Estimating Parameters of Pareto Distribution Under Interval and Fuzzy Uncertainty

Nitaya Buntao

****

Department of Applied Statistics
King Mongkut's University of Technology North Bangkok
Thailand
Email: taltanot@hotmail.com

February 16, 2011

Table of Contents

Estimating Parameters of the Pareto Distribution

The Pareto distribution is a power law probability distribution: the probability that X is greater than some number x is given by

$$f_X(x) = \begin{cases} \alpha \cdot \dfrac{x_0^{\alpha}}{x^{\alpha+1}} & ; \quad \text{if } x > x_0 \\ 0 & ; \quad \text{if } x < x_0 \ . \end{cases}$$

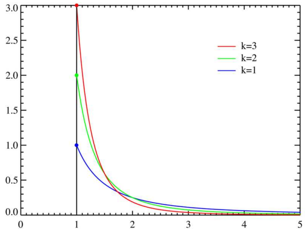

Figure: Pareto probability density functions for various α with $x_0 = 1$.

Estimating Parameters of the Pareto Distribution

- Reminder: the Pareto distribution is a power law probability distribution: the probability that X is greater than some number x is given by

$$f_X(x) = \begin{cases} \alpha \cdot \frac{x_0^\alpha}{x^{\alpha+1}} & ; \quad \text{if } x > x_0 \\ 0 & ; \quad \text{if } x < x_0 \end{cases}.$$

- Estimators of parameters x_0 and α based on the observed data values $x_1, \ldots, x_n$ come from applying the Maximum Likelihood techniques:

$$\hat{x}_0 = \min(x_1, \ldots, x_n),$$

and

$$\hat{\alpha} = n \cdot \left(\sum_{i=1}^{n} \ln \left(\frac{x_i}{\min(x_1, \ldots, x_n)} \right) \right)^{-1}.$$

Need to take into Account Interval and Fuzzy Uncertainty

- In practice, we rarely know the exact values of x_i. For example, in financial situations, we can take, as x_i, the price of the financial instrument at the i-th moment of time – e.g., on the i-th day. However, the price does not remain stable during the day.

Need to take into Account Interval and Fuzzy Uncertainty

- In practice, we rarely know the exact values of x_i. For example, in financial situations, we can take, as x_i, the price of the financial instrument at the i-th moment of time – e.g., on the i-th day. However, the price does not remain stable during the day.
- It is more reasonable to consider the whole range $[\underline{x}_i, \overline{x}_i]$ of the daily prices instead of a single value x_i.

Need to take into Account Interval and Fuzzy Uncertainty

- In practice, we rarely know the exact values of x_i. For example, in financial situations, we can take, as x_i, the price of the financial instrument at the i-th moment of time – e.g., on the i-th day. However, the price does not remain stable during the day.
- It is more reasonable to consider the whole range $[\underline{x}_i, \overline{x}_i]$ of the daily prices instead of a single value x_i.
- We need to find the range of all resulting values of x_0 and α.

Need to take into Account Interval and Fuzzy Uncertainty

- In practice, we rarely know the exact values of x_i. For example, in financial situations, we can take, as x_i, the price of the financial instrument at the i-th moment of time – e.g., on the i-th day. However, the price does not remain stable during the day.
- It is more reasonable to consider the whole range $[\underline{x}_i, \overline{x}_i]$ of the daily prices instead of a single value x_i.
- We need to find the range of all resulting values of x_0 and α.
- Estimating this range under interval uncertainty is a particular case of a general problem of *interval computations.*

Interval Uncertainty: $x_i \in [\underline{x}_i, \overline{x}_i]$

- Some of these values x_i may be flukes caused by accidental errors.

Interval Uncertainty: $x_i \in [\underline{x}_i, \overline{x}_i]$

- Some of these values x_i may be flukes caused by accidental errors.
- Experts can usually tell which values x_i are possible. However, experts used word from a natural language.

Interval Uncertainty: $x_i \in [\underline{x}_i, \overline{x}_i]$

- Some of these values x_i may be flukes caused by accidental errors.
- Experts can usually tell which values x_i are possible. However, experts used word from a natural language.
- To describe these natural-language statements, it is reasonable to use *fuzzy logic*.

Interval Uncertainty: $x_i \in [\underline{x}_i, \overline{x}_i]$

- Some of these values x_i may be flukes caused by accidental errors.
- Experts can usually tell which values x_i are possible. However, experts used word from a natural language.
- To describe these natural-language statements, it is reasonable to use *fuzzy logic*.
- It is desirable to conclude what is the degree of possibility of different values x_0 and α from the corresponding intervals.

From the Computational Viewpoint, Fuzzy Data Processing can be Reduced to Interval Data Processing

- The set
$$X_i(\alpha) = \{x_i : \mu_i(x_i) \geq \alpha\}$$
is called an *alpha-cut*.

From the Computational Viewpoint, Fuzzy Data Processing can be Reduced to Interval Data Processing

- The set
$$X_i(\alpha) = \{x_i : \mu_i(x_i) \geq \alpha\}$$
is called an *alpha-cut*.

- For every function $R(x_1, \ldots, x_n$ and for $R = R(X_1, \ldots, X_n)$, we have:
$$R(\alpha) = \{R(x_1, \ldots, x_n) : x_i \in X_i(\alpha)\}.$$

From the Computational Viewpoint, Fuzzy Data Processing can be Reduced to Interval Data Processing

- The set
  $$X_i(\alpha) = \{x_i : \mu_i(x_i) \geq \alpha\}$$
  is called an *alpha-cut*.
- For every function $R(x_1, \ldots, x_n$ and for $R = R(X_1, \ldots, X_n)$, we have:
  $$R(\alpha) = \{R(x_1, \ldots, x_n) : x_i \in X_i(\alpha)\}.$$
  Thus, $\forall \alpha$, finding the alpha-cut of the resulting membership function $\mu(R)$ is equivalent to applying interval computations to the corresponding intervals $X_1(\alpha)$, ..., $X_n(\alpha)$.

From the Computational Viewpoint, Fuzzy Data Processing can be Reduced to Interval Data Processing

- The set
$$X_i(\alpha) = \{x_i : \mu_i(x_i) \geq \alpha\}$$
is called an *alpha-cut*.
- For every function $R(x_1, \ldots, x_n$ and for $R = R(X_1, \ldots, X_n)$, we have:
$$R(\alpha) = \{R(x_1, \ldots, x_n) : x_i \in X_i(\alpha)\}.$$
Thus, $\forall\alpha$, finding the alpha-cut of the resulting membership function $\mu(R)$ is equivalent to applying interval computations to the corresponding intervals $X_1(\alpha)$, ..., $X_n(\alpha)$.
- We will thus only consider the case of interval uncertainty.

Estimating x_0 Under Interval Uncertainty

Problem: Reminder

Let $x_1 \in [\underline{x}_1, \overline{x}_1], \ldots, x_n \in [\underline{x}_n, \overline{x}_n]$.
The range of the estimator $x_0 = \min(x_1, \ldots, x_n)$.

Consider: The function $x_0 = \min(x_1, \ldots, x_n)$ is a (non-strictly) increasing function of each of its variables.

Estimating x_0 Under Interval Uncertainty

Problem: Reminder

Let $x_1 \in [\underline{x}_1, \overline{x}_1], \ldots, x_n \in [\underline{x}_n, \overline{x}_n]$.
The range of the estimator $x_0 = \min(x_1, \ldots, x_n)$.

Consider: The function $x_0 = \min(x_1, \ldots, x_n)$ is a (non-strictly) increasing function of each of its variables.

The interval of possible values of x_0:
the largest possible value of x_0 is equal to

$$\overline{x}_0 = \min(\overline{x}_1, \ldots, \overline{x}_n),$$

and the smallest possible value of x_0 is equal to

$$\underline{x}_0 = \min(\underline{x}_1, \ldots, \underline{x}_n).$$

Reducing the Problem:

Problem:

We want to find the range $[\underline{\alpha}, \overline{\alpha}]$ of the estimate α.

Reducing the Problem:

Problem:

We want to find the range $[\underline{\alpha}, \overline{\alpha}]$ of the estimate α.

First step:

according to the description of the estimate α, this estimate has the form

$$\alpha = \frac{n}{r},$$

where we denote

$$r = \sum_{i=1}^{n} \ln\left(\frac{x_i}{\min(x_1, \ldots, x_n)}\right).$$

Reducing the Problem: First Step

Since the function $\frac{n}{r}$ is decreasing,

- the largest possible value $\overline{\alpha}$ of $\alpha = \frac{n}{r}$ is attained when r takes the smallest possible value, and

Reducing the Problem: First Step

Since the function $\dfrac{n}{r}$ is decreasing,

- the largest possible value $\overline{\alpha}$ of $\alpha = \dfrac{n}{r}$ is attained when r takes the smallest possible value, and
- the smallest possible value $\underline{\alpha}$ of $\alpha = \dfrac{n}{r}$ is attained when r takes the largest possible value.

Relation between α and r:

$$(\alpha = \overline{\alpha}) \Leftrightarrow (r = \underline{r}) \text{ and } (\alpha = \underline{\alpha}) \Leftrightarrow (r = \overline{r})$$

Reducing the Problem: First Step

Since the function $\dfrac{n}{r}$ is decreasing,

- the largest possible value $\overline{\alpha}$ of $\alpha = \dfrac{n}{r}$ is attained when r takes the smallest possible value, and
- the smallest possible value $\underline{\alpha}$ of $\alpha = \dfrac{n}{r}$ is attained when r takes the largest possible value.

Relation between α and r:

$$(\alpha = \overline{\alpha}) \Leftrightarrow (r = \underline{r}) \text{ and } (\alpha = \underline{\alpha}) \Leftrightarrow (r = \overline{r})$$

Then we can then find the range $[\underline{\alpha}, \overline{\alpha}]$ for α as follows:

$$\underline{\alpha} = \frac{n}{\overline{r}}; \quad \overline{\alpha} = \frac{n}{\underline{r}}.$$

Reducing the Problem:

Second step:

we can use the fact that r is the sum of several logarithms, and the sum of the logarithms is equal to the logarithm of the product:

$$r = \ln(S),$$

where we denote

$$S \stackrel{\text{def}}{=} \prod_{i=1}^{n} \frac{x_i}{\min(x_1, \ldots, x_n)} = \frac{\prod\limits_{i=1}^{n} x_i}{(\min(x_1, \ldots, x_n))^n}.$$

Reducing the Problem: Second Step

Since the function $\ln(S)$ is increasing,

- the largest possible value $\overline{r}$ of $r = \ln(S)$ is attained when S takes the largest possible value, and
- the smallest possible value $\underline{r}$ of $r = \ln(S)$ is attained when S takes the smallest possible value.

Relation between r and S:

$$(r = \overline{r}) \Leftrightarrow (S = \overline{S}) \text{ and } (r = \underline{r}) \Leftrightarrow (S = \underline{S})$$

Reducing the Problem: Second Step

Since the function $\ln(S)$ is increasing,

- the largest possible value $\overline{r}$ of $r = \ln(S)$ is attained when S takes the largest possible value, and
- the smallest possible value $\underline{r}$ of $r = \ln(S)$ is attained when S takes the smallest possible value.

Relation between r and S:

$$(r = \overline{r}) \Leftrightarrow (S = \overline{S}) \text{ and } (r = \underline{r}) \Leftrightarrow (S = \underline{S})$$

Then the range of r is

$$[\underline{r}, \overline{r}] = [\ln(\underline{S}), \ln(\overline{S})].$$

Further Reduction

When we know that x_j is the smallest of n values $x_1, \ldots, x_n$, then the expression for S can be simplified even further:

$$S = \frac{\prod_{i=1}^{n} x_i}{x_j^n}.$$

Further Reduction

When we know that x_j is the smallest of n values $x_1, \ldots, x_n$, then the expression for S can be simplified even further:

$$S = \frac{\prod_{i=1}^{n} x_i}{x_j^n}.$$

Then we cam further simplify this expression into

$$S = \frac{\prod_{i \neq j} x_i}{x_j^{n-1}}.$$

Computing $\overline{S}$

The expression for S is increasing as a function of all the variables x_i with $i \neq j$ and decreasing as a function of the remaining variable x_j. Thus, its largest possible value is attained when:

- all the variables x_i with $i \neq j$ attain their largest possible value $\overline{x}_i$, while
- the variable x_i attains its smallest possible value $\underline{x}_j$.

Computing $\overline{S}$

The expression for S is increasing as a function of all the variables x_i with $i \neq j$ and decreasing as a function of the remaining variable x_j. Thus, its largest possible value is attained when:

- all the variables x_i with $i \neq j$ attain their largest possible value $\overline{x}_i$, while
- the variable x_i attains its smallest possible value $\underline{x}_j$.

The corresponding expression is equal to

$$S_j = \frac{\prod_{i=1}^{n} \overline{x}_i}{\overline{x}_j \cdot \underline{x}_j^{n-1}} = \frac{\prod_{i \neq j} \overline{x}_i}{\underline{x}_j^{n-1}}.$$

This expression is only possible when $x_j \leq x_i$ for all $i \neq j$, i.e., when $\underline{x}_j \leq \overline{x}_i$ for all i.

$$\underline{x}_j \leq \min(\overline{x}_1, \ldots, \overline{x}_n) = \overline{x}_0.$$

Computing $\underline{\alpha}$

Therefore

$$\underline{\alpha} = \frac{n}{\overline{r}} = \frac{n}{\ln\left(\overline{S}\right)} = \frac{n}{\ln\left(\frac{\prod\limits_{i \neq j} \overline{x}_i}{\underline{x}_j^{n-1}}\right)},$$

where

$$S \stackrel{\text{def}}{=} \frac{\prod\limits_{i=1}^{n} x_i}{\left(\underline{x}_j\right)^n}$$

and

$$\underline{x}_j = \min(\underline{x}_1, \ldots, \underline{x}_n).$$

Computing $\underline{S}$

Consider,

Case I:

all the values x_i are equal to each other: $x_1 = \ldots = x_n$.

Computing $\underline{S}$

Consider,

Case I:

all the values x_i are equal to each other: $x_1 = \ldots = x_n$.

Since $S = \frac{\prod_{i=1}^{n} x_i}{x_j^n}$, $x_j = \min(x_1, \ldots, x_n)$

then $S = 1$

and we can increase all the values until we reach the upper endpoint $\overline{x}_i$ of one of the intervals,
where $\overline{x}_i = \min(\overline{x}_1, \ldots, \overline{x}_n)$ $(= \overline{x}_0)$.

- For this i, we have $x_i = \overline{x}_i$,
- and for all other $k \neq i$, we get $x_k = \max(\overline{x}_i, \underline{x}_k)$.

Computing $\underline{S}$

Case II:

not all the coordinates of the optimizing vector $(x_1, \ldots, x_n)$ are equal to each other.

Computing $\underline{S}$

Case II:

not all the coordinates of the optimizing vector $(x_1, \ldots, x_n)$ are equal to each other.

Since $$S = \frac{\prod\limits_{i \neq j} x_i}{x_j^{n-1}}, \qquad x_j = \min(x_1, \ldots, x_n)$$
and S is increasing as a function of all the variable x_i with $i \neq j$ and decreasing as a function of the remaining variable x_j.
We increase x_j where $x_j \leq \overline{x}_j$ and $x_j \leq x_i$.

Computing $\underline{S}$

Case II:

not all the coordinates of the optimizing vector $(x_1, \ldots, x_n)$ are equal to each other.

Since $$S = \frac{\prod\limits_{i \neq j} x_i}{x_j^{n-1}}, \qquad x_j = \min(x_1, \ldots, x_n)$$

and S is increasing as a function of all the variable x_i with $i \neq j$ and decreasing as a function of the remaining variable x_j.
We increase x_j where $x_j \leq \overline{x}_j$ and $x_j \leq x_i$.

Thus, we either have

- we either have $x_j = \overline{x}_j$,
- or we have $x_j < \overline{x}_j$ and $x_j = x_i$ for some $i \neq j$.

Computing $\underline{S}$

Thus, we either have

- we either have $x_j = \overline{x}_j$,
- **or we have $x_j < \overline{x}_j$ and $x_j = x_i$ for some $i \neq j$.** ✓

Let us consider the second case, when we have several values x_i for which $x_j = x_i$.

Let n_j be the total number of such values x_j. We conclude that

$$S = \frac{\prod\limits_{i: x_i > x_j} x_i}{x_j^{n-n_j}}.$$

If $\forall i$ for which $x_i = x_j$, we have $x_i < \overline{x}_i$, then we can increase value $x_j = x_i = \ldots$ without changing any other value x_k and thus, further decrease S. So, at least for one i, we have $x_j = x_i = \overline{x}_i$.

Computing $\underline{S}$

- Thus, for the minimizing vector, the smallest value $\min(x_1, \ldots, x_n)$ is attained at one of the upper endpoints $\overline{x}_i$.
- Since this value $\overline{x}_i$ is the smallest, we get $\overline{x}_i \le x_k$ for all $k \neq i$, and since $x_k \le \overline{x}_k$, we conclude that $\overline{x}_i \le \overline{x}_k$ for all k
- Thus, the minimal value $\overline{x}_i = \min(x_1, \ldots, x_n)$ is the smallest of n upper endpoints:

$$x_j = \min(\overline{x}_1, \ldots, \overline{x}_n) = \overline{x}_0.$$

- For every $k \neq i$, we select the smallest possible value $x_k \in [\underline{x}_k, \overline{x}_k]$ for which $x_k \ge \overline{x}_k$, i.e., the value $x_k = \max(\overline{x}_i, \underline{x}_k)$.
- The smallest value $\underline{S}$ of S corresponds to the smallest value $\underline{r}$ of r and thus, to the largest value $\overline{\alpha}$ of α.

Algorithm for Computing $\underline{\alpha}$: Stages to Find $\underline{\alpha}$

First stage:

We compute the value $\overline{x}_0 = \min(\overline{x}_1, \ldots, \overline{x}_n)$.
$*$ If we have the range $[\underline{x}_0, \overline{x}_0]$ then we just borrow the value $\overline{x}_0$.

Algorithm for Computing $\underline{\alpha}$: Stages to Find $\underline{\alpha}$

First stage:

We compute the value $\overline{x}_0 = \min(\overline{x}_1, \ldots, \overline{x}_n)$.
$*$ If we have the range $[\underline{x}_0, \overline{x}_0]$ then we just borrow the value $\overline{x}_0$.

Second stage:

Find $\min(\overline{x}_j \cdot \underline{x}_j^{n-1})$ where $\underline{x}_j \leq \overline{x}_0, j = 1, \ldots, n$.
$\Rightarrow$ We get the index j.

Algorithm for Computing $\underline{\alpha}$: Stages to Find $\underline{\alpha}$

First stage:

We compute the value $\overline{x}_0 = \min(\overline{x}_1, \ldots, \overline{x}_n)$.
$*$ If we have the range $[\underline{x}_0, \overline{x}_0]$ then we just borrow the value $\overline{x}_0$.

Second stage:

Find $\min(\overline{x}_j \cdot \underline{x}_j^{n-1})$ where $\underline{x}_j \leq \overline{x}_0, j = 1, \ldots, n$.
$\Rightarrow$ We get the index j.

Final formula:

$$\underline{\alpha} = \frac{n}{\ln\left(\frac{\prod\limits_{i \neq j} \overline{x}_i}{\underline{x}_j^{n-1}}\right)} = n \cdot \left(\sum_{i \neq j} \ln\left(\frac{\overline{x}_i}{\underline{x}_j}\right)\right)^{-1}.$$

Computation Time

- At each stage, this algorithm takes the linear number of steps,
  - i.e., the number of steps bounded by the number of variables n.
  - Indeed, we need to take into account each of the intervals $[\underline{x}_i, \overline{x}_i]$.

Computation Time

- At each stage, this algorithm takes the linear number of steps,
  - i.e., the number of steps bounded by the number of variables n.
  - Indeed, we need to take into account each of the intervals $[\underline{x}_i, \overline{x}_i]$.
- Thus, overall, we have a linear-time algorithm.
  - Thus, the overall number of computation steps cannot be smaller than n.
  - So, our algorithm that takes times $\leq \text{const} \cdot n$ is asymptotically optimal.

This computation time ($\leq \text{const} \cdot n$) is asymptotically optimal.

To Find $\overline{\alpha}$

First stage:

we compute the value $\overline{x}_0 = \min(\overline{x}_1, \ldots, \overline{x}_n)$. $*$ If we already have the range $[\underline{x}_0, \overline{x}_0]$ then we just borrow the value $\overline{x}_0$.

To Find $\overline{\alpha}$

First stage:

we compute the value $\overline{x}_0 = \min(\overline{x}_1, \ldots, \overline{x}_n)$. $*$ If we already have the range $[\underline{x}_0, \overline{x}_0]$ then we just borrow the value $\overline{x}_0$.

Second stage: for each $k = 1, \ldots, n$,

we take $x_k = \max(\overline{x}_0, \underline{x}_k)$, and then compute the value α as

$$\overline{\alpha} = n \cdot \left(\sum_{k=1}^{n} \ln \left(\frac{\max(\overline{x}_0, \underline{x}_k)}{\overline{x}_0} \right) \right)^{-1}.$$

To Find $\overline{\alpha}$

First stage:

we compute the value $\overline{x}_0 = \min(\overline{x}_1, \ldots, \overline{x}_n)$. $*$ If we already have the range $[\underline{x}_0, \overline{x}_0]$ then we just borrow the value $\overline{x}_0$.

Second stage: for each $k = 1, \ldots, n$,

we take $x_k = \max(\overline{x}_0, \underline{x}_k)$, and then compute the value α as

$$\overline{\alpha} = n \cdot \left(\sum_{k=1}^{n} \ln \left(\frac{\max(\overline{x}_0, \underline{x}_k)}{\overline{x}_0} \right) \right)^{-1}.$$

Computation time.

This algorithm also takes linear time and is, thus, also asymptotically optimal.

Acknowledgments

The author would like to thank

- Ass. Prof. Sa-aat Niwitpong (KMUTNB, Thailand),
- Prof. Hung T. Nguyen (CMU, Thailand),
- Prof. Tony Wang (NMSU, USA),
- and prof. Vladik Kreinovich (UTEP, USA)

for their encouragement and advise.

END

Thank you very much.