Lower quantile estimation with censored Weibull MLE

Yang Liu (Seagle)

Department of Statistics, UBC

March 17, 2012

Acknowledgements

- Supervised by Matías Salibián-Barrera and Ruben H. Zamar
- A project in the joint research program among SFU, Forrest Products Innovation and UBC
- Research group led by Jim Zidek and William Welch

Overview

- Background and motivation
- Problems of empirical quantile and parametric quantile estimates
- Semi-parametric model and censored Weibull MLE
- Simulation comparison
- Discussion

Background and Motivation

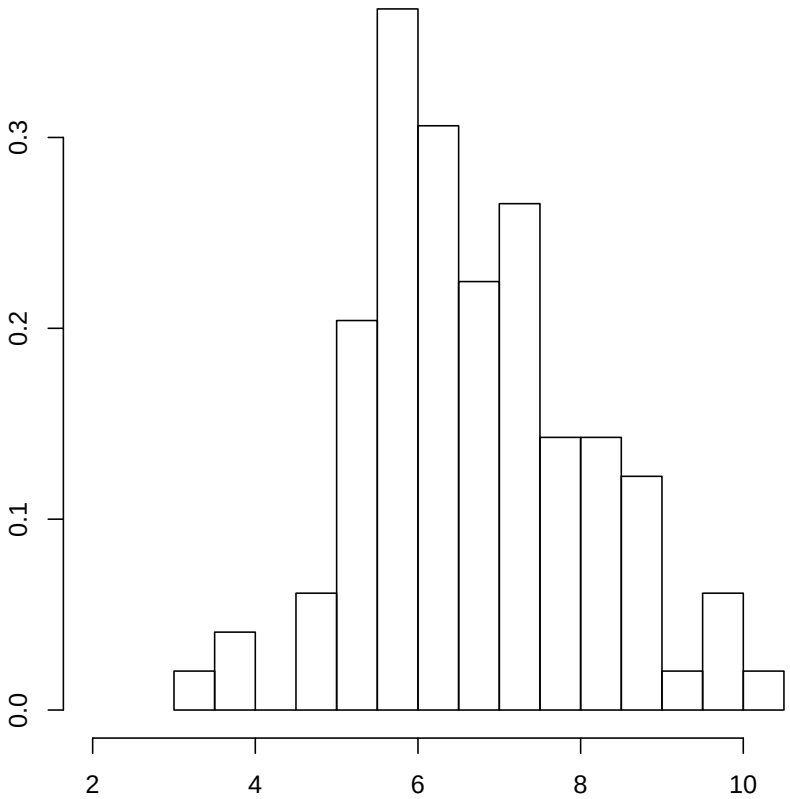
Importance of Lower Quantile Estimate

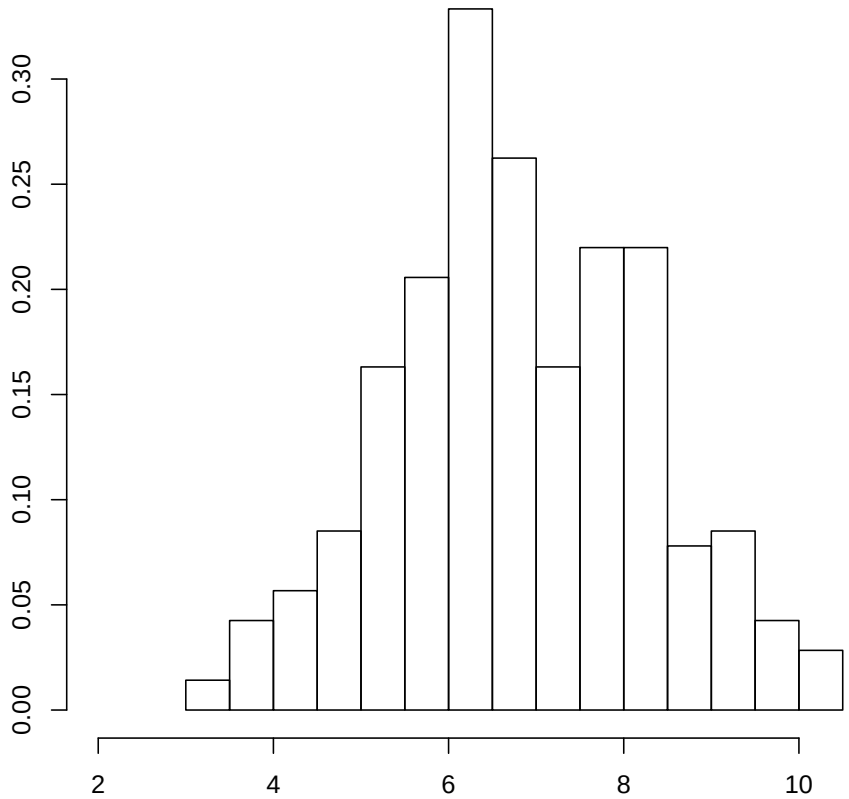
Importance of Lower Quantile Estimate

How much can we load on those wood board?

- Randomness in the wood materials
- Need to find a "Design Value" of the wood boards
- Randomly sample 100-300 wood boards and test their strength (modules of rupture MOR)
- 5% quantile: the probability of failure will be smaller than 5% if the load is not over it

Data


0.0
0.1
0.2
0.3
2
4
6
8
10
MOR (Size: 98)


0.00
0.05
0.10
0.15
0.20
0.25
0.30
2
4
6
8
10
MOR2 (Size: 282)

How to evaluate the quantile estimate?

- Accuracy: almost unbiased estimate
- Efficiency: small variance in for a sample of $100 - 300$
- The quantile q, its estimate $\tilde{q}_n$ and model G

Mean Squared Error

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

Non-parametric and Parametric Quantile Estimates

Non-parametric: Empirical Quantile

- Empirical CDF:

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

- Empirical Quantile:

$$\hat{q}_n = \hat{F}_n^{-1}(p) = \inf\{x : \hat{F}_n(x) \geq p\}$$

- Better empirical quantile by linear interpolation between two order statistics $p_i = \frac{i-a}{n+b}$

$$a = 1/2, b = 0;\ a = b = \tfrac{1}{3};\ a = \tfrac{3}{8}, b = 0.25;$$

- Based on our simulations, Type IX in R ($a = \frac{3}{8}, b = 0.25$) is chosen.

Non-parametric: Kernel Quantile

Inverse of kernel density estimate

- Kernel density estimate:

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

- $\tilde{K}_b(x) = \int_{-\infty}^{x} \tilde{k}_b(u)du$ and then inverse it numerically

Asymptotic Variance and Problem

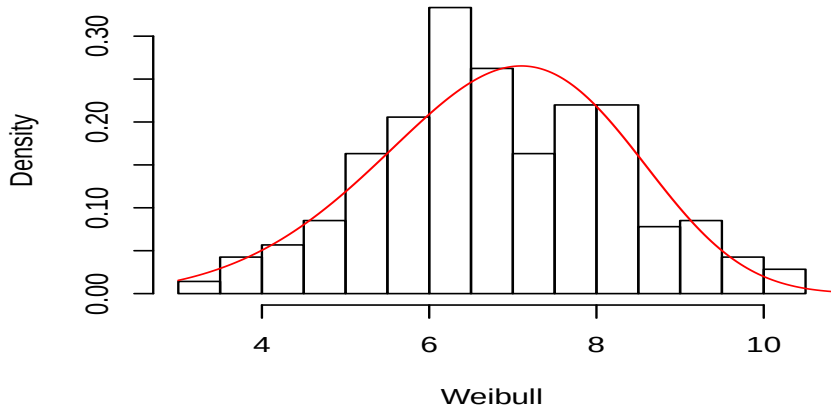
$$\sqrt{n}(\hat{q}_n - q) \rightarrow_d p(1-p)/f^2(q)$$

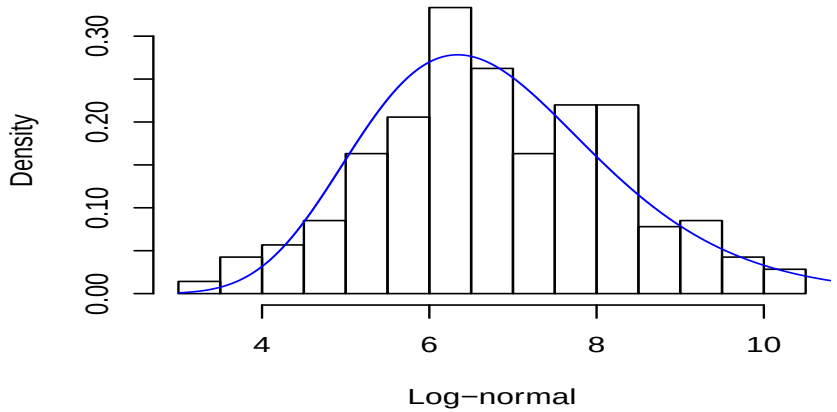
Need a large data set to achieve a good quantile estimate

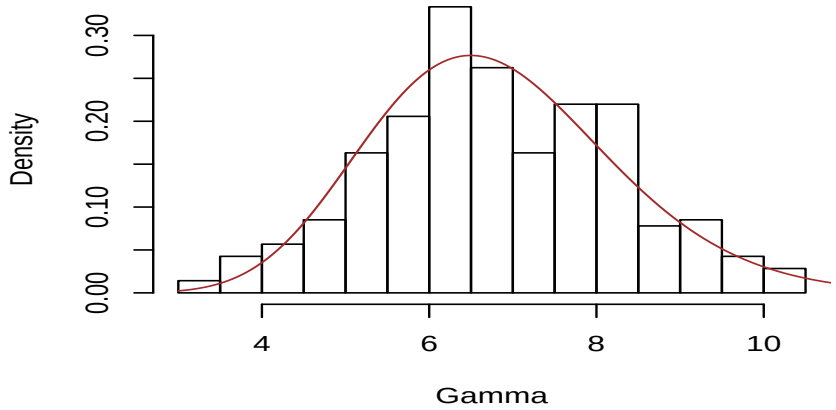
Parametric Quantile Estimate: MLE

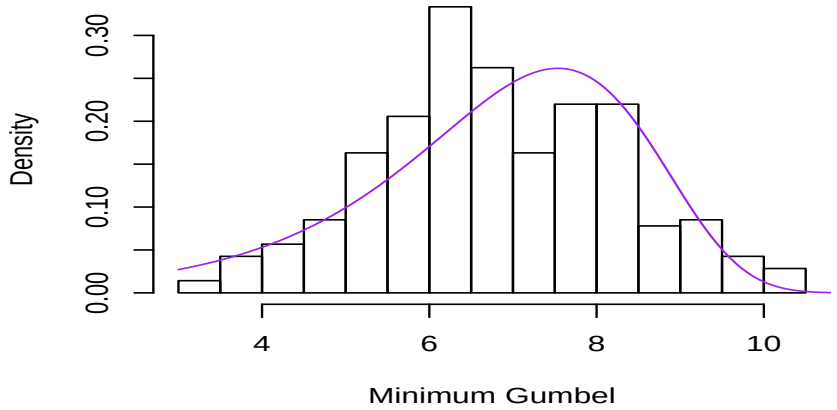
Fit a parametric distribution to the data by MLE

Fully efficient but problematic under mis-specified models









Semi-parametric: Censored Weibull MLE

A Semi-parametric Tail Model

- Unnecessary to have a complete density estimate
- Quantile can be sufficiently decided by the density below it

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

- $f(x;\boldsymbol{\theta})$: the parametric tail
- X_0: changing point
- $h(x)$: unspecified non-parametric part
  1. $h(x)\geq 0, x>X_0$
  2. $\int_{X_0}^{\infty} h(x)dx = 1-G(X_0) = 1-F(X_0,\boldsymbol{\theta})$.

Subjective Censoring

- It is impossible to obtain $\tilde{X}_0$ by MLE
- When $q < X_0$, there is no need to estimate X_0
- To estimate θ without X_0:

> Focus on the left tail of the data and subjectively censor the rest

- Choose a threshold C, larger than the quantile of interest and supposedly smaller than X_0
- The corresponding likelihood:

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

$$\delta_i = \mathbf{1}\{X_i \leq C\}$$

Consistency of Subjective Censoring

- Treat it as an M-estimator and its score function

$$\psi(x) = \begin{cases} \dfrac{\partial \log(f(x))}{\partial \boldsymbol{\theta}}, & x \leq C \\ \dfrac{\partial \log(1 - F(C))}{\partial \boldsymbol{\theta}}, & x > C \end{cases},$$

- When the distribution family of $f(x; \boldsymbol{\theta})$ is known and $C \leq X_0$:

$$E_G \psi(x) = \mathbf{0}$$

- Unbiased estimating equation leads to consistent parameter estimate and thus consistent quantile estimate

Choice of the Parametric Tail: Weibull

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

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

- Industrial custom: Standard 5457-04a, censored Weibull with C as the 10% empirical quantile
- Statistical reason:
  - Type III minimum value distribution: limiting distribution for the minimum of distributions bounded on the left
  - Flexible to approximate the left tails of other distributions

Calculation of Censored Weibull MLE

For convenience, $X_1, X_2, \ldots, X_m \leq C$ and $X_{m+1}, \ldots, X_n$ are censored

- Solving score functions equal 0:

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

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

- An iteration algorithm, slower but more robust than Newton Raphson:

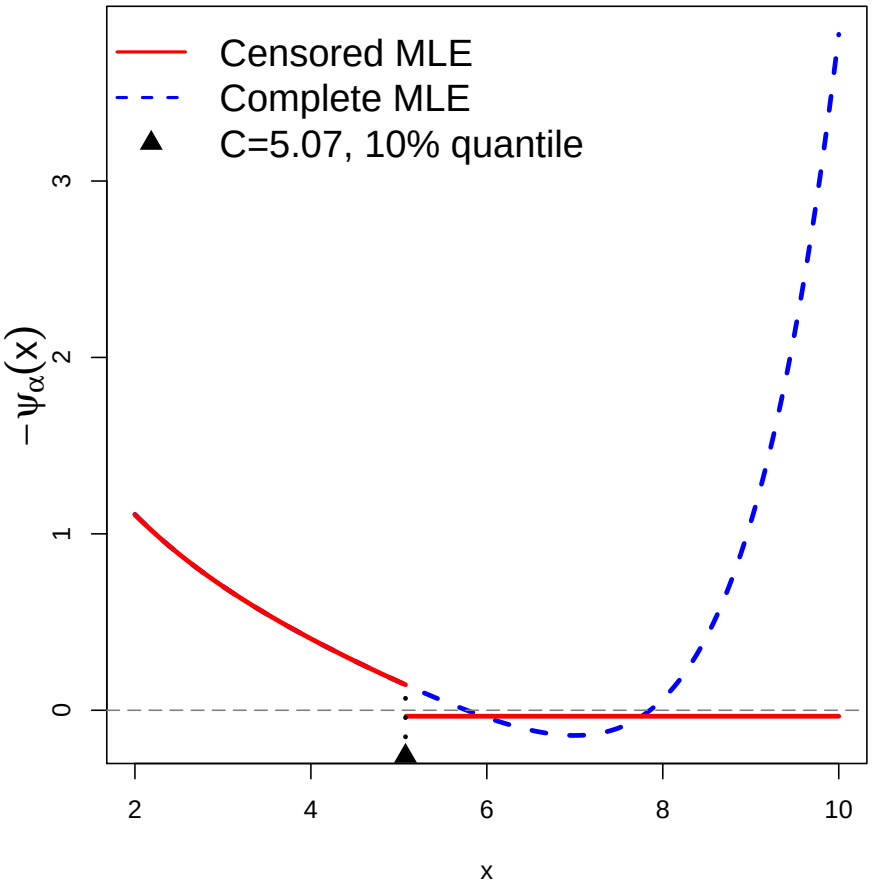
$$\frac{1}{\alpha} = \frac{\sum_{i=1}^{m} X_i^{\alpha} \ln(X_i) + (n-m) C^{\alpha} \ln(C)}{\sum_{i=1}^{m} X_i^{\alpha} + (n-m) C^{\alpha}} - \frac{\sum_{i=1}^{m} \ln(X_i)}{m}$$

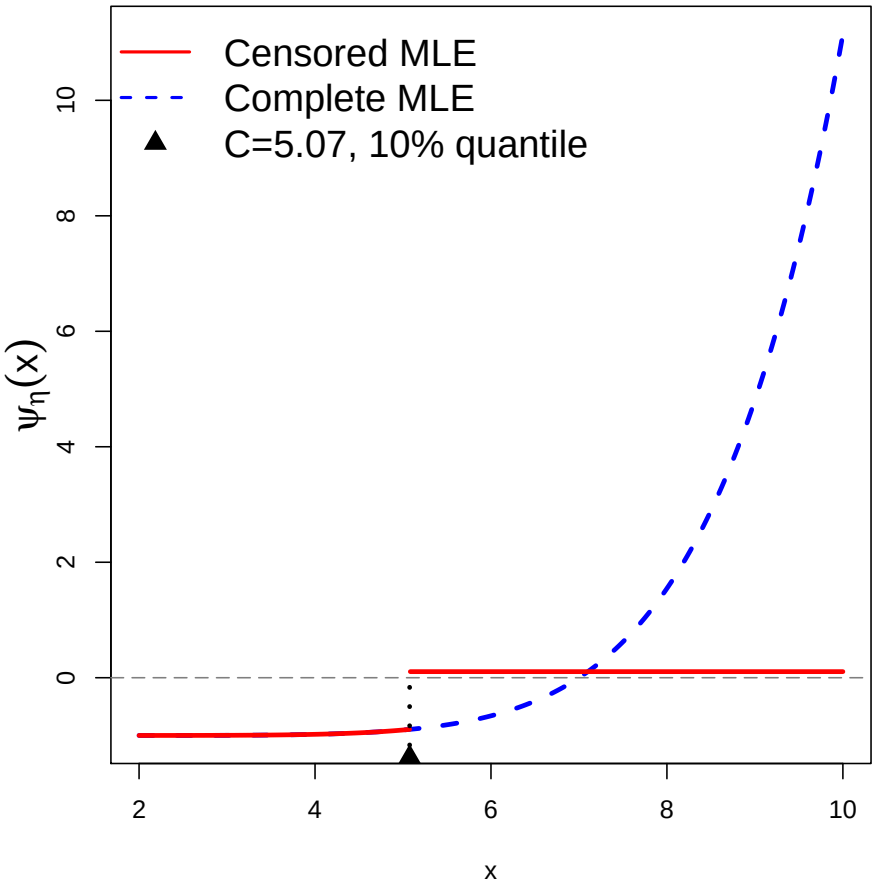
- $\tilde{q}_n = \tilde{\eta}_n \left[-\log(1-p)\right]^{1/\tilde{\alpha}_n}$

How Censoring Works

via Influence (Score) function, Weibull(7, 7)

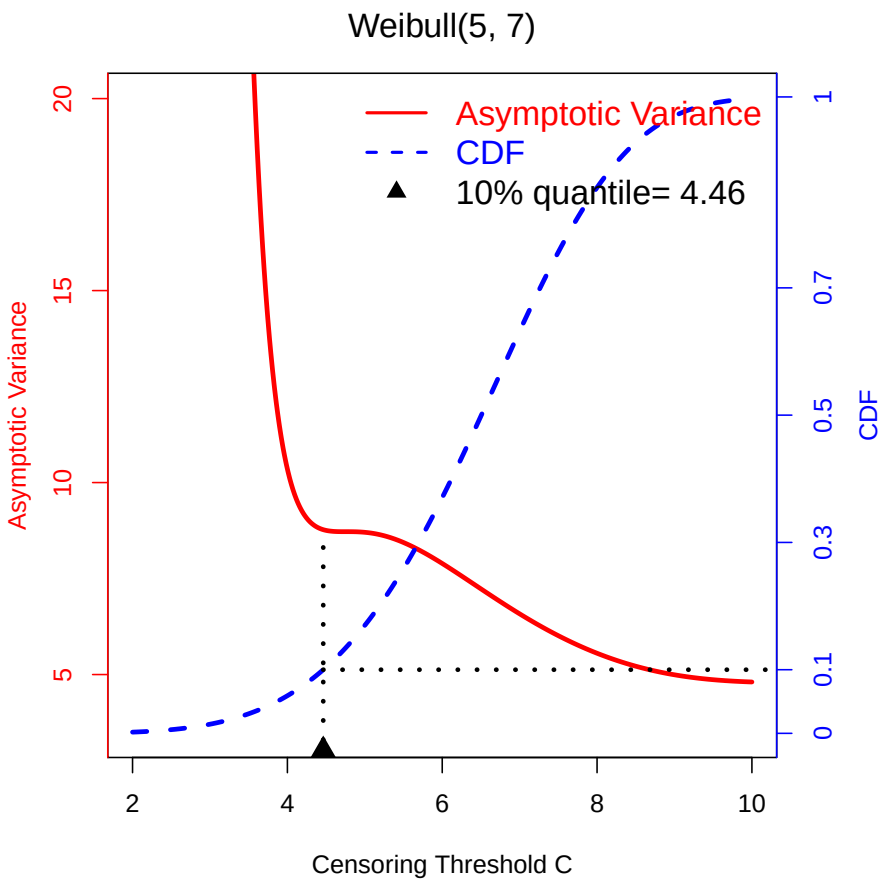
Reduce the influence of upper tail and focus on the left tail:

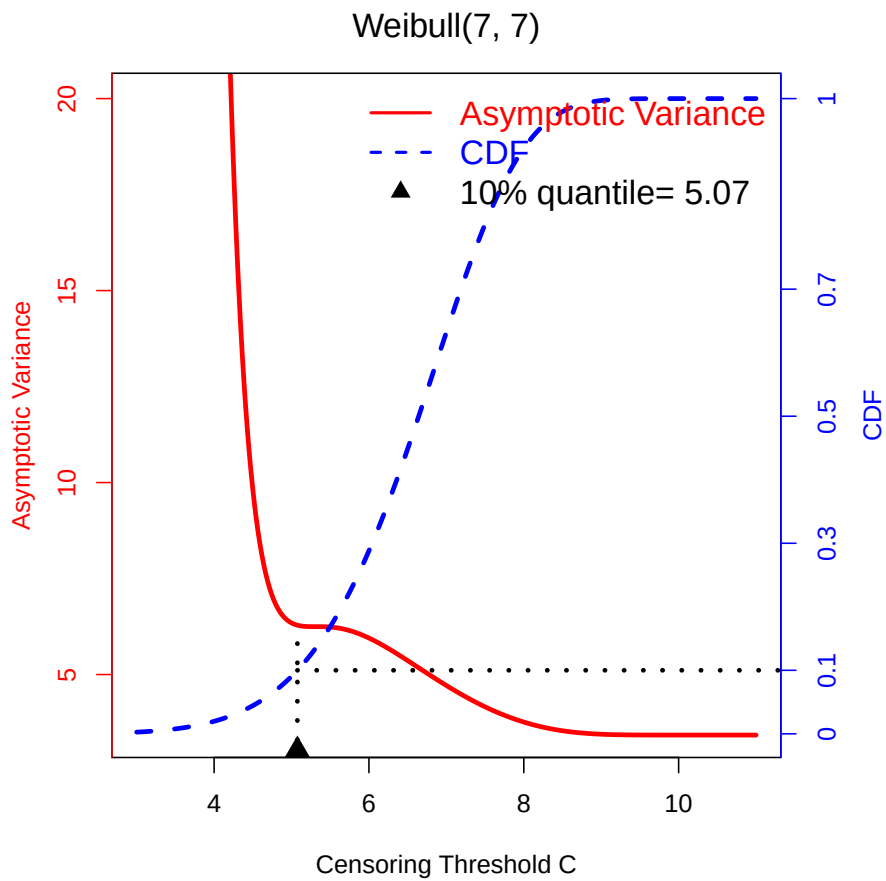




Efficiency Lost in Censoring

Asymptotic Variance of $\sqrt{n}(\tilde{q}_n - q)$





Accuracy Gained by Censoring

Goodness-of-fit in the left tail

- Truncated Kolmogorov-Smirnov statistic: $D^{\dagger}(\tilde{F}_n) = \sup_{x \leq C}\{|\tilde{F}_n(x) - G(x)|\}$
- Extremely difficult to derive its distribution
- Compare it to the same statistic achieved by empirical CDF and kernel density estimate

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

- Approximate the distribution of $d(\tilde{F}_n, \hat{F}_n)$ and $d(\tilde{F}_n, \tilde{K}_n)$ by simulation, to see $Pr(d(\tilde{F}_n, \hat{F}_n) < 0)$ and $Pr(d(\tilde{F}_n, \tilde{K}_n) < 0)$.

Simulation and Results

Model Setting: Parametric Models

Censored MLE on MOR2 with $C = 10\%$ empirical quantile

Four popular models in survival analysis:

Model	Parameter Setting	
Weibull	$\alpha = 7.378$	$\eta = 6.738$
Log-normal	$\mu = 1.976$	$\sigma = 0.2916$
Gamma	$k = 16.16$	$s = 0.4407$
Minimum Gumbel	$a = 6.315$	$b = 0.5997$

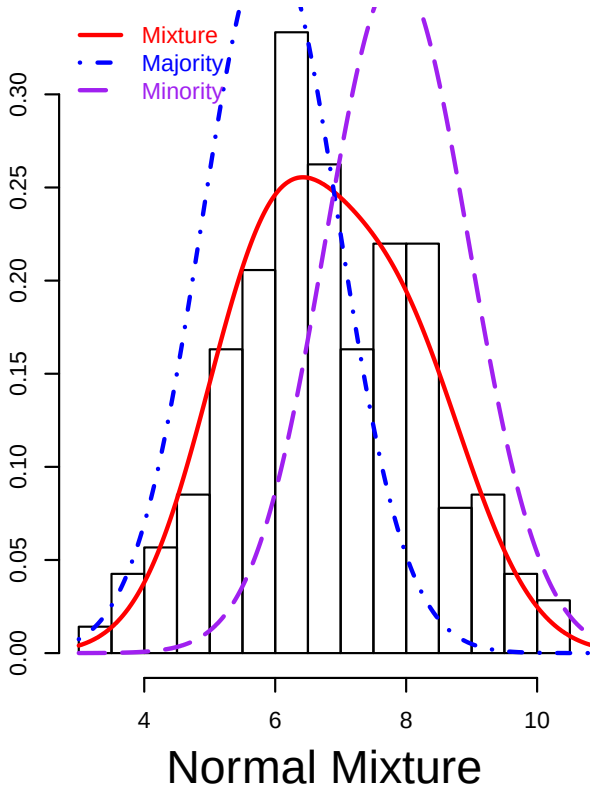
Gumbel distribution is the type I extreme value distribution, whose PDF is

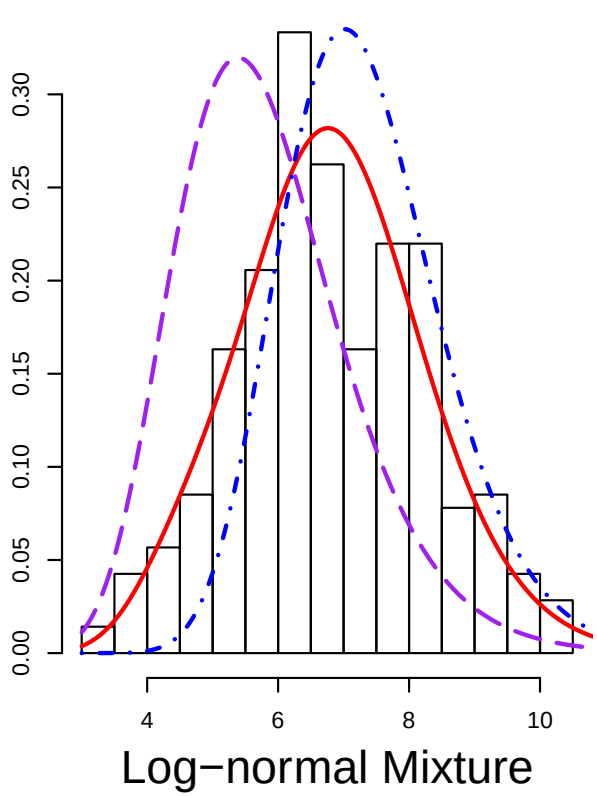
$$f(x; a, b) = \frac{1}{b} \exp\left(\frac{x-a}{b} - \exp\left(\frac{x-a}{b}\right)\right)$$

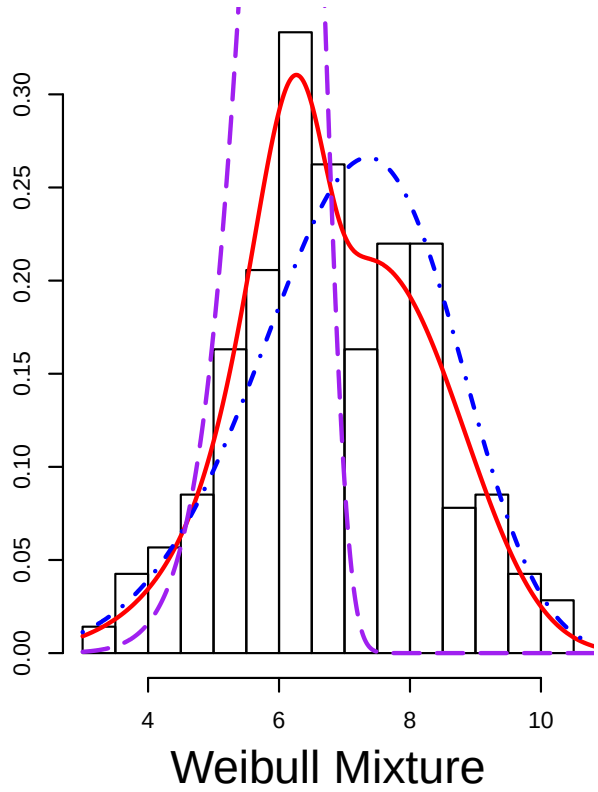
Model Setting: Mixture Models

Obtained by EM algorithm on MOR2 without censoring

Model	p	Majority Population		Minority Population	
Normal Mixture	0.5406	$\mu_1 = 5.924$	$\sigma_1 = 1.042$	$\mu_2 = 7.859$	$\sigma_2 = 1.095$
Log-normal Mixture	0.6649	$\mu_1 = 1.976$	$\sigma_1 = 0.167$	$\mu_2 = 1.736$	$\sigma_2 = 0.226$
Weibull Mixture	0.7932	$\alpha_1 = 5.427$	$\eta_1 = 7.642$	$\alpha_2 = 12.01$	$\eta_2 = 6.186$



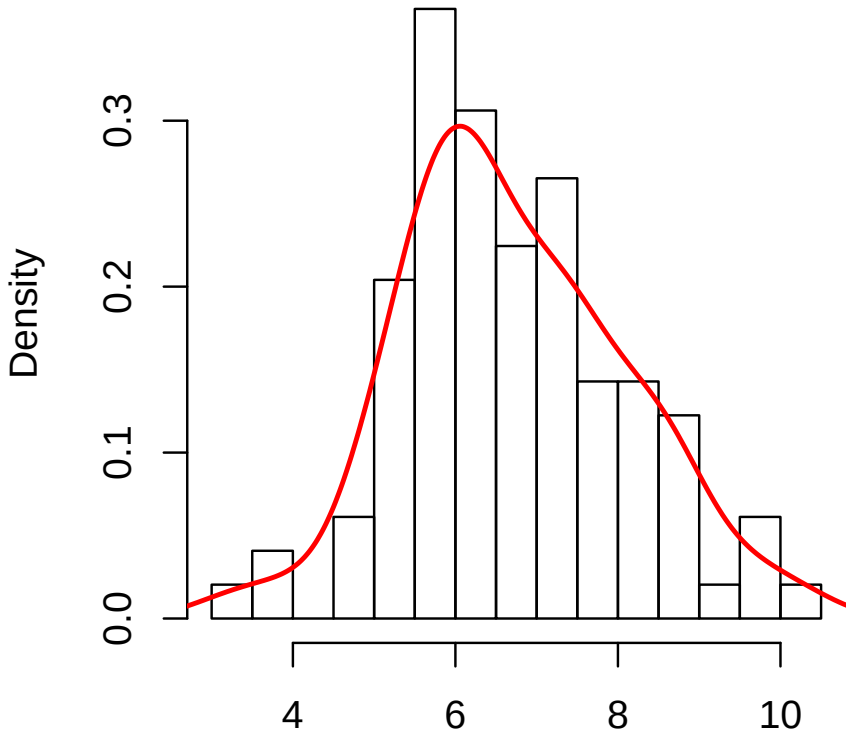


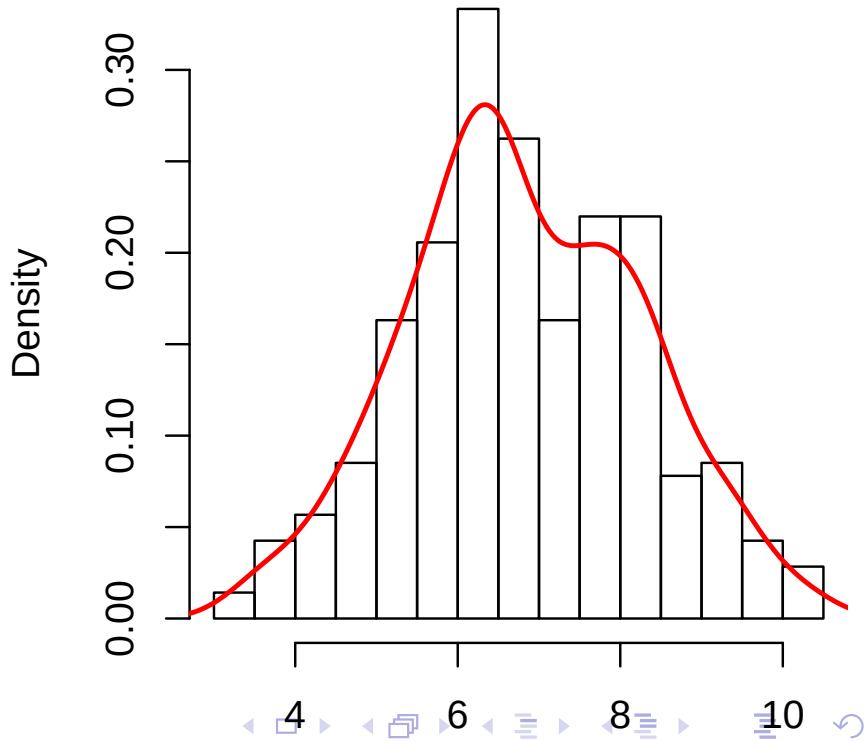


Model Setting: KDE Models

Simulate data from the kernel density estimate of MOR and MOR2

Data	Sample Size	Bandwidth
MOR	98	0.4909
MOR2	282	0.4056





Other Settings

- Sample Size
  - Same as the original sample size in the KDE models
  - 300 in all the others
- Replicates: 2500
- Quantile estimates to be compared:
  - Weibull MLE (MLE)
  - Censored Weibull MLE (CMLE)
  - Empirical Quantile (EMP, Type IX)
  - Kernel Quantile (KDE)

Results: Goodness-of-fit

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

Model	$Pr(d(\tilde{F}_n, \hat{F}_n) < 0)$	$Pr(d(\tilde{F}_n, \tilde{K}_n) < 0)$
Weibull	97.48%	50.36%
Log-normal	94.16%	66.64%
Gamma	96.00%	63.48%
Minimum Gumbel	95.56%	46.83%
Normal Mixture	97.20%	65.08%
Log-normal Mixture	94.00%	54.84%
Weibull Mixture	97.96%	60.96%
MOR2 (KDE Model)	73.44%	44.04%
MOR (KDE Model)	72.84%	48.80%

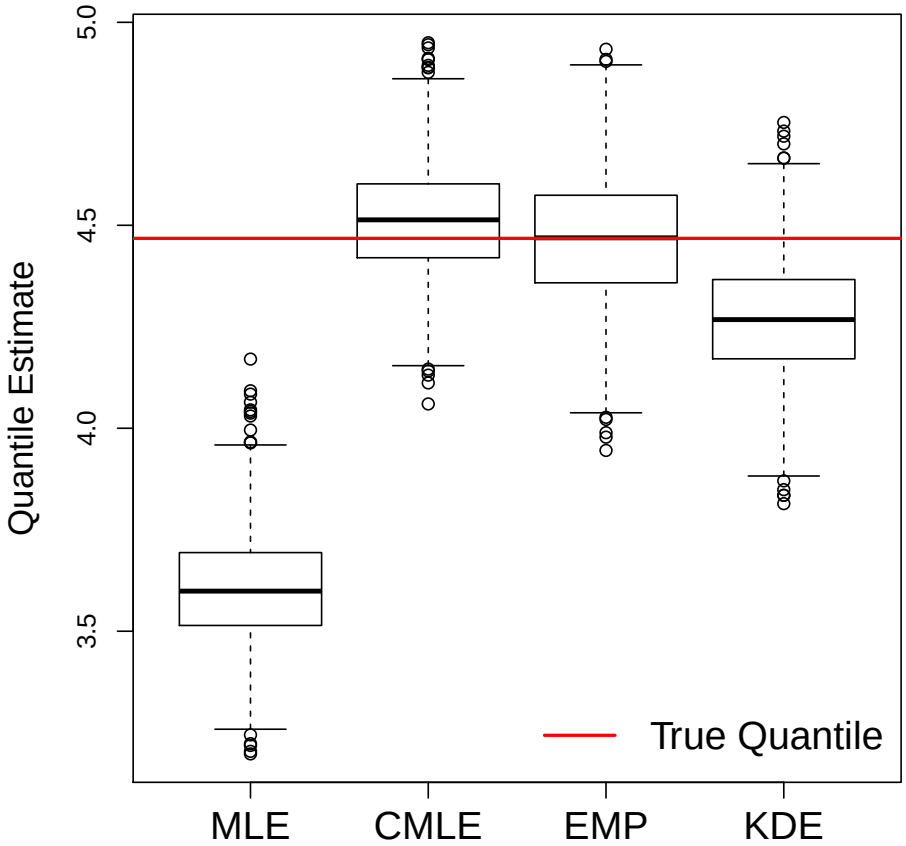
Results: Bias

Empirical Quantile $<$ Censored Weibull MLE $<$ KDE $<<$ Weibull MLE

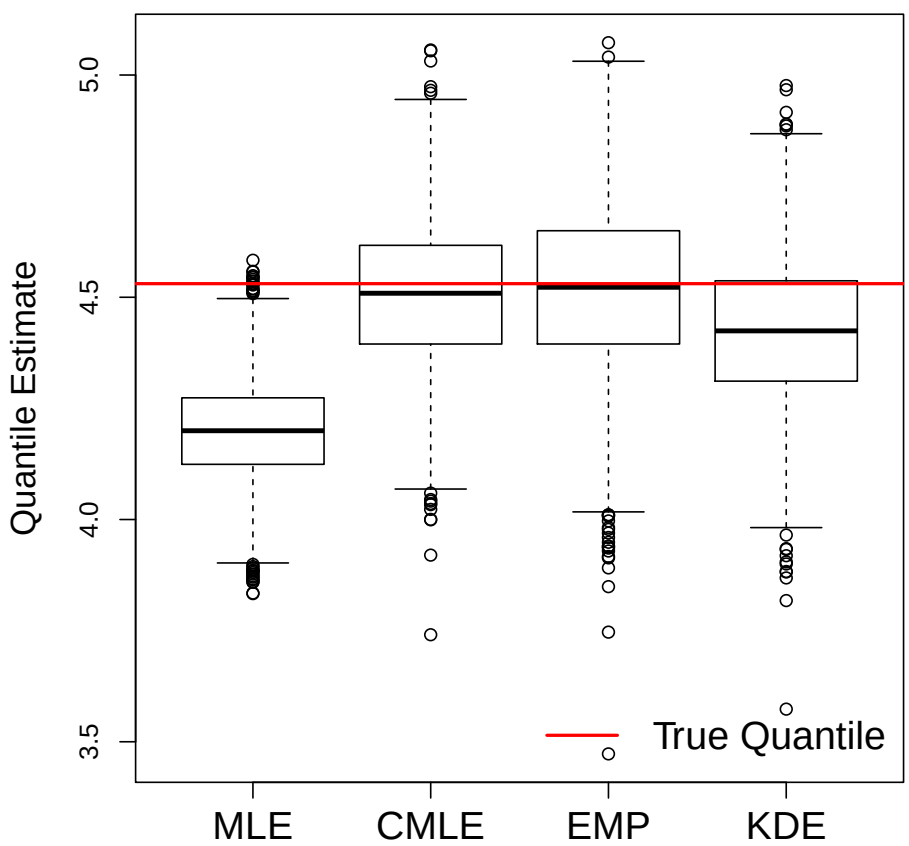
Model	MLE	CMLE	EMP	KDE
Weibull	7.17e-6	1.49e-5	8.50e-6	0.0043
Log-normal	0.7429	0.00215	2.26e-6	0.0390
Gamma	0.3503	0.00141	5.12e-7	0.0262
Minimum Gumbel	0.0190	0.00187	9.54e-5	0.0014
Normal Mixture	0.1237	0.00052	2.99e-6	0.0201
Log-normal Mixture	0.1251	0.00254	2.38e-6	0.0124
Weibull Mixture	0.1096	0.00057	1.42e-4	0.0119
MOR2 (KDE Model)	0.0678	0.00208	3.45e-7	0.0127
MOR (KDE Model)	0.2453	0.0019	0.0014	0.0487

Results

Log-normal and Mixture of Weibull



Log-normal



Weibull Mixture

Results: Variance

Weibull MLE $<$ Censored Weibull MLE $<$ KDE $<$ Empirical Quantile

Model	MLE	CMLE	EMP	KDE
Weibull	0.0010	0.0181	0.0242	0.0189
Log-normal	0.0195	0.0182	0.0250	0.0210
Gamma	0.0173	0.0182	0.0246	0.0204
Minimum Gumbel	0.0079	0.0216	0.024	0.0203
Normal Mixture	0.0128	0.0146	0.0203	0.0160
Log-normal Mixture	0.0149	0.0161	0.0225	0.0177
Weibull Mixture	0.0130	0.0272	0.0369	0.0288
MOR2 (KDE Model)	0.0139	0.0213	0.0331	0.0256
MOR (KDE Model)	0.0360	0.0717	0.1078	0.0780

Results: MSE

Censored Weibull MLE $<$ Empirical Quantile $<$ Weibull MLE and KDE

Model	MLE	CMLE	EMP	KDE
Weibull	0.0010	0.0181	0.0242	0.0231
Log-normal	0.7629	0.0204	0.0250	0.0599
Gamma	0.3676	0.0196	0.0246	0.0466
Minimum Gumbel	0.0269	0.0235	0.0241	0.0216
Normal Mixture	0.1364	0.0151	0.0203	0.0361
Log-normal Mixture	0.1400	0.0186	0.0225	0.0301
Weibull Mixture	0.1226	0.0278	0.037	0.0406
MOR2 (KDE Model)	0.0816	0.0233	0.0331	0.0383
MOR (KDE Model)	0.2813	0.0735	0.1092	0.1267

Advantage of Censored Weibull MLE

Balance between Accuracy and Efficiency

- Censored Weibull MLE is not the most efficient
- Nor it is not the most accurate
- The efficiency sacrificed by censoring enables us to focus on the left tail to gain accuracy

Something More

- The censoring threshold C is ad-hoc
- Estimate the MSE by bootstrap and then select the optimal threshold
- Consider more complex model for more accuracy, e.g. Weibull Mixture

Thank you!