Censored Data

Maximum Likelihood Estimation

Nikita Orlov*

Faculty of Economic Sciences, Higher School of Economics, Moscow

E-mail: naorlov_2@edu.hse.ru

Abstract

The current study focuses on the distribution parametric estimation using Maximum Likelihood and applying this approach to left and right-censored data. This paper can be divided into three parts. At first, the proof of the asymptotic properties of Maximum Likelihood Estimators will be provided. Then we are going to apply MLE with censored data for one-dimensional distributions. Finally, we construct MLE for two-dimensional data and see if the results hold.

Motivation

The inspiration for this study is borrowed from the paper Mechanism Choice in Scoring Auctions by Pasha Andreyanov.[1] The problem of data censoring is very common in auctions due to the existence of reserve prices. When the task is to define the real distribution of values, we face partially left-censored distribution of bids (when the reserve price takes place).

Furthermore this approach can be useful for solving other problems in microeconomics and econometrics as censoring is almost everywhere. Certainly the current study focuses only

on parametric methods of dealing censored data. The task of this research is to derive the scheme of using MLE to estimate the distribution of data and giving examples applying this scheme to simulations using python (here is the link to all python simulations including Figures 2-5).

Maximum Likelihood

Maximum likelihood estimators have strong asymptotic properties: efficiency, normality and consistency. That is why MLE is widely-used to compute real data distribution based on observations. Moreover Maximum likelihood estimators are easy to find and do not require extra manipulations with data. But everything becomes a little more difficult when talking about censored data.

The current part of the research focuses on proving asymptotic properties of classical Maximum likelihood. The idea of the proof is also basic and is published in various articles. The current proof is similar to the one studied at Stamford University.[2]

Theorem

Let $\theta_0 \in \Theta\,, X_1, \ldots, X_n \overset{i.i.d.}{\sim} f(x|\theta_0)$ and $\hat{\theta}$ is a maximum-likelihood estimator.

Regularity conditions:

(1) Probability density functions $f(x|\theta)$ have the same support.

(2) θ_0 is not on the boundary of Θ.

(3) $l = \ln(L)(\theta)$, where $L(\theta)$ is a likelihood function is differentiable in θ.

(4) $\hat{\theta}$ is unique and solves $l^{'}(\theta) = 0$

Then $\hat{\theta}$ is consistent and asymptotically normal estimator for θ.

$$\text{Consistency:} \qquad \operatorname*{plim}_{n\to\infty} \hat{\theta} = \theta_0 \Leftrightarrow \forall \varepsilon > 0 : \lim_{n\to\infty} \mathbb{P}(|\hat{\theta} - \theta_0| > \varepsilon) = 0$$

$$\text{Asymptotic normality:} \qquad \sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow[n\to\infty]{F} \mathcal{N}\Big(0, \frac{1}{I(\theta_0)}\Big)$$

$s(x,\theta) = \frac{\partial}{\partial\theta}\ln f(x|\theta)$ is the score function and $I(\theta_0)$ is the total Fisher information and can be calculated this way:

$$I(\theta) = \mathrm{Var}_\theta\big[s(x,\theta)\big] = -\mathbb{E}_\theta\big[s^{'}(x,\theta)\big]$$

For large n this theorem also states the following:

(a) $\hat{\theta}$ is asymptotically unbiased

(b) $\mathrm{Var}\big[\hat{\theta}\big] \approx \frac{1}{nI(\theta)}$

(c) Assuming θ_0 is the real parameter: $\hat{\theta} \overset{as}{\approx} \mathcal{N}\big(\theta_0, \frac{1}{nI(\theta_0)}\big)$

Properties of the score for $\theta \in \Theta$:

$$\mathbb{E}_\theta\big[s(X,\theta)\big] = 0 \qquad \mathrm{Var}_\theta\big[s(X,\theta)\big] = -\mathbb{E}_\theta\big[s^{'}(X,\theta)\big]$$

Proof. Using some simple operations with differentiation:

$$s(x,\theta)f(x|\theta) = \Big(\frac{\partial}{\partial\theta}\ln f(x|\theta)\Big)f(x|\theta) = \frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}f(x|\theta) = \frac{\partial}{\partial\theta}f(x|\theta)$$

$$\text{Since the square under the PDF is always : } \int f(x|\theta)\,dx = 1$$

$$\mathbb{E}\big[s(X,\theta)\big] = \int s(x,\theta)f(x|\theta)\,dx = \int \frac{\partial}{\partial\theta}f(x|\theta)\,dx = \frac{\partial}{\partial\theta}\int f(x|\theta)\,dx = 0$$

Let's prove the second property of score based on the previous results.

$$\mathbb{E}\big[s^{'}(X,\theta)\big] + \mathrm{Var}\big[s(X,\theta)\big] = \mathbb{E}\big[s^{'}(X,\theta)\big] + \mathbb{E}\big[s^2(X,\theta)\big] =$$

$$= \int s^{'}(x,\theta)f(x|\theta)\,dx + \int s(x,\theta)s(x,\theta)f(x|\theta)\,dx =$$

$$= \int (s^{'}(x,\theta)f(x|\theta) + s(x,\theta)\frac{\partial}{\partial\theta}f(x|\theta))\,dx = \frac{\partial}{\partial\theta}\int s(x,\theta)f(x|\theta)\,dx =$$

$$= \frac{\partial}{\partial\theta}\mathbb{E}\big[s(x,\theta)\big] = 0$$

Consistency Proof

To prove the consistency of MLE note that $\hat{\theta}$ is found this way:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \ln f(x|\theta)$$

If the real parameter is θ_0, then by The Law of Large Numbers:

$$\forall\, \theta \in \Theta : \frac{1}{n} \sum_{i=1}^{n} \ln f(x|\theta) \xrightarrow[n\to\infty]{\text{P}} \mathbb{E}_{\theta_0}\big[\ln f(X|\theta)\big]$$

Thus under Regularity conditions the value $\hat{\theta}$, which maximizes the left side of the equation, converges in probability to θ_0, which maximizes the right side.

$$\forall\, \theta \in \Theta : \mathbb{E}_{\theta_0}\big[\ln f(X|\theta)\big] - \mathbb{E}_{\theta_0}\big[\ln f(X|\theta_0)\big] = \mathbb{E}_{\theta_0}\Big[\ln \frac{f(X|\theta)}{f(X|\theta_0)}\Big]$$

The logarithm function is strictly concave, so we can apply Jensen's Inequality: $\forall$ positive variable $X : \mathbb{E}\big[\ln X\big] \le \ln \mathbb{E}\big[X\big]$. Applying inequality to our expression:

$$\mathbb{E}_{\theta_0}\Big[\ln \frac{f(X|\theta)}{f(X|\theta_0)}\Big] \le \ln \mathbb{E}_{\theta_0}\Big[\frac{f(X|\theta)}{f(X|\theta_0)}\Big] = \ln \int \frac{f(x|\theta)}{f(x|\theta_0)} f(x|\theta_0)\, dx = \ln \int f(x|\theta)\, dx = 0$$

Consistency of MLE is based on the fact: $\theta_0 = \arg\max \mathbb{E}_{\theta_0}\big[f(x|\theta)\big]$

Asymptotic Normality Proof

The log-likelihood function is maximized at $\theta = \hat{\theta}$, therefore $l^{'}(\hat{\theta}) = 0$. In the previous point we proved that $\hat{\theta}$ converges in probability to the real parameter θ_0. This lets us use the first-order Taylor expansion to $l^{'}(\hat{\theta}) = 0$ around $\hat{\theta} = \theta_0$.

$$l^{'}(\theta_0) + (\hat{\theta} - \theta_0) l^{''}(\theta) \approx 0 \Rightarrow \sqrt{n}(\hat{\theta} - \theta_0) \approx -\sqrt{n} \frac{l^{'}(\theta_0)}{l^{''}(\theta_0)} = -\frac{\frac{1}{\sqrt{n}} l^{'}(\theta_0)}{\frac{1}{n} l^{''}(\theta_0)}$$

By The Law of Large Numbers and the definition of Fisher Information:

$$\frac{1}{n}l^{''}(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\ln L(f(x_i|\theta))\Big|_{\theta=\theta_0} = \frac{1}{n}\sum_{i=1}^{n}s^{'}(x_i,\theta_0) \xrightarrow[n\to\infty]{\text{P}} \mathbb{E}_{\theta_0}\big[s^{'}(X,\theta_0)\big] = -I(\theta_0)$$

Recalling the Properties of the score: $\text{Var}_\theta\big[s(X,\theta)\big] = I(\theta_0)$ if $X \sim f(x|\theta_0)$.

Then by The Central Limit Theorem:

$$\frac{1}{\sqrt{n}}l^{'}(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f(x_i|\theta)\Big|_{\theta=\theta_0} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}s(x_i,\theta_0) \xrightarrow[n\to\infty]{\text{F}} \mathcal{N}\big(0, I(\theta_0)\big)$$

Now we can apply Slutsky's Lemma: if $X_n \xrightarrow[n\to\infty]{\text{P}} c$ and $Y_n \xrightarrow[n\to\infty]{\text{F}} Y$, then $X_nY_n \xrightarrow[n\to\infty]{\text{F}} cY$ and the Continuous Mapping Theorem to get the result:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow[n\to\infty]{\text{F}} \frac{1}{I(\theta_0)}\mathcal{N}\big(0, I(\theta_0)\big) = \mathcal{N}\big((0, I^{-1}(\theta_0))\big)$$

Example: Exponential Distribution

Consider $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f(x|\lambda)$, where $f(x|\lambda) = \lambda\exp(-\lambda x), x \geq 0$.

The Likelihood function:

$$L(\lambda) = \prod_{i=1}^{n}f(x|\lambda) = \prod_{i=1}^{n}\lambda e^{-\lambda x} = \lambda^n\exp\Big(-\lambda\sum_{i=1}^{n}x_i\Big) \to \max_{\lambda}$$

$$l(\lambda) = \ln L(\lambda) = n\ln\lambda - \lambda\sum_{i=1}^{n}x_i \to \max_{\lambda}$$

$$s(x,\lambda) = l^{'}(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n}x_i \Rightarrow \hat{\lambda} = \frac{n}{\sum_i x_i}$$

$$s^{'}(x,\lambda) = l^{''}(\lambda) = -\frac{n}{\lambda^2} < 0 \Rightarrow \hat{\lambda} \Rightarrow I(\lambda) = -\mathbb{E}_\lambda\big[z^{'}(X,\lambda)\big] = \frac{n}{\lambda^2}$$

Using asymptotic normality, we can state that $\hat{\lambda} \xrightarrow[n\to\infty]{\text{F}} \mathcal{N}\left(\lambda, \frac{\lambda^2}{n}\right)$

To prove the consistency we must use The Law of Large Numbers:

$$\operatorname*{plim}_{n\to\infty} \hat{\lambda}_n = \operatorname*{plim}_{n\to\infty} \frac{n}{\sum_i x_i} = \frac{n}{n\mathbb{E}\big[X_i\big]} = \lambda$$

Therefore $\hat{\theta} = n/\sum_i x_i$ - is asymptotically normal and consistent estimator for λ

Censored MLE

Now we can continue our study with introducing likelihood function in censored data case. The idea of using maximum likelihood for estimating partially censored data is described in the paper by N. Balakrishnan and M. Kateri (2008).[3] In this research they define the likelihood function which includes censoring and use it for Weibull Distribution.

Let index i denote the observable data. Suppose that r samples are observable and others are not. Then in case of left-censored data the likelihood function looks this way:

$$L(x|\theta) = \prod_{i=1}^{r} f(x_i|\theta) \prod_{j=r+1}^{n} F(x_j|\theta) \to \max_{\theta}$$

$$l(x) = \sum_{i=1}^{r} \ln f(x_i|\theta) + \sum_{j=r+1}^{n} \ln F(x_j|\theta) \to \max_{\theta}$$

$$l^{'}(x) = \sum_{i=1}^{r} \frac{\partial}{\partial\theta} \ln f(x_i|\theta) + \sum_{j=r+1}^{n} \frac{\partial}{\partial\theta} \ln F(x_j|\theta)\bigg|_{\theta=\hat{\theta}} = 0$$

Gradient Methods of Optimization

When introducing special conditions like censoring, the optimization problem becomes more complicated and cannot be solved explicitly. In this case gradient methods can be used to find the solution.

Suppose single parameter $\theta \in \Theta$ and $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f(x|\theta)$. At first stage we initialize the

starting point: $\hat{\theta}_0 = \gamma_0$. After that we repeat the same procedure until convergence:

$$\hat{\theta}_n = \hat{\theta}_{n-1} + \eta \frac{\partial}{\partial \theta} l(x, \theta)\Big|_{\theta=\hat{\theta}_{n-1}}$$

This method is called the Gradient Descent and the idea is to move in the direction of function's greatest increasement speed. Gradient Descent is often used in machine learning to optimize loss functions.

Example: Normal Distribution

Finally we can apply the MLE algorithm to some real distribution. The motivation for using Normal Distribution to estimate data has come from the Central Limit Theorem which guarantees that independent identically drawn samples converge in distribution to Gaussian. To find MLE we must recall the PDF and CDF of normal distribution:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(x-\mu)^2}{2\sigma^2} \right) \qquad F(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \exp\left(- \frac{(t-\mu)^2}{2\sigma^2} \right) dt$$

The likelihood function with censored data:

$$L\big(x|\mu, \sigma^2\big) = \prod_{i=1}^{r} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(x_i-\mu)^2}{2\sigma^2} \right) \prod_{j=r+1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x_j} \exp\left(- \frac{(t-\mu)^2}{2\sigma^2} \right) dt$$

Then the log-likelihood function is:

$$\begin{aligned} l(x) = &\sum_{i=1}^{r} \left(- \frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2 - \frac{(x_i-\mu)^2}{2\sigma^2} \right) + \\ &+ \sum_{j=r+1}^{n} \left(- \frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2 + \ln\left(\int_{-\infty}^{x_j} \exp\left(- \frac{(t-\mu)^2}{2\sigma^2} \right) dt \right) \right) = \\ &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \sum_{i=1}^{r} \frac{(x_i-\mu)^2}{2\sigma^2} + (n-r) \cdot \ln\left(\int_{-\infty}^{x_j} \exp\left(- \frac{(t-\mu)^2}{2\sigma^2} \right) dt \right) \end{aligned}$$

Now we must find the derivatives with respect to μ and σ^2. Our function differs from the standard likelihood of normal distribution by the last term. Finding its partial derivatives gives us (under certain conditions we can switch the integration and differentiation):

$$\psi = \ln\left(\int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt\right)$$

$$\frac{\partial\psi}{\partial\mu} = \frac{\dfrac{\partial}{\partial\mu}\displaystyle\int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt}{\displaystyle\int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt} = \frac{\displaystyle\int_{-\infty}^{x} \frac{t-2\mu}{2\sigma^2} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt}{\displaystyle\int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt}$$

$$\frac{\partial\psi}{\partial\sigma^2} = \frac{\dfrac{\partial}{\partial\sigma^2}\displaystyle\int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt}{\displaystyle\int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt} = \frac{\displaystyle\int_{-\infty}^{x} \frac{(t-\mu)^2}{2\sigma^4} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt}{\displaystyle\int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt}$$

Now we can find the derivatives of log-likelihood function and apply Gradient Methods to find MLEs for μ and σ^2

$$\frac{\partial l}{\partial\mu} = \sum_{i=1}^{r} \frac{x_i-\mu}{\sigma^2} + (n-r)\cdot\frac{\partial\psi}{\partial\mu} \qquad \frac{\partial l}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \sum_{i=1}^{r} \frac{(x_i-\mu)^2}{2\sigma^4} + (n-r)\cdot\frac{\partial\psi}{\partial\sigma^2}$$

Thus the algorithm of finding the optimal parameters looks this way:

$$\hat{\mu}_t = \hat{\mu}_{t-1} + \eta\frac{\partial l}{\partial\mu}\bigg|_{\mu=\hat{\mu}_{t-1},\sigma^2=\hat{\sigma}^2_{t-1}} \qquad \hat{\sigma}^2_t = \hat{\sigma}^2_{t-1} + \eta\frac{\partial l}{\partial\sigma^2}\bigg|_{\mu=\hat{\mu}_{t-1},\sigma^2=\hat{\sigma}^2_{t-1}}$$

Later we will return to estimating Normal distribution using some other algorithm with python simulation. And now it is important to mention the scheme for right-censoring

Right-censored Data: Exponential Distribution

In case the right side of data is censored, we do not see the values that are greater than $\max x_r$, where index r denotes the observable data. That is why instead of $F(x)$ we use $1 - F(x)$ to include non-observable data in Maximum likelihood function.

$$L(x|\theta) = \prod_{i=1}^{r} f(x_i|\theta) \prod_{j=r+1}^{n} \big(1 - F(x_j|\theta)\big) \to \max_{\theta \in \Theta}$$
$$l(x|\theta) = \sum_{i=1}^{r} \ln f(x_i|\theta) + \sum_{j=r+1}^{n} \ln \big(1 - F(x_j|\theta)\big) \to \max_{\theta \in \Theta}$$

For exponential distribution the scheme is:

$$f(x|\lambda) = \lambda \exp\big(-\lambda x\big) \qquad F(x|\lambda) = 1 - \exp\big(-\lambda x\big)$$
$$l(x|\lambda) = \sum_{i=1}^{r} \big(\ln \lambda - \lambda x_i\big) + \sum_{j=r+1}^{n} -\lambda x_j = n \ln \lambda - \lambda \sum_{i=1}^{r} x_i - \lambda \sum_{j=r+1}^{n} x_j$$
$$\frac{\partial l}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_k \Longrightarrow \hat{\lambda} = \frac{r}{\sum_k x_k}$$

Therefore no matter whether we are facing right or left-censored data. The only thing that differs is the function used as a supplement for censoring.

Nelder-Mead Method

Motivation

Gradient methods are widely-used for optimization, but they are based on finding the function derivatives with respect to all parameters, which can take long time and be rather complicated to calculate. Nelder-Mead Algorithm[4] does not require finding the derivatives and is often used for optimizing functions. It is also implemented in scipy.optimize as one of the default methods.

Algorithm

Initialization: Consider the task of minimization of a function with n parameters. At first we must initialize $n+1$ points: $x_1, \ldots, x_{n+1}$ (the current simplex) and compute the function's value in these points $f(x_1), \ldots, f(x_{n+1})$. Each iteration consists of three phases:
(1) Ordering: We determine the indexes h - the highest function value, s - the second highest value and l - the lowest value.

$$y_h = \max_{i=1,\ldots,n+1}\{f(x_i)\} \qquad y_s = \max_{i\neq h}\{f(x_i)\} \qquad y_l = \min_{i=1,\ldots,n+1}\{f(x_i)\}$$

(2) Centroid: Calculating the centoroid for every x besides x_h.

$$c = \frac{1}{n}\sum_{i\neq h} x_i$$

(3) Transform: At this stage we compute the new vertex using methods from Figure 1:

- At first stage we are replacing only the worst vertex x_h using the reflection, expansion or contraction.
- If this succeeds replace x_h with the new point
- If this fails use shrinkage for the best point x_l.

Nelder-Mead method has four main parameters for reflection, expansion, contraction and shrinkage:

$$\alpha > 0, \quad 0 < \beta < 1, \quad \gamma > 1, \quad \gamma > \alpha, \quad 0 < \delta < 1$$

- Reflection: Calculating the point $x_r = c + \alpha(c + x_h)$. If $f(x_l) \leq f(x_r) < f(x_s)$ - accept the point.

- Expansion: If reflection failed: $f(x_r) > f(x_s)$, compute the point $x_e = c + \gamma(x_r - c)$. There are two cases: 1) $f(x_r) > f(x_e)$ - accept x_e or $f(x_r) < f(x_e)$ - accept x_r.

- Contraction: If $f(x_r) \geq f(x_s)$: compute the point x_c depending on the situation:
  1) Outside: $f(x_s) \leq f(x_r) < f(x_h)$; $x_c = c + \beta(x_r - c)$. If $f(x_c) \leq f(x_r)$ accept x_c.
  2) Inside: $f(x_r) \geq f(x_h)$; $x_c = c + \beta(x_h - c)$. If $f(x_c) < f(x_h)$ accept x_c.
  If none of these conditions is satisfied use shrinkage.

- Shrinkage: Introducing n new points: $x_i = x_l + \delta(x_i - x_l) \, \forall i \neq l$.

This procedure is repeated until convergence.

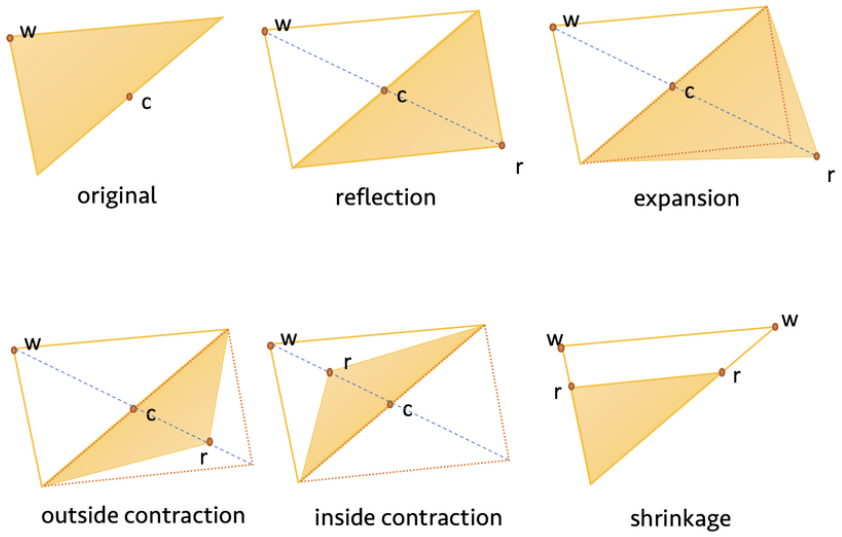

Figure 1: Operations in Nelder-Mead Optimization.

Picture from the paper by Jiang, Yang, Bierner, Li, Wang, Moghtaderi[5]

Results: Normal Distribution

In this section we demonstrate the results of applying the Nelder-Mead Method for the task of maximizing log-likelihood of normal distribution using scipy.minimize. The solution can be seen in code part!!!.

The results are rather optimistic: for smaller number of observations the estimation results are very positive for comparatively small sigma and $10 - 15\%$ of censored data (Figure 2).

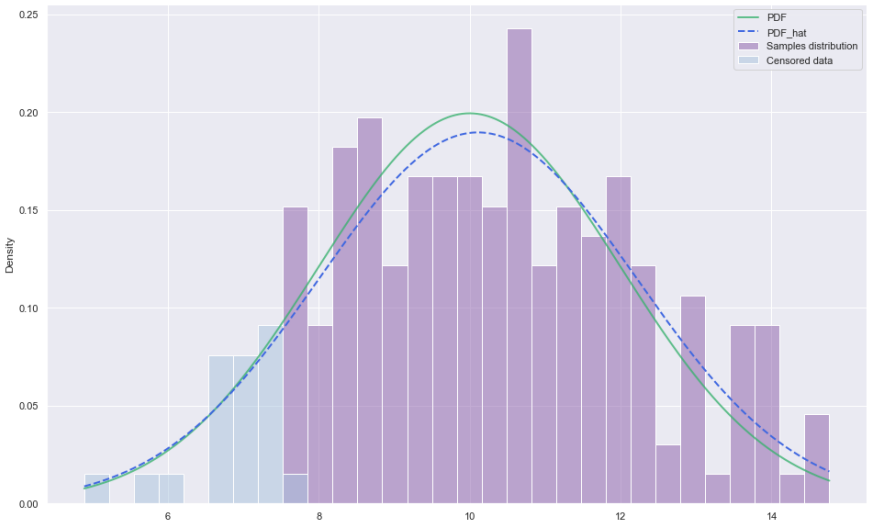

Figure 2: Data Censoring 10% , $\sigma = 2$, 200 observations

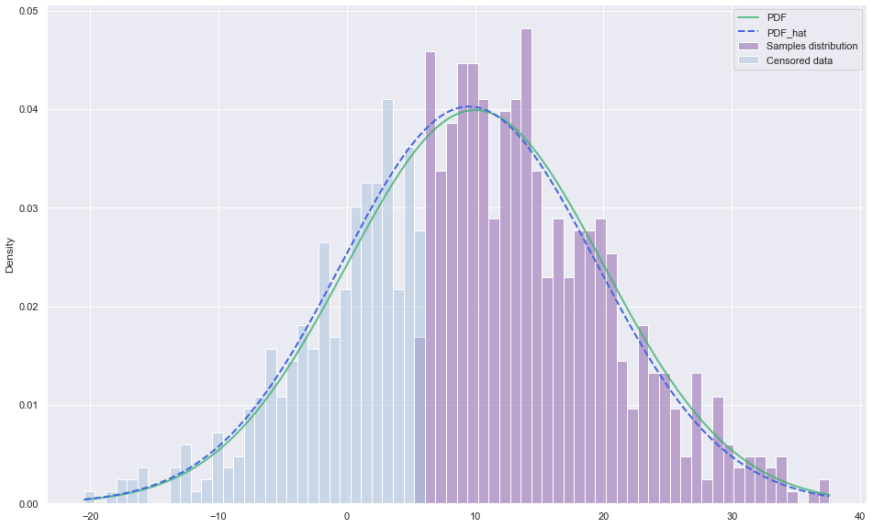

Figure 3: Data Censoring 35% , $\sigma = 10$, 1000 observations

When we increase the number of observations, the deviations from the real distribution are small even for large sigmas and bigger percentage of data censoring (Figure 3).

Two-dimensional Distributions

Supposing the situation with two parameters where one of them is censored is more realistic but more difficult at the same time. This is exactly the problem faced by Pasha Andreyanov in his study.[1] In this case it is obvious to change the censored part of the likelihood function into something different.

Optimization Task

Actually, when one parameter x is observable and the other one y is left-censored, we must add the probability that y is smaller than the censored value according to the value of observable x. That is the conditional cumulative probability function.

$$F_Y(y|X=x) = \int_{-\infty}^{y} f_Y(t|x)\,dt \qquad f_Y(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

When r observations are seen and others are censored the likelihood function looks this way:

$$L(x,y|\theta) = \prod_{i=1}^{r} f_{X,Y}(x,y|\theta) \prod_{j=r+1}^{n} (1 - F_Y(y_j|x_j,\theta)) \to \max_{\theta}$$

$$l = \ln L(x,y|\theta) = \sum_{i=1}^{r} \ln f_{X,Y}(x,y|\theta) + \sum_{j=r+1}^{n} \ln(1 - F_Y(y_j|x_j,\theta)) \to \max_{\theta}$$

Assuming right-censoring is rather obvious as we are going to use Pareto distribution, where the mass center is within low values and it is more rational to suppose that greater values are unseen.

Bivariate Pareto Distribution

Pareto Distribution is a good example to illustrate the work of MLE in two dimensions as it does not face the problem of undefined integrals (as in Gaussian distribution). At the same time it can be helpful because Pareto distribution takes place in lots of processes, especially in economics and physics.

The PDF of 2d Pareto distribution is has been borrowed from the paper by N.H. Abdel-Alla, H.N. Abd-Ellah[6]

$$f_{X,Y}(x, y, \gamma, \sigma, \alpha) = \alpha(\alpha + 1)(\gamma\sigma)^{\alpha+1}\lambda^{-(\alpha+2)}, \quad x \geq \gamma > 0\,,\, y \geq \sigma > 0$$

$$\lambda = \sigma x + \gamma y - \gamma\sigma \text{ - Pareto distribution of the first-kind}$$

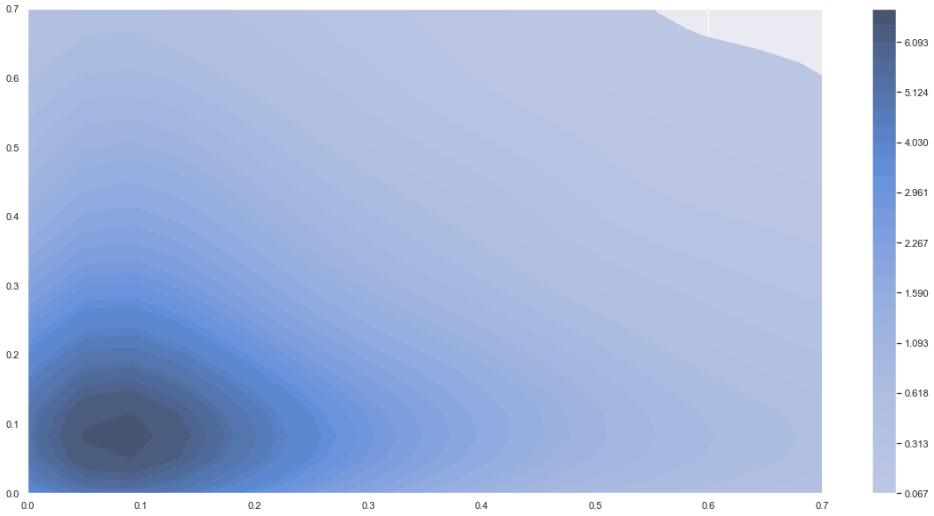

Figure 4: Bivariate Pareto Distribution PDF

Before calculating the conditional PDF and CDF, we should get the marginal PDF of X.

$$f_X(x) = \int_{\sigma}^{+\infty} f_{X,Y}(x,y)\,dy = \\ = \int_{\sigma}^{+\infty} \alpha(\alpha+1)(\gamma\sigma)^{\alpha+1}(\sigma x+\gamma y-\gamma\sigma)^{-(\alpha+2)}\,dy = \\ = -\alpha(\gamma\sigma)^{\alpha+1}\gamma^{-1}(\sigma x+\gamma y-\gamma\sigma)^{-(\alpha+1)}\Big|_{y=\sigma}^{y=+\infty} = \\ = \alpha(\gamma\sigma)^{\alpha+1}\gamma^{-1}(\sigma x)^{-(\alpha+1)} = \alpha\gamma^{\alpha}x^{-(\alpha+1)}, \quad x \geq \gamma$$

Now everything is done to calculate the conditional PDF:

$$f_Y(y|x) = \frac{\alpha(\alpha+1)(\gamma\sigma)^{\alpha+1}(\sigma x+\gamma y-\gamma\sigma)^{-(\alpha+2)}}{\alpha\gamma^{\alpha}x^{-(\alpha+1)}} = (\alpha+1)\gamma(\sigma x)^{\alpha+1}\lambda^{-(\alpha+2)}, \quad y \geq \sigma$$

Then the conditional CDF is:

$$F_Y(y|x) = \int_{\sigma}^{y} (\alpha+1)\gamma(\sigma x)^{\alpha+1}(\sigma x+\gamma t-\gamma\sigma)^{-(\alpha+2)}\,dt = \\ = -(\sigma x)^{\alpha+1}(\sigma x+\gamma t-\gamma\sigma)^{-(\alpha+1)}\Big|_{t=\sigma}^{t=y} = \\ = -(\sigma x)^{\alpha+1}(\sigma x+\gamma y-\gamma\sigma)^{-(\alpha+1)} + (\sigma x)^{\alpha+1}(\sigma x)^{-(\alpha+1)} = \\ = 1-(\sigma x)^{\alpha+1}(\sigma x+\gamma y-\gamma\sigma)^{-(\alpha+1)}$$

Thus the optimization task is:

$$L = \prod_{i=1}^{r}\left(\alpha(\alpha+1)(\gamma\sigma)^{\alpha+1}\lambda^{-(\alpha+2)}\right)\prod_{j=r+1}^{n}\left((\sigma x)^{\alpha+1}(\sigma x+\gamma y-\gamma\sigma)^{-(\alpha+1)}\right) \to \max_{\gamma,\sigma,\alpha>0}$$

$$l = \sum_{i=1}^{r}\ln\left(\alpha(\alpha+1)(\gamma\sigma)^{\alpha+1}\lambda^{-(\alpha+2)}\right) + \sum_{j=r+1}^{n}\ln\left((\sigma x)^{\alpha+1}(\sigma x+\gamma y-\gamma\sigma)^{-(\alpha+1)}\right) \to \max_{\gamma,\sigma,\alpha>0}$$

Generating Samples

To generate samples from the bivariate Pareto distribution the Inverse Transform Sampling can be used. If X is a random variable with CDF $F_X(x)$:

- Generate u - a random sample from the standard uniform distribution $[0, 1]$
- Compute the inverse function of CDF $F_X^{-1}(x)$
- Calculate the desired sample as $x = F_X^{-1}(u)$

We are working with two-dimensional data thus it is rationale to make two steps described in the paper by S. Olver and A Townsend:[7]

(1) Get samples for x from its marginal CDF $F_X(x)$.

(2) Based on the value of x draw samples for y using its conditional CDF $F_Y(y|x)$.

$$F_X(x) = \int_\gamma^x \alpha\gamma^\alpha t^{-(\alpha+1)}\, dt = -\gamma^\alpha t^{-\alpha}\Big|_{t=\gamma}^{t=x} = 1 - \left(\frac{\gamma}{x}\right)^\alpha$$

$$F_X^{-1}(x) = \frac{\gamma}{x^{1/\alpha}} \quad \Rightarrow \quad x = \frac{\gamma}{u^{1/\alpha}}, \quad u \sim \mathrm{U}[0,1]$$

Moving on to the procedure of sampling the other variable:

$$F_Y(y|x) = 1 - \left(\frac{\sigma x}{\sigma x + \gamma y - \gamma\sigma}\right)^{\alpha+1} = 1 - \left(\frac{\frac{\sigma}{\gamma}x}{\frac{\sigma}{\gamma}x + y - \sigma}\right)^{\alpha+1}$$

$$F_Y^{-1}(y|x) = \sigma + \frac{\sigma}{\gamma}x\left(\frac{1}{y^{1/(\alpha+1)} - 1}\right)$$

$$y = \sigma + \frac{\sigma}{\gamma}x\left(\frac{1}{u^{1/(\alpha+1)} - 1}\right), \quad u \sim \mathrm{U}[0,1]$$

The results of sampling and finding MLE can be found in the code part. Figure 5 demonstrates the original sample distribution with right-censoring and the estimation results.

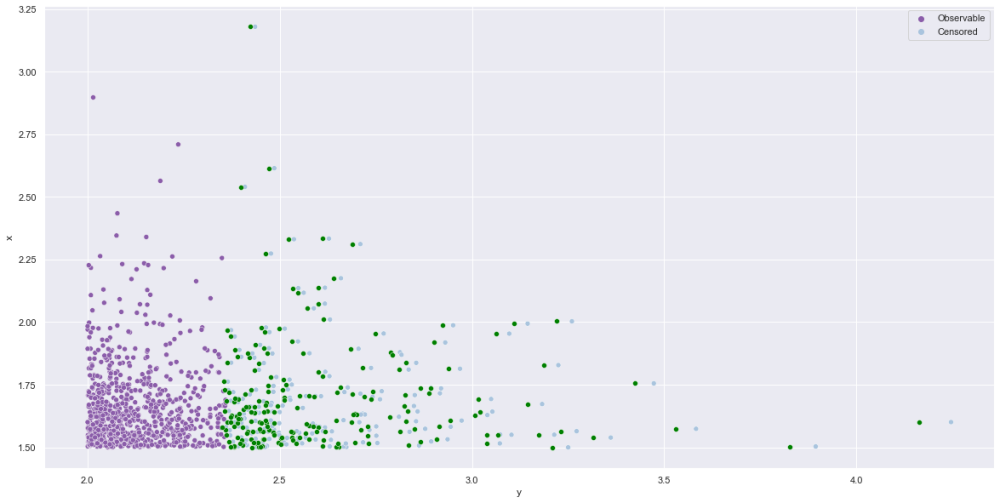

Figure 5: Data Censoring 20% , $\alpha = 10$, $\gamma = 1.5$, $\sigma = 2$, 1000 observations

Results and Further Research Problems

The current study concentrates on the maximum likelihood estimation of censored data. The main goals of the study have been achieved:

- The consistency and asymptotic normality of MLE are proven in the first chapter of the research.
- The likelihood function for the censored data is constructed and derived for Normal distribution showing strong results.
- The implementation of Nelder-Mead algorithm is successfully used to solve the optimization problem.
- The likelihood function to deal with censoring in case of two dimensional data is introduced and solved for bivariate Pareto distribution.

However there are a lot of things which can be the aim of future research. They include:

- Dealing with multidimensional data, including multidimensional normal distribution.
- Testing the algorithm on real data in scoring auction
- Applying non-parametric approach to deal with censored data

References

(1) P. Andreyanov, UCLA Job Market Paper, Mechanism Choice in Scoring Auctions

(2) Stamford University (2016), Introduction to Statistical Inference, Lecture 14 — Consistency and asymptotic normality of the MLE. Link

(3) N. Balakrishnan, M. Kateri (2008), On the maximum likelihood estimation of parameters of Weibull distribution based on complete and censored data

(4) J. A. Nelder, R. Mead (1965), A simplex method for function minimization

(5) P. Jiang, Y. Yang, G. Bierner, F. A. Li, R. Wang, A. Moghtaderi (2019), *Ranking in Genealogy: Search Results Fusion at Ancestry*

(6) N.H. Abdel-Alla,b, H.N. Abd-Ellah (2015), Geometric visualization of parallel bivariate Pareto distribution surfaces

(7) S. Olver, A.Townsend (2013), Fast inverse transform sampling in one and two dimensions