Distributed Inference in Extreme Value Index

December 22, 2020

This talk is based on following paper.

- Liujun Chen, Deyuan Li, and Chen Zhou(2020). Distributed Inference for Extreme Value Index.

Table of Contents

Motivation

Why should we do distributed Inference?

- Restriction on data sharing, e.g. GDPR, CCPA
- Datasets may be stored in multiple machines.

Divide and Conquer Algorithm

How to do distributed inference:

- Distributed Optimization.

- Divide and Conquer Algorithm.
  - Estimate a desired quantity or parameter on each machine.
  - Transmit the results to a central machine.
  - The central machine combines all the result, often by a simple averaging, to obtain a computationally feasible estimator.

Oracle Property of Distributed Inference

Definition (Oracle Property)

Speed of convergence and asymptotic distribution coincides with the oracle estimator when applying the same statistical procedure to the hypothetically combined dataset.

- For a broader set of statistical procedures, under mild conditions, the divide and conquer algorithm possesses the oracle property.
- Nevertheless, the oracle property may not hold for some specific statistical methods, or requires additional conditions. e.g. quantile regression (see e.g. Volgushev et al. (2019))

Distributed Inference for Extremes

- Extreme value analysis focuses on statistical inference regarding the tail of a distribution.
- Similar to quantile regression, the oracle property of a standard DC algorithm based on extreme value methods is not guaranteed by the general theory in distributed inference.
- For example, considering a distribution with a finite endpoint, the standard DC algorithm based on averaging fails.

Table of Contents

Model Setting

- Consider a distribution function $F \in D(G_\gamma)$ with $\gamma > 0$ (heavy tailed distribution).
- This is equivalent to $U := (1/(1-F))^{\leftarrow}(t)$ is a regular varying function:

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^\gamma.$$

- A key question in extreme value analysis is to estimate the extreme value index γ.

Model Setting

- Assume that the i.i.d. observations $X_1, \ldots, X_n$ are stored in k machines with m observations each and $n = mk$.
- Assume that $m \to \infty, k \to \infty$ as $n \to \infty$
- Assume that only one result can be transmitted from each machine to the central machine.
- Practically, we cannot apply statistical procedures to the oracle sample, i.e. the hypothetically combined dataset $\{X_1, \ldots, X_n\}$.

Oracle Hill estimator

If we can use the oracle sample, he oracle Hill estimator is defined as

$$\hat{\gamma}_H := \frac{1}{l}\sum_{i=1}^{l}(\log M^{(i)} - \log M^{(l+1)}),$$

where

$$l = l(n) \to \infty, l/n \to 0 \text{ as } n \to \infty.$$

Note that, the oracle Hill estimator involves the top $l+1$ highest order statistics, or in other words, top l exceedance ratios $M^{(i)}/M^{(l+1)}$.

Distributed Hill estimator

Following a divide and conquer Algorithm,

- Apply the Hill estimator at each machine

$$\hat{\gamma}_j := \frac{1}{d_j} \sum_{i=1}^{d_j} \left(\log M_j^{(i)} - \log M_j^{(l+1)} \right),$$

  where $M_j^{(1)} \geq \cdots \geq M_j^{(m)}$ denote the order statistics within the machine j.

- Take the average of the Hill estimates from all machines

$$\hat{\gamma}_{DH} := \frac{1}{k} \sum_{j=1}^{k} \hat{\gamma}_j.$$

Table of Contents

Conditions

(A.) $k = k(n) \to \infty, m = m(n) \to \infty$ and $m/\log k \to \infty$ as $n \to \infty$

(B.) There exist an eventually positive or negative function A with $\lim_{t\to\infty} A(t) = 0$ and a real number $\rho \leq 0$ such that

$$\lim_{t\to\infty} \frac{\frac{U(tx)}{U(t)} - x^{\gamma}}{A(t)} = x^{\gamma} \frac{x^{\rho} - 1}{\rho},$$

for all $x > 0$.

Asymptotic properties in IID case

- Homogenous case: $d_1 = \cdots = d_k = d < \infty$
- Heterogeneous case: d_j are different but uniformly bounded $\sup_{j \in \mathbb{N}} d_j < \infty$
- Homogenous case: $d_1 = \cdots = d_k = d \to \infty$

Homogeneous case and $d < \infty$

Theorem

Suppose $F \in D(G_\gamma)$ with $\gamma > 0$ and Conditions A and B hold. Let $d_1 = d_2 = \cdots = d_k = d$, where $d \geq 1$ is a fixed integer. If $\sqrt{kd}A(m/d) = O(1)$ as $n \to \infty$, then

$$\sqrt{kd}\left(\hat{\gamma}_{DH} - \gamma - A(m/d)g(d, m, \rho)\right) \xrightarrow{d} N(0, \gamma^2),$$

where

$$g(d, m, \rho) = \frac{1}{1-\rho}\left(\frac{m}{d}\right)^{-\rho} \frac{\Gamma(m+1)\Gamma(d-\rho+1)}{\Gamma(m-\rho+1)\Gamma(d+1)}.$$

Oracle property

- Assume that $\sqrt{kd}A(n/(kd) = \sqrt{kd}A(m/d) \to \lambda \in \mathbb{R}$, the oracle Hill estimator possesses the asymptotic normality

$$\sqrt{kd}(\hat{\gamma}_H - \gamma) \xrightarrow{d} N(\lambda/(1-\rho), \gamma^2)$$

- Under the same condition, we have that

$$\sqrt{kd}(\hat{\gamma}_{DH} - \gamma) \xrightarrow{d} N\left(\lambda \frac{d^\rho}{1-\rho} \frac{\Gamma(d-\rho+1)}{\Gamma(d+1)}, \gamma^2\right)$$

Corollary

The oracle property holds only when $\rho = 0$ or $\lambda = 0$.

d_j are different but uniformly bounded

Theorem

Suppose $F \in D(G_\gamma)$ with $\gamma > 0$ and Conditions A and B hold. Let $d_1, d_2, \ldots, d_k$ be uniformly bounded positive integers. If $\sqrt{k}\bar{d}A(m/\bar{d}) = O(1)$ as $n \to \infty$, then

$$\sqrt{k}\bar{d}\left(\hat{\gamma}_{DH} - \gamma - A(m/\bar{d})\frac{1}{k}\sum_{j=1}^{k}\left(\frac{\bar{d}}{d_j}\right)^{\rho} g(d_j, m, \rho)\right) \xrightarrow{d} N(0, \gamma^2),$$

where $\bar{d} = k^{-1}\sum_{j=1}^{k} d_j$.

Oracle property

- Assume that $\sqrt{k\bar{d}}A(m/\bar{d}) \to \lambda$, the oracle Hill estimator possesses the asymptotic normality:

$$\sqrt{k\bar{d}}(\hat{\gamma}_H - \gamma) \xrightarrow{d} N(\lambda/(1-\rho), \gamma^2)$$

- The asymptotic bias for the distributed Hill estimator is

$$\frac{1}{k}\sum_{j=1}^{k}\left(\frac{\bar{d}}{d_j}\right)^{\rho} g(d_j, m, \rho) \sim \frac{\bar{d}^{\rho}}{1-\rho}\frac{1}{k}\sum_{j=1}^{k}\frac{\Gamma(d_j - \rho + 1)}{\Gamma(d_j + 1)}.$$

The distributed Hill estimator achieves the oracle property when $\lambda = 0$ or $\rho = 0$. Nevertheless, it is not guaranteed that this condition is also necessary.

Homogeneous case and $d \to \infty$

Theorem

Suppose $F \in D(G_\gamma)$ with $\gamma > 0$ and Conditions A and B hold. Let $d_1 = d_2 = \cdots = d_k = d$, $d = d(m) \to \infty$ and $d/m \to 0$ as $n \to \infty$. If $\sqrt{kd}A(m/d) = O(1)$ as $n \to \infty$, then

$$\sqrt{kd}\left(\hat{\gamma}_{DH} - \gamma - A(m/d)g(d, m, \rho)\right) \xrightarrow{d} N(0, \gamma^2).$$

Remark: In this case, the condition $m \to \infty$ and $k \to \infty$ can be relaxed.

Oracle property

- Assume that $\sqrt{kd}A(n/(kd) = \sqrt{kd}A(m/d) \to \lambda \in \mathbb{R}$, the oracle Hill estimator possesses the asymptotic normality

$$\sqrt{kd}(\hat{\gamma}_H - \gamma) \xrightarrow{d} N(\lambda/(1-\rho), \gamma^2)$$

- Under the same condition, $g(d, m, \rho) \to 1$. So, the oracle property always holds.

Table of Contents

Non Identically Distributed Case

- Assume all observations are independent, but only observations on the same machine follow the same distribution.
- Denote the distribution function of observations in machine j as $F_{k,j}$ for $j = 1, 2, \ldots, k$.

Assumption: heteroscedastic extremes

(C.) There exists a continuous distribution function F such that

$$\lim_{x\to\infty} \frac{1-F_{k,j}(x)}{1-F(x)} = c_{k,j},$$

uniformly for all $1 \leq j \leq k$ and all $k \in \mathbb{N}$ with $c_{k,j}$ uniformly bounded away from 0 and ∞.

Define $U_{k,j}(t) = (/(1-F_{k,j}))^{\leftarrow}(t)$. Assume that $F \in D(G_\gamma)$. It is straight forward to show that condition C leads to

$$\lim_{t\to\infty} \frac{U_{k,j}(t)}{U(t)} = c_{k,j}^{\gamma}.$$

Assumptions: second order heteroscedastic extremes

(D.) There exists an eventually positive or negative function $A_1(t) \in RV(\tilde{\rho})$ with index $\tilde{\rho} \leq 0$ and $\lim_{t\to\infty} A_1(t) = 0$ such that

$$\sup_{k\in\mathbb{N}} \max_{1\leq j\leq k} \left| \frac{U_{k,j}(t)}{U(t)} - c_{k,j}^{\gamma} \right| = O\left(A_1(t)\right),$$

Under the heteroscedastic extremes setup, Einmahl et al. (2016) shows that one could apply the Hill estimator to the oracle example while discarding the fact that they are not from the same distribution.

Asymptotic property

Theorem

Suppose $F \in D(G_\gamma)$ *with* $\gamma > 0$ *and Conditions A-D hold. Let* $d_1 = d_2 = \cdots = d_k = d$, *where* $d \geq 1$ *is a fixed integer. If* $\sqrt{kd}A(m/d) = O(1)$ *and* $\sqrt{kd}A_1(m/d) \to 0$ *as* $n \to \infty$, *then*

$$\sqrt{kd}\left(\hat{\gamma}_{DH} - \gamma - A(m/d)g(d, m, \rho)\right) \xrightarrow{d} N(0, \gamma^2).$$

Oracle property

- The above theorem can be easily extended to the heterogeneous case where d_j are uniformly bounded and the homogeneous case where $d = d(m)$ is an intermediate sequence.

- The heteroscedastic extremes setup does not affect the oracle properties of the distributed Hill estimators.

Table of Contents

Simualtion Setting

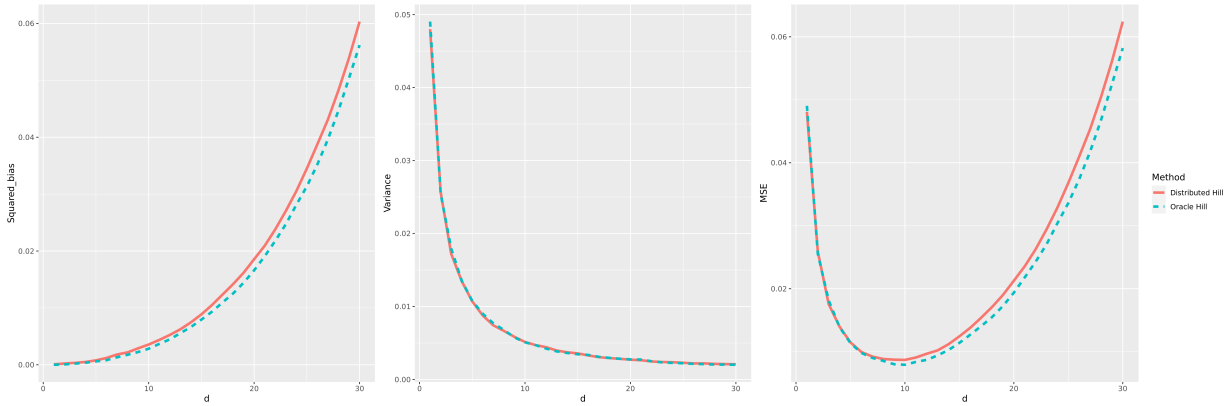
We consider three distributions:

- Fréchet distribution: $F(x) = e^{-x^{-1}}, x > 0$
- Pareto(1) distribution $F(x) = 1 - x^{-1}, x > 1$
- Absolute Cauchy distribution: $f(x) = 2(\pi(1 + x^2))^{-1}, x > 0.$

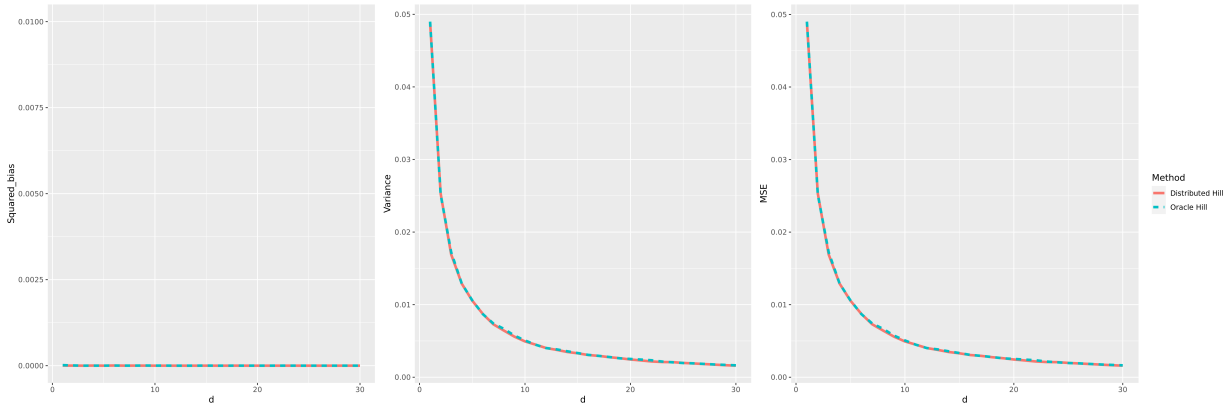
The sample size is $n = 1000$. The bias, variance and MSE are based on $r = 1000$ Monte Carlo repetitions.

Simulation 1: Comparison for different level of d

Fix $k = 20$ and $m = 50$ and compare the finite sample performance for different values of d.

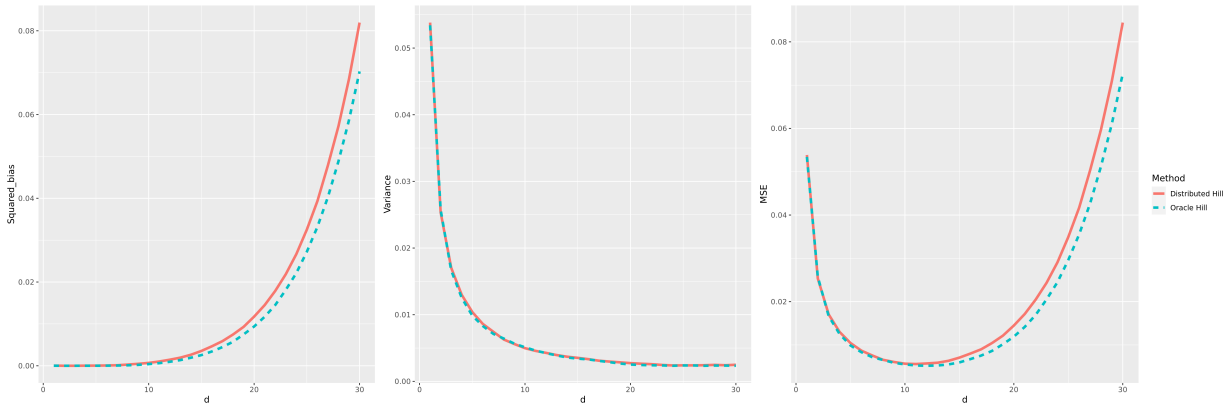


(a) Fréchet(1) Distribution



(b) Pareto(1) Distribution

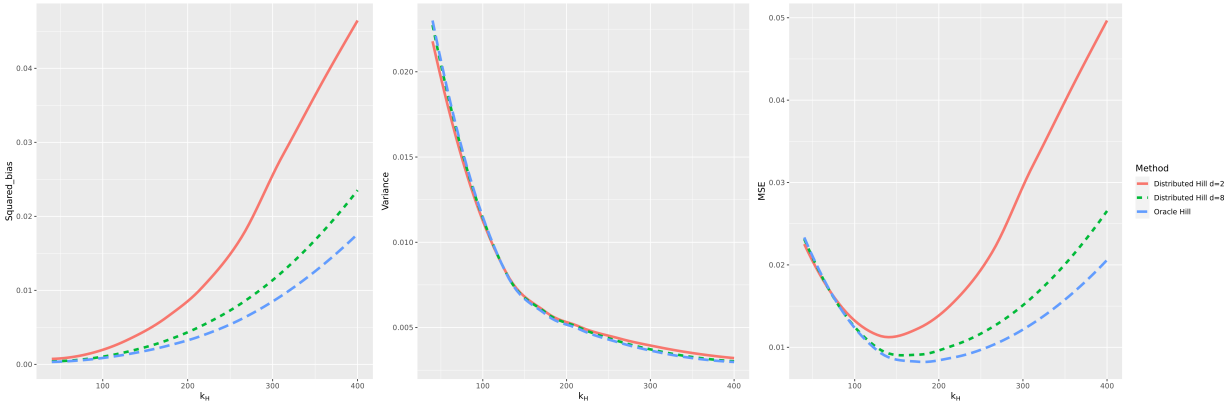
Continue



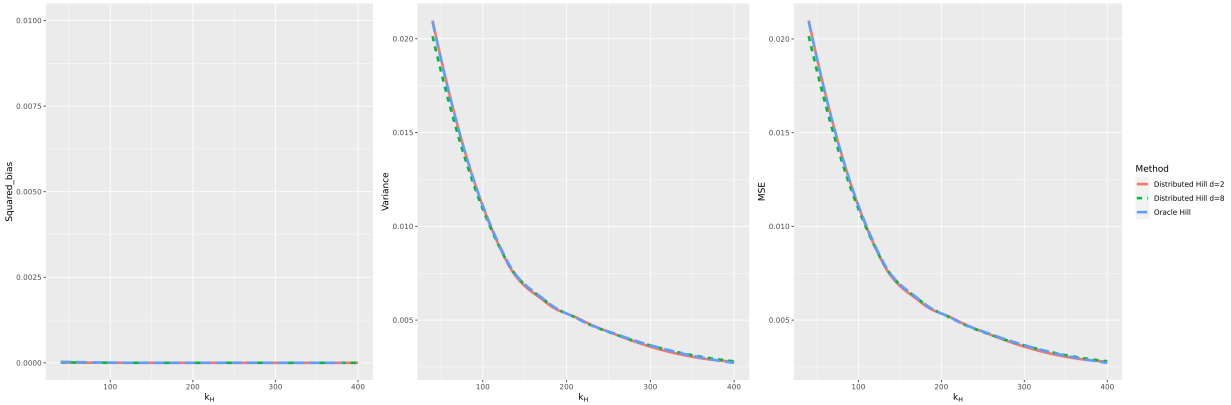
(c) Absolute Distribution

Simu2: Comparison for different level of k

Fix $d = 2$ (a low level) and $d = 8$ (an intermediate level.) The x-axis is the total number of exceedance ratios.

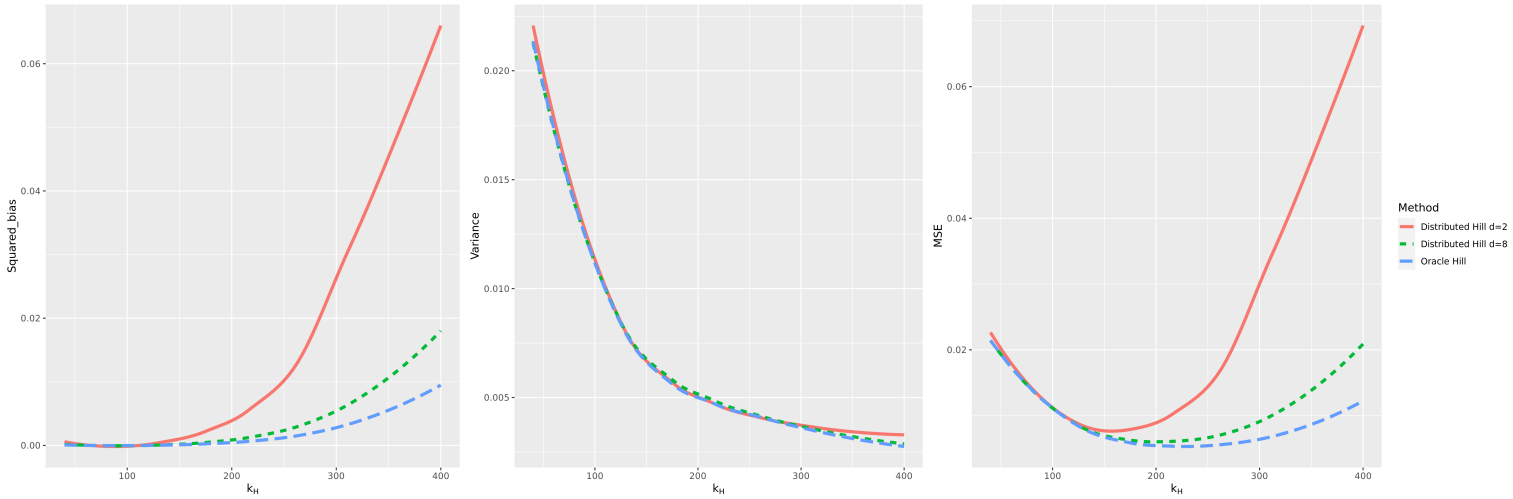


(d) Fréchet(1) Distribution



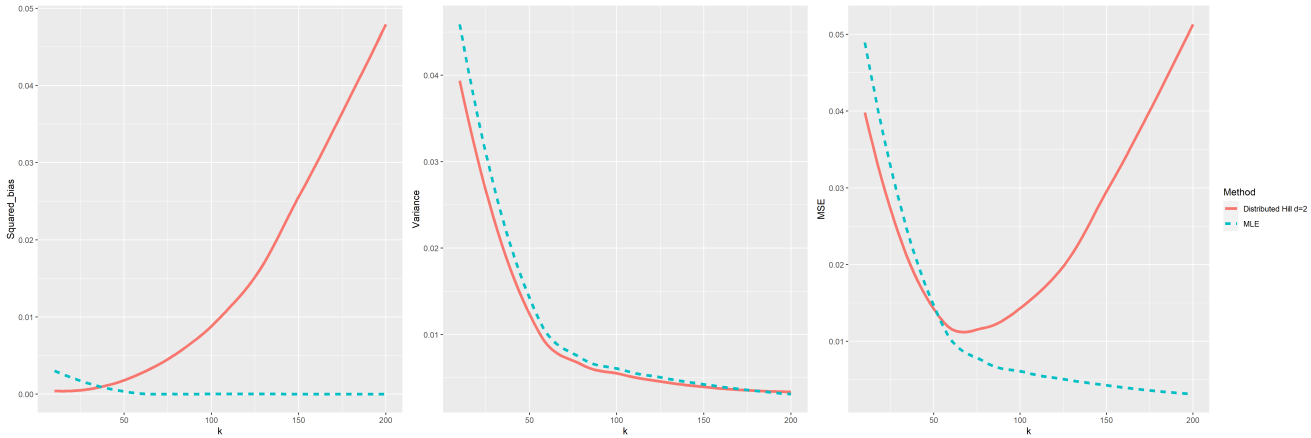
(e) Pareto(1) Distribution

Continue

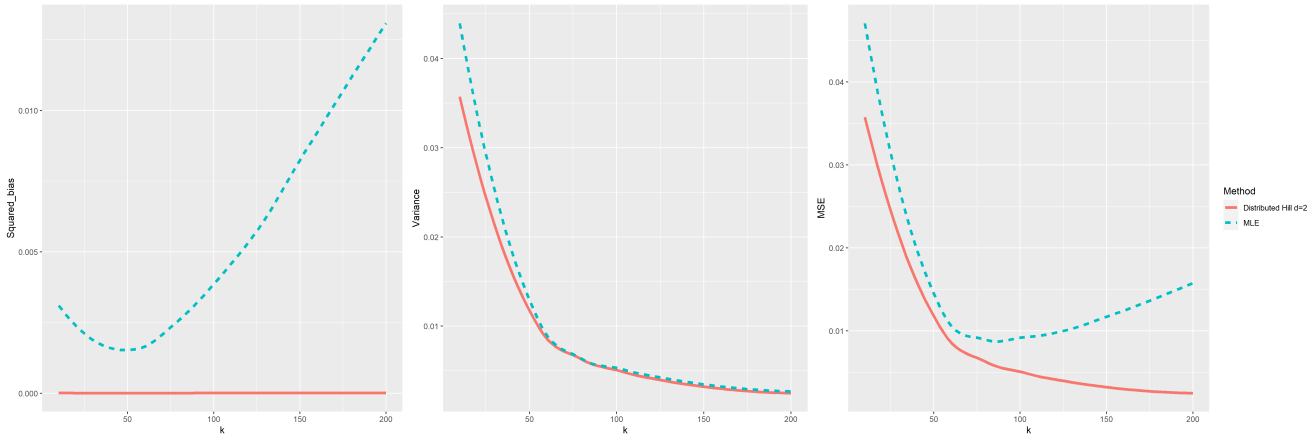


(f) Absolute Cauchy Distribution

Simu3: Comparison with block maxima MLE

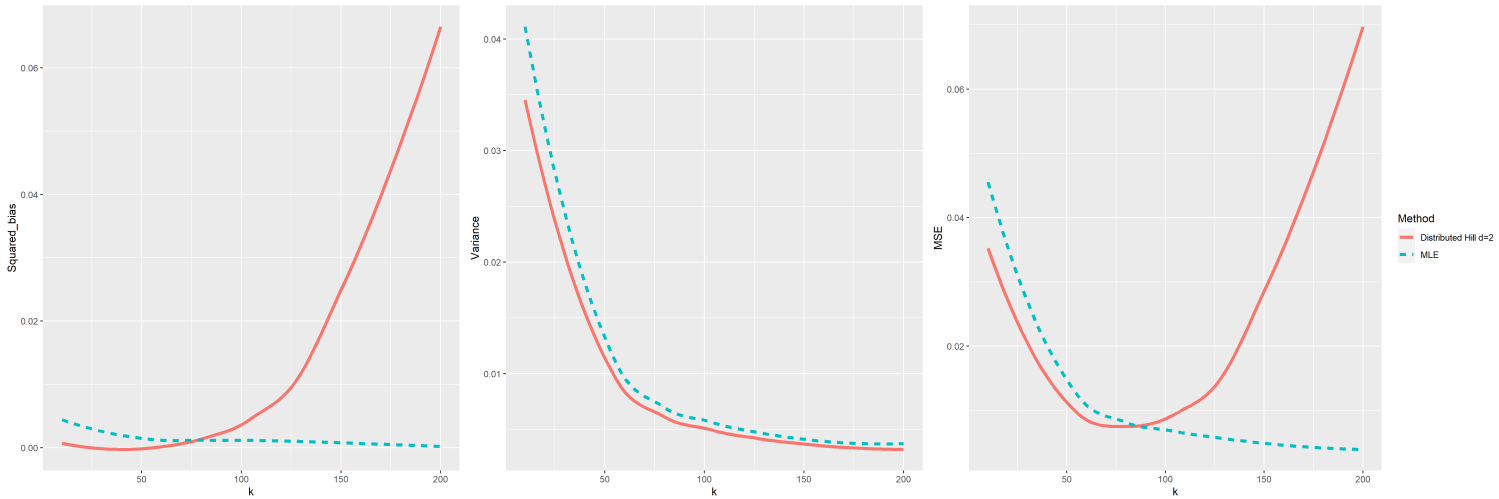


(g) Fréchet(1) Distribution



(h) Pareto(1) Distribution

Continue



(i) Absolute Cauchy Distribution

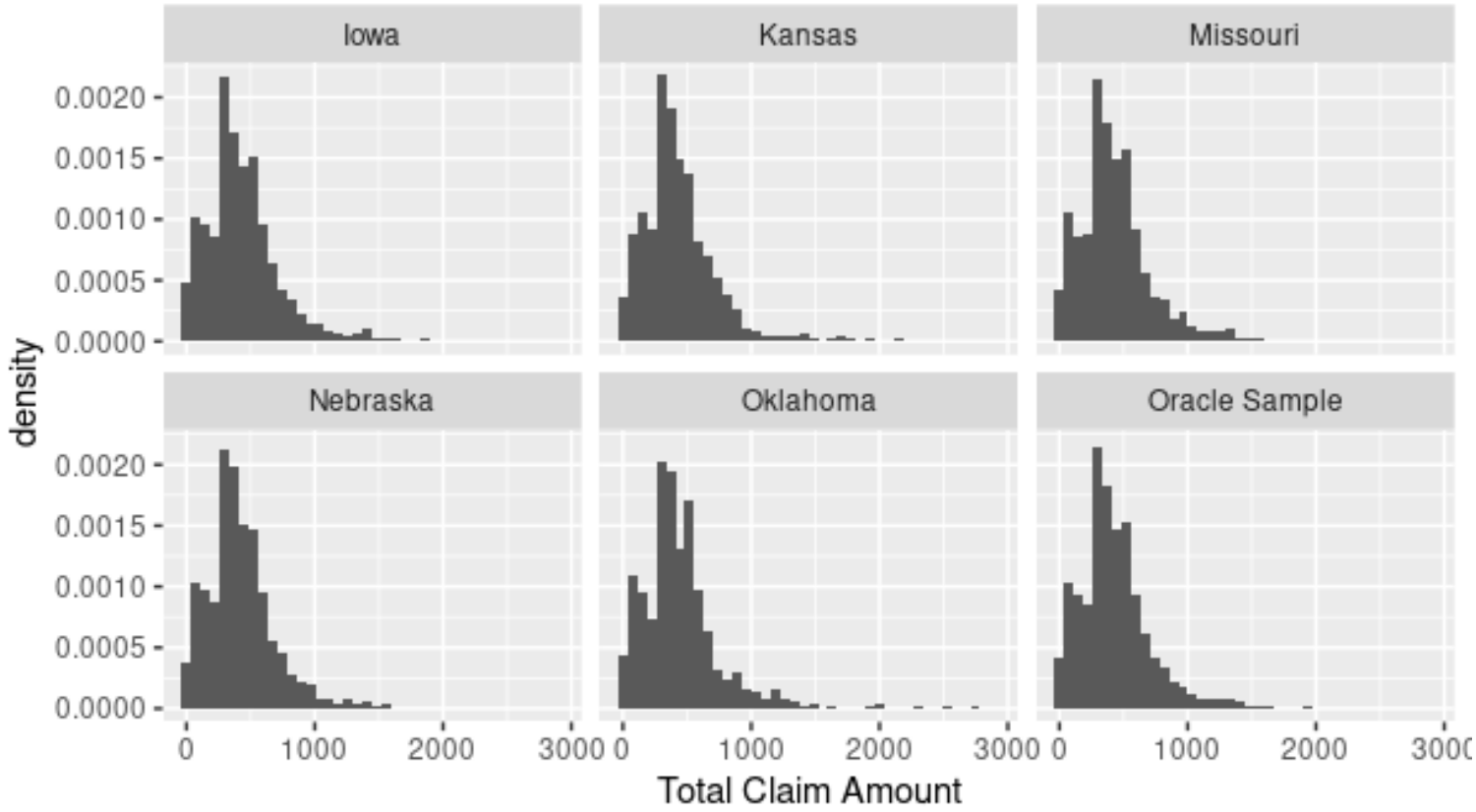
Table of Contents

Dataset Description

- Car insurance claims in five states of the United States during January and February 2011.
- We consider five hypothetical insurance companies, one in each state, where each company monopolies all the car insurance in its own state.
- Due to business privacy, companies cannot share their data to others but they are willing to share their statistical results.

Dataset Description

The dataset contains 9134 observations, with 2601, 798, 3150, 1703 and 882 observations for Iowa,Kansas, Missouri, Nebraska and Oklahoma, repsectively.



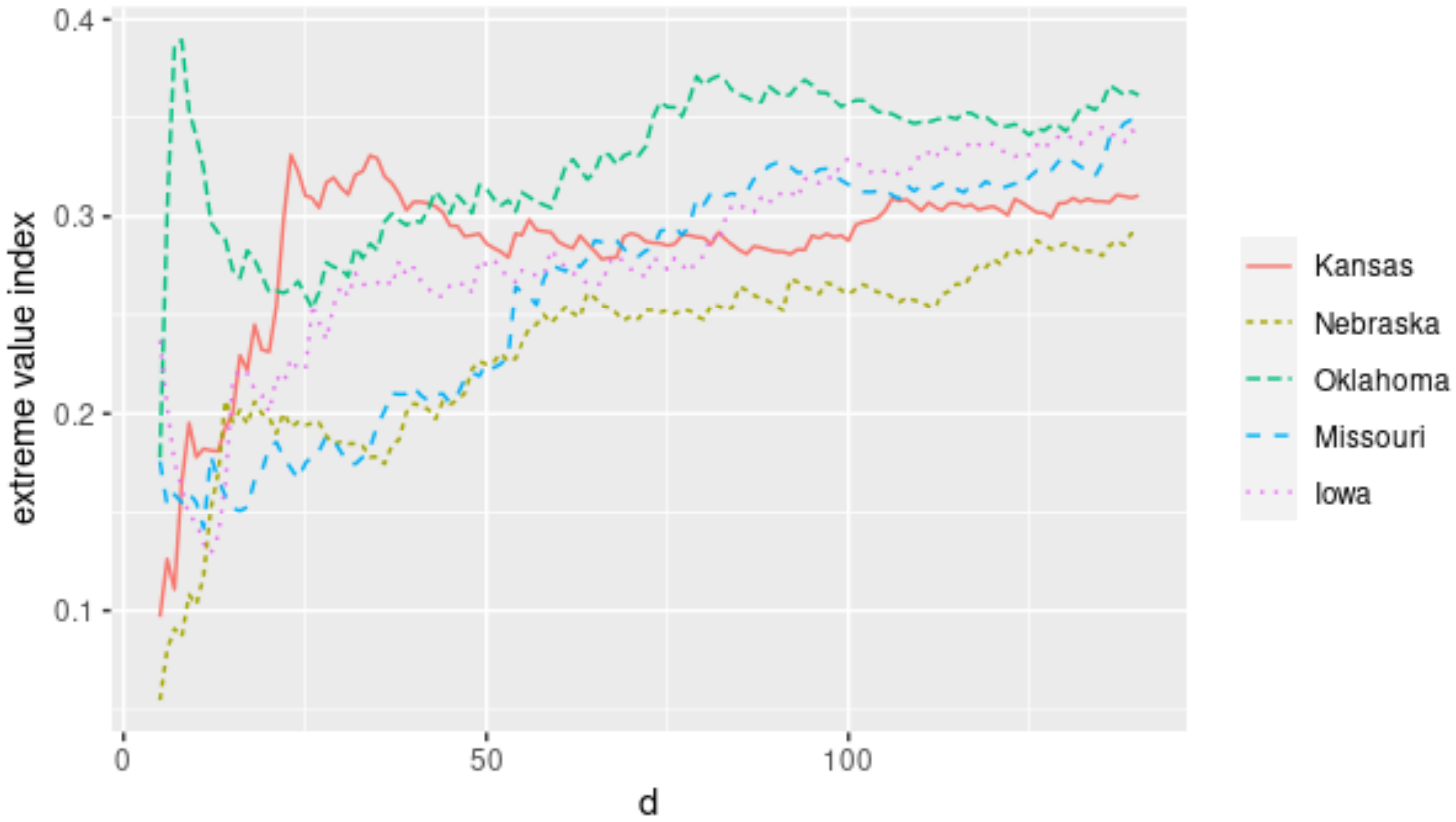


Figure: Estimation for the extreme value index of total claim amount for each state.

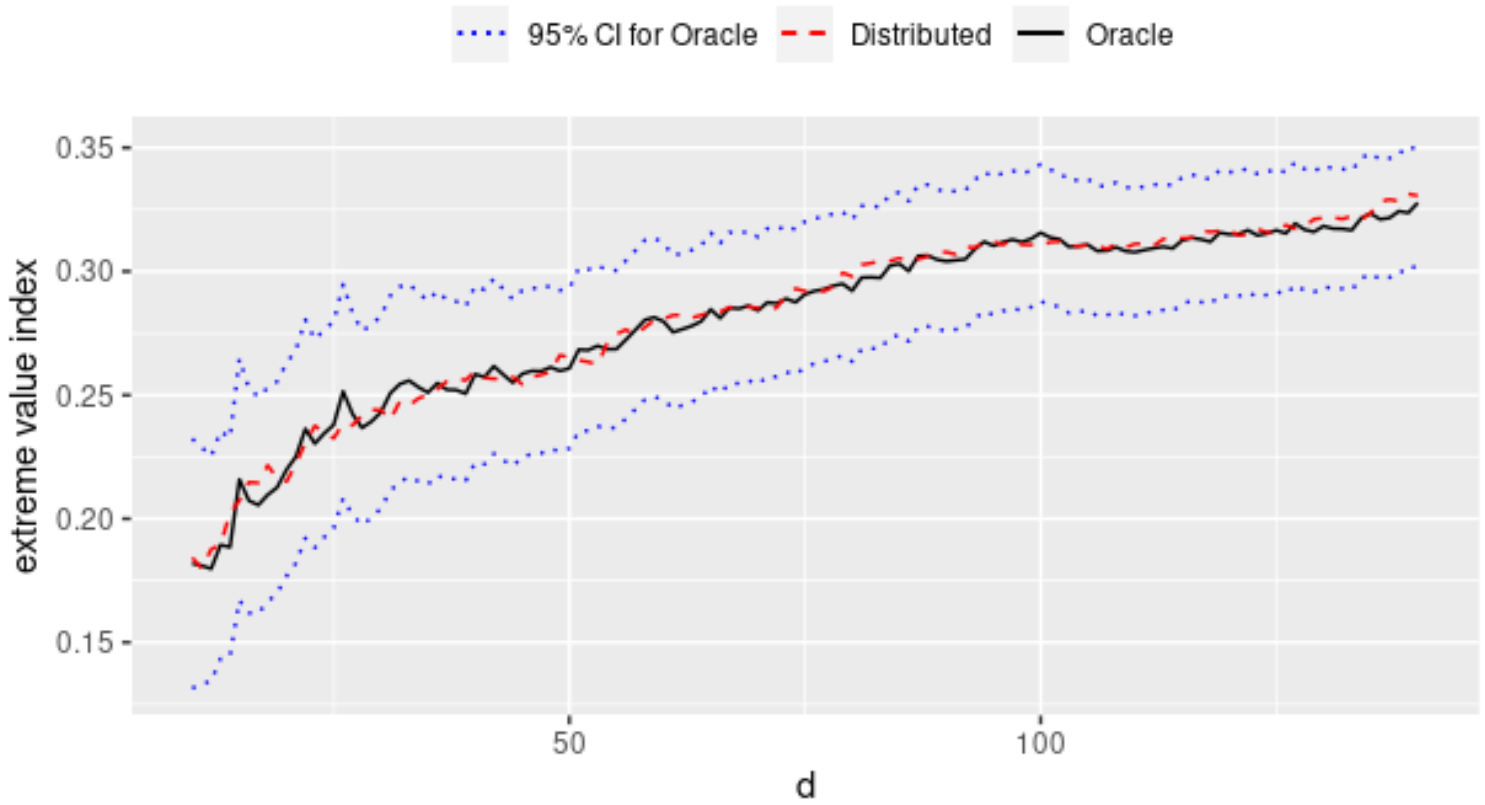


Figure: The point estimation and the 95% confidence interval of the distributed Hill estimator and the oracle Hill estimator for γ of total claim amount.