M. (1996). A general bahadur representation of m-estimators and its application to linear regression with nonstochastic designs. *The Annals of Statistics*, 24(6):2608–2630.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174.

Huber, P. J. (1981). *Robust Statistics.* New York: John Wiley and Sons.

Hyndman, R. J. and Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician*, 50(4):361–365.

Izenman, A. J. (1991). Recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, 86(413):205–224.

Johnson, R. A., Evans, J. W., and Green, D. W. (2003). Confidence intervals for predicting lumber strength properties based on ratios of percentiles from two weibull populations. *USDA Forest Services Research Paper*, FPL-RP:606.

Jones, M. C. (1992). Estimating densities, quantiles, quantile densities and density quantiles. *Annals of the Institute of Statistical Mathematics*, 44(4):721–727.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433):401–407.

Jureckova, J. and Sen, P. K. (1987). A second-order asymptotic distributional representation of m-estimators with discontinuous score functions. *The Annals of Probability*, 15(2):814–823.

Lawless, J. F. (1982). *Statistical models and methods for lifetime data*. Hoboken, N.J. : Wiley-Interscience.

Lawless, J. F. (2003). *Statistical models and methods for lifetime data*. Hoboken, N.J. : Wiley-Interscience.

Li, P., Chen, J., and Marriott, P. (2009). Non-finite fisher information and homogeneity: an em approach. *Biometrika*, 96(2):411–426.

Lio, Y. L. and Padgett, W. J. (1991). A note on the asymptotically optimal bandwidth for nadaraya's quantile estimator. *Statistics & Probability Letters*, 11(3):243–249.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36(3):318–324.

Parzen, E. (1979). Nonparametric statistical data modeling. *Journal of the American Statistical Association*, 74(365):105–121.

Rinne, H. (2009). *The Weibull distribution : a handbook*. CRC Press.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. New York : Wiley.

Shao, J. (1993). Differentiability of statistical functionals and consistency of the jackknife. *The Annals of Statistics*, 21(1):61–75.

Shao, J. and Tu, D. (1995). *The jackknife and bootstrap.* Springer Verlag, New York.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):683–690.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, New York.

Warren, W. (1973). Estimation of the exclusion limit for dimension lumber. In *IUFRO Division V meeting, South Africa, Vol. 2: 1148.1154.*

Weibull, W. (1951). A statistical distribution function of wide application. *Journal of applied mechanics*, 18:293–297.

Wolfe, J. H. (1971). A monte carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions. Technical report, Technical Bulletin STB 72-2, Naval Personnel and Training Research Laboratory, San Diego.

Wu, C. F. J. (1983). On the convergence properties of the em algorithm. *The Annals of Statistics*, 11(1):pp. 95–103.

Appendix A

Supporting Materials

A.1   Introduction of some other distributions

Here are the PDF and some brief introduction about the distributions other than Weibull used in our study.

- Log-normal distribution is the exponential transformation of normal distribution, which means that the logarithm of a log-normal random variable will be normally distribution. Its PDF is,

$$f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log(x) - \mu)^2}{2\sigma^2}\right), x \geq 0. \qquad (A.1)$$

- Gamma distribution is also parameterized by a shape parameter k and a scale parameter s,

$$f(x; k, s) = \frac{x^{k-1}}{\Gamma(k)s^k} \exp\left(-\frac{x}{s}\right), x \geq 0. \qquad (A.2)$$

- Gumbel distribution (maximum or minimum) is the type I extreme value distribution introduced by Frechét (1927) and Fisher and Tippet (1928). Only the Gumbel distribution of the minimum is considered in our simulation. It is necessary to point out that people usually treat the maximum type Gumbel as the default Gumbel distribution.

So we will specify the Gumbel distribution we use here as "minimum Gumbel", whose PDF is:

$$f(x; a, b) = \frac{1}{b} \exp\left(\frac{x-a}{b} - \exp\left(\frac{x-a}{b}\right)\right) \tag{A.3}$$

A.2 Derivation of the asymptotic variance of censored Weibull MLE

The score function for shape parameter α is

$$\psi_\alpha(x) = \begin{cases} \dfrac{\partial \log(f(x))}{\partial \alpha} = \dfrac{1}{\alpha} + \log(x) - \log(\eta) - \left(\dfrac{x}{\eta}\right)^\alpha \log\left(\dfrac{x}{\eta}\right) & , x \leq C \\[3mm] \dfrac{\partial \log(1 - F(C))}{\partial \alpha} = -\left(\dfrac{C}{\eta}\right)^\alpha \log\left(\dfrac{C}{\eta}\right) & , x > C \end{cases}.$$
$$\tag{A.4}$$

Similarly the score function for the scale parameter η is

$$\psi_\eta(x) = \begin{cases} \dfrac{\partial \log(f(x))}{\partial \eta} = -\dfrac{\alpha}{\eta} + \dfrac{\alpha x^\alpha}{\eta^{\alpha+1}} & , x \leq C \\[3mm] \dfrac{\partial \log(1 - F(C))}{\partial \eta} = \dfrac{\alpha C^\alpha}{\eta^{\alpha+1}} & , x > C \end{cases}. \tag{A.5}$$

As the information matrix

$$I(\boldsymbol{\theta}) = -E_G\left(\frac{\partial \psi}{\partial \boldsymbol{\theta}}\right)$$

we need to differentiate the score function against the parameters, which are:

$$\frac{\partial \psi_\alpha(x)}{\partial \alpha} = \begin{cases} -\dfrac{1}{\alpha^2} - \left(\dfrac{x}{\eta}\right)^\alpha \log^2\left(\dfrac{x}{\eta}\right) & , x \leq C \\[3mm] -\left(\dfrac{C}{\eta}\right)^\alpha \log^2\left(\dfrac{C}{\eta}\right) & , x > C \end{cases}, \tag{A.6}$$

$$\frac{\partial \psi_\eta(x)}{\partial \eta} = \begin{cases} \dfrac{\alpha}{\eta^2} - \dfrac{\alpha(\alpha+1)x^\alpha}{\eta^{\alpha+2}} & , x \le C \\[2ex] -\dfrac{\alpha(\alpha+1)C^\alpha}{\eta^{\alpha+2}} & , x > C \end{cases} \tag{A.7}$$

and

$$\frac{\partial \psi_\alpha(x)}{\partial \eta} = \frac{\partial \psi_\eta(x)}{\partial \alpha} = \begin{cases} -\dfrac{1}{\eta} + \dfrac{x^\alpha}{\eta^{\alpha+1}} + \dfrac{\alpha x^\alpha}{\eta^{\alpha+1}} & , x \le C \\[2ex] \dfrac{C^\alpha}{\eta^{\alpha+1}} + \dfrac{\alpha C^\alpha}{\eta^{\alpha+1}} \log\left(\dfrac{C}{\eta}\right) & , x > C \end{cases}. \tag{A.8}$$

Then take the expectation of (A.6), (A.7) and (A.8), we will have

$$-E_G \frac{\partial \psi_\alpha(x)}{\partial \alpha} = \int_0^C \left[\frac{1}{\alpha^2} + \left(\frac{x}{\eta}\right)^\alpha \log^2\left(\frac{x}{\eta}\right) \right] f(x; \alpha, \eta) dx + \left(\frac{C}{\eta}\right)^\alpha \log^2\left(\frac{C}{\eta}\right) \int_C^\infty g(x) dx$$

$$= \int_0^C \left[\frac{1}{\alpha^2} + \left(\frac{x}{\eta}\right)^\alpha \log^2\left(\frac{x}{\eta}\right) \right] f(x; \alpha, \eta) dx + \left(\frac{C}{\eta}\right)^\alpha \log^2\left(\frac{C}{\eta}\right) \int_C^\infty f(x; \alpha, \eta) dx$$

$$= \int_0^C \left(\frac{1}{\alpha^2} + \left(\frac{x}{\eta}\right)^\alpha \log^2\left(\frac{x}{\eta}\right) \right) \frac{\alpha x^{\alpha-1}}{\eta^\alpha} \exp\left(-\left(\frac{x}{\eta}\right)^\alpha\right) dx$$

$$+ \left(\frac{C}{\eta}\right)^\alpha \log^2\left(\frac{C}{\eta}\right) \exp\left(-\left(\frac{C}{\eta}\right)^\alpha\right) \tag{A.9}$$

Similarly, we have,

$$-E_G \frac{\partial \psi_\eta(x)}{\partial \eta} = \frac{\alpha^2}{\eta^2} \left(1 - \exp\left(-\left(\frac{C}{\eta}\right)^\alpha\right) \right) \tag{A.10}$$

$$-E_G \frac{\partial \psi_\alpha(x)}{\partial \eta} = \int_0^C \left(\frac{1}{\eta} - \frac{x^\alpha}{\eta^{\alpha+1}} - \frac{\alpha x^\alpha}{\eta^{\alpha+1}} \right) \frac{\alpha x^{\alpha-1}}{\eta^\alpha} \exp\left(-\left(\frac{x}{\eta}\right)^\alpha\right) dx$$

$$+ \left(\frac{C^\alpha}{\eta^{\alpha+1}} + \frac{\alpha C^\alpha}{\eta^{\alpha+1}} \log\left(\frac{C}{\eta}\right) \right) \exp\left(-\left(\frac{C}{\eta}\right)^\alpha\right) \tag{A.11}$$

In practice, the evaluation of these integrals require numeric techniques, but the "integrate" function can handle them easily.