Ozonmasters

Statistics

Home work

Kozhemyak Vitaly

October 3, 2020

Contents

In this article we denote $X_1, \dots X_n$ as a sequence of i.i.d random variables and $x_1, \dots, x_n$ as a sample.

1 Task 1

1.1 Theoretical solution

Let $(1, 2, 3)$ be the events like "rock, scissors, paper". Consider $\eta, \eta_1, \eta_2 \sim P$ on $[0, 1]$ and random variables

$$X = \begin{cases} 1, & p, \\ 2, & q, \\ 3, & 1 - p - q, \end{cases}$$

$$Y = \begin{cases} 1, & \eta_{(1)}, \\ 2, & \eta_{(2)} - \eta_{(1)}, \\ 3, & 1 - \eta_{(2)}. \end{cases}$$

$$\mathbb{P}(1st\ win) = \mathbb{P}(\{X = 1, Y = 2\} \cup \{X = 2, Y = 3\} \cup \{X = 3, Y = 1\}) =$$
$$= \mathbb{P}(X = 1)\mathbb{P}(Y = 2) + \mathbb{P}(X = 2)\mathbb{P}(Y = 3) + \mathbb{P}(X = 3)\mathbb{P}(Y = 1) =$$

$$= p \int_0^1 \mathbb{P}(Y = 2 | \eta_{(2)} - \eta_{(1)} = x) f_{\eta_{(2)} - \eta_{(1)}}(x) dx +$$

$$+ q \int_0^1 \mathbb{P}(Y = 3 | 1 - \eta_{(2)} = x) f_{1 - \eta_{(2)}}(x) dx +$$

$$+ (1 - p - q) \int_0^1 \mathbb{P}(Y = 1 | \eta_{(1)} = x) f_{\eta_{(1)}}(x) dx =$$

$$= p\,\mathbb{E}[\eta_{(2)} - \eta_{(1)}] + q\,\mathbb{E}[1 - \eta_{(2)}] + (1 - p - q)\,\mathbb{E}[\eta_{(1)}] =$$

$$= (\mathbb{E}[\eta_{(2)}] - 2\,\mathbb{E}[\eta_{(1)}])p + (1 - \mathbb{E}[\eta_{(2)}] - \mathbb{E}[\eta_{(1)}])q + \mathbb{E}[\eta_{(1)}]. \quad (1)$$

Lemma 1. *Consider $X_1, \dots X_n$ as a sequence of i.i.d random variables with some distribution P. Then*

$$F_{X_{(n)}}(x) = F_X^n(x),$$
$$F_{X_{(1)}}(x) = 1 - (1 - F_X(x))^n$$

where X has the same distribution P on $[0, 1]$.

Proof. Let's deduce the **distribution of maximum**

$$F_{X_{(n)}}(x) = \mathbb{P}(X_{(n)} \leqslant x) = \mathbb{P}(X_1 \leqslant x, \ldots, X_n \leqslant x) =$$

$$= \prod_{i=1}^{n} \mathbb{P}(X_i \leqslant x) = F_X^n(x),$$

and **distribution of minimum**

$$F_{X_{(1)}}(x) = \mathbb{P}(X_{(1)} \leqslant x) = 1 - \mathbb{P}(X_{(1)} > x) = \mathbb{P}(X_1 > x, \ldots, X_n > x) =$$

$$= 1 - \prod_{i=1}^{n} \mathbb{P}(X_i > x) = 1 - \prod_{i=1}^{n} (1 - \mathbb{P}(X_i \leqslant x)) = 1 - (1 - F_X(x))^n.$$

□

Now we can calculate the expected values:

$$\mathbb{E}[\eta_{(1)}] = \int_{-\infty}^{+\infty} x f_{\eta_{(1)}}(x) dx = 2 \int_{-\infty}^{+\infty} x(1 - F_X(x)) f_X(x) dx = 2(\mathbb{E}[X] - \mathbb{E}[X F_X(X)]),$$

$$\mathbb{E}[\eta_{(2)}] = \int_{-\infty}^{+\infty} x f_{\eta_{(2)}}(x) dx = 2 \int_{-\infty}^{+\infty} x F_X(x) f_X(x) dx = 2\,\mathbb{E}[X F_X(X)].$$

Remember the last equation in (1) and make the substitution of expected values we have found above

$$\mathbb{P}(1st\ win) = (6\,\mathbb{E}[X F_X(X)] - 4\,\mathbb{E}[X])p + (1 - 2\,\mathbb{E}[X])q + 2(\mathbb{E}[X] - \mathbb{E}[X F_X(X)]).$$

We want to maximize $\mathbb{P}(1st\ win)$ in p, q, so consider the next linear optimization problem:

$$\begin{cases} \alpha p + \beta q \to \max\limits_{p,q}, \\ p + q \leqslant 1, \\ p \geqslant 0, q \geqslant 0. \end{cases}$$

The solution of this problem is

1. $\alpha < \beta \Rightarrow q = 1, p = 0$;
2. $\alpha > \beta \Rightarrow q = 0, p = 1$;
3. $\alpha = \beta \geqslant 0 \Rightarrow p + q = 1$;

4. $\alpha = \beta < 0 \Rightarrow p = q = 0$;

5. $\beta < \alpha = 0 \Rightarrow q = 0, p = [0, 1]$;

6. $\alpha < \beta = 0 \Rightarrow q = [0, 1], p = 0$.

So, put $\alpha = 6\,\mathbb{E}[XF_X(X)] - 4\,\mathbb{E}[X], \beta = 1 - 2\,\mathbb{E}[X]$. If the players have played n rounds before then we can estimate

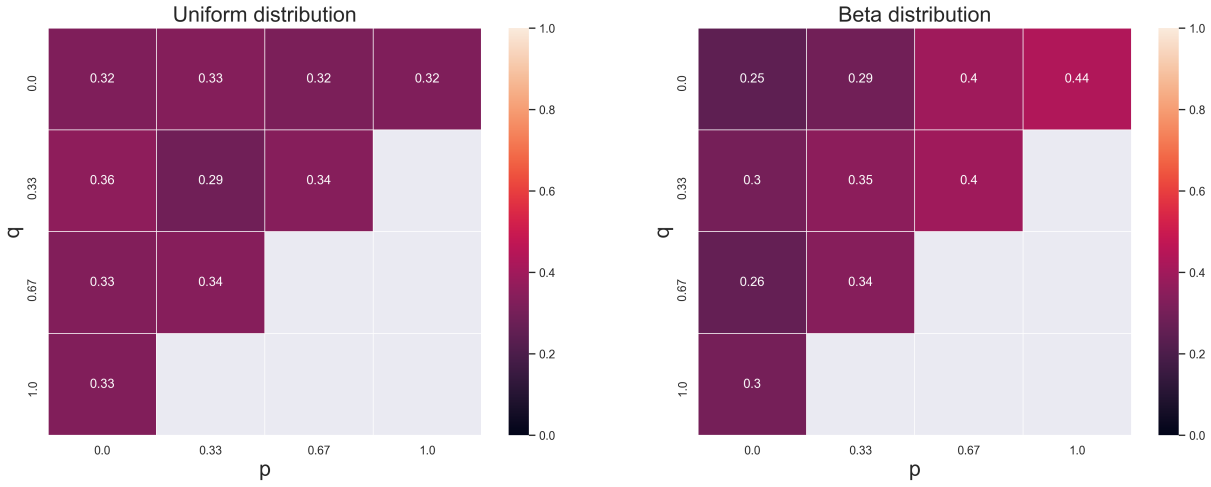
$$\mathbb{E}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Split last n games as follows $\{1, \dots m\} + \{m+1, \dots, n\}$. Then

$$F_X(x) \approx F_n(x) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{I}[x_j \leqslant x],$$

$$\mathbb{E}[XF_X(X)] = \frac{1}{n-m}\sum_{i=m+1}^{n} x_i F_n(x_i).$$

1.2 Numerical solution



2 Task 2

2.1 Theoretical solution

2.1.1 Problem 1

Consider $X_1, \dots X_n$ — i.i.d such that $X_i \sim U[0,1], \forall i = 1, \dots, n$. As we know

$$\mathbb{E}X_{(r)} = \int\limits_{-\infty}^{+\infty} x dF_{X_{(r)}}(x),$$

where $F_{X_{(r)}}(x)$ is CDF of $X_{(r)}$.

Theorem 1. *Let $F_X(x)$ be the common CDF of the r.v. X. For $r = 1, \dots, n$,*

$$F_{X_{(r)}}(x) = \mathbb{P}(X_{(k)} \leqslant x) = \sum_{k=r}^{n} C_n^k (F_X(x))^k (1 - F_X(x))^{n-k}.$$

Proof. Note, that in a sample $\{X_1, \dots, X_n\}$ of n independent values the number of observations not bigger than x has $Bin(n, F_X(x))$ distribution. By breaking the event $\{X_{(r)} \leqslant x\}$ into simple disjoint events, we get

$$\{X(r) \leqslant x\} = A_n(x) \bigcup \dots \bigcup A_r(x) =$$

$$= \{X_{(n)} \leqslant x\} \bigcup \{X_{(n-1)} \leqslant x, X_{(n)} > x\} \bigcup$$

$$\dots$$

$$\bigcup \{X_{(r)} \leqslant x, X_{(n)} > x, \dots X_{(r-1)} > x\}.$$

Then

$$F_{X_{(r)}}(x) = \mathbb{P}(X_{(r)} \leqslant x) = \mathbb{P}\left(\bigcup_{k=r}^{n} A_k(x)\right) = \sum_{k=r}^{n} \mathbb{P}(A_k(x)) =$$

$$= \sum_{k=r}^{n} C_n^k (F_X(x))^k (1 - F_X(x))^{n-k}.$$

□

We also know that

$$\frac{dF_X(x)}{dx} = f_X(x),$$

where $f_X(x)$ is density function of X. If $X \sim U[0,1]$, then we derive the next equation

$$\frac{dF_{X_{(r)}}(x)}{dx} = \sum_{k=r}^{n} C_n^k [kx^{k-1}(1-x)^{n-k} - (n-k)x^k(1-x)^{n-k-1}] = f_{X_{(r)}}(x).$$

Finally, we have

$$\mathbb{E}\, X_{(r)} = \int\limits_{-\infty}^{+\infty} x f_{X_{(r)}}(x)dx =$$

$$= \sum_{k=r}^{n} kC_n^k \int\limits_0^1 [x^k(1-x)^{n-k}]dx - \sum_{k=r}^{n}(n-k)C_n^k \int\limits_0^1 [x^{k+1}(1-x)^{n-k-1}]dx =$$

$$= \sum_{k=r}^{n} kC_n^k B(k+1, n-k+1) - \sum_{k=r}^{n}(n-k)C_n^k B(k+2, n-k) =$$

$$= \sum_{k=r}^{n} \frac{k \cdot n!}{(n-k)!k!} \frac{k!(n-k)!}{(n+1)!} - \sum_{k=r}^{n} \frac{(n-k)\cdot n!}{(n-k)!k!} \frac{(k+1)!(n-k-1)!}{(n+1)!} =$$

$$= \frac{r}{n+1}.$$

2.1.2 Problem 2

We derive the next equation using the result from **Problem 1**:

$$\mathbb{E}\, X_{(r)}^2 = \int\limits_{-\infty}^{+\infty} x^2 f_{X_{(r)}}(x)dx =$$

$$= \sum_{k=r}^{n} kC_n^k \int\limits_0^1 [x^{k+1}(1-x)^{n-k}]dx - \sum_{k=r}^{n}(n-k)C_n^k \int\limits_0^1 [x^{k+2}(1-x)^{n-k-1}]dx =$$

$$= \sum_{k=r}^{n} kC_n^k B(k+2, n-k+1) - \sum_{k=r}^{n}(n-k)C_n^k B(k+3, n-k) =$$

$$= \sum_{k=r}^{n} \frac{k \cdot n!}{(n-k)!k!} \frac{(k+1)!(n-k)!}{(n+2)!} - \sum_{k=r}^{n} \frac{(n-k)\cdot n!}{(n-k)!k!} \frac{(k+2)!(n-k-1)!}{(n+2)!} =$$

$$= \frac{r(r+1)}{(n+1)(n+2)}.$$

2.1.3 Problem 3

Transform the next equation

$$f_{X_{(r)}}(x) = \sum_{k=r}^{n} C_n^k [kx^{k-1}(1-x)^{n-k} - (n-k)x^k(1-x)^{n-k-1}] =$$

$$= C_n^{k-1} x^{r-1}(1-x)^{n-r} + \sum_{k=r+1}^{n} C_n^k [kx^{k-1}(1-x)^{n-k}] - \sum_{k=r}^{n-1} C_n^k [(n-k)x^k(1-x)^{n-k-1}].$$

The last two terms above cancel, since using the change of variables $j = k-1$. Now we can find the maximum of $f_{X_{(r)}}(x)$:

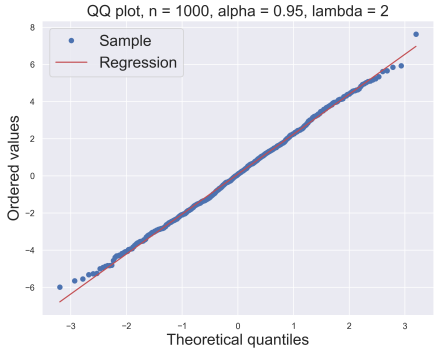
$$\frac{df_{X_{(r)}}(x)}{dx} = C_n^{k-1}[(r-1)x^{r-2}(1-x)^{n-r} - (n-r)x^{r-1}(1-x)^{n-r-1}] = 0.$$

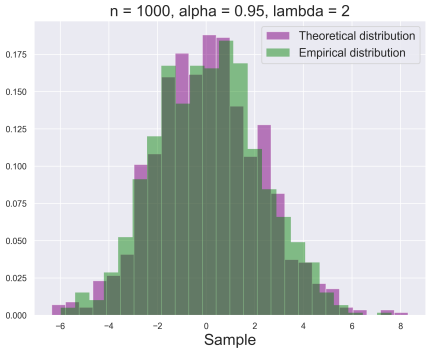
The solutions of this equation are

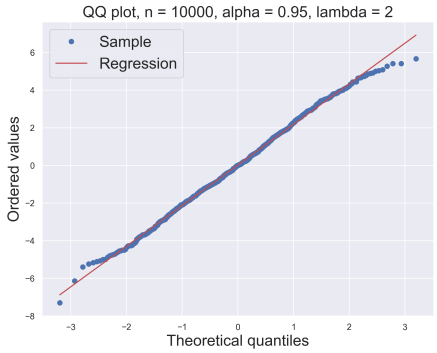
$$x = 0, x = 1, x = \frac{r-1}{n-1},$$

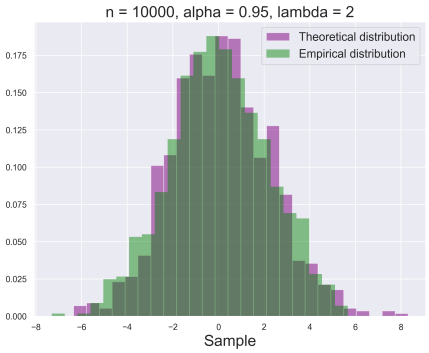
but the maximum of $f_{X_{(r)}}(x)$ is attained on $x = \frac{r-1}{n-1}$.

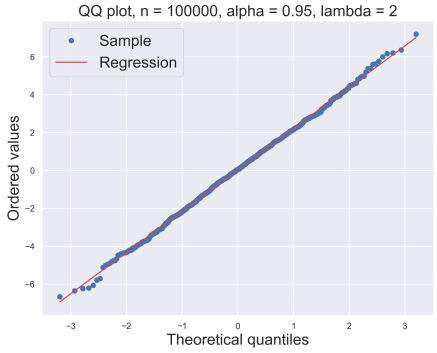
2.2 Numerical solution

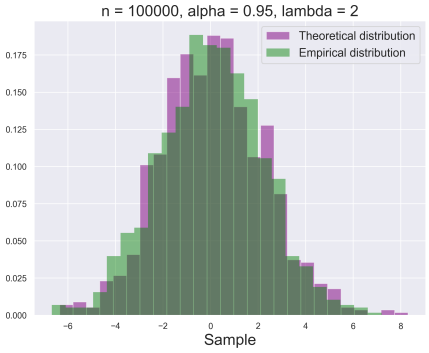












3 Task 3

3.1 Theoretical solution

3.1.1 Problem 1 (MLE method)

Let $X_1, \ldots, X_n$ be a sequence of i.i.d random variables. Consider log-likelihood function

$$l(\mu, \sigma) = \log p(x_1, \ldots, x_n; \mu, \sigma) = \sum_{i=1}^{n} \log p(x_i; \mu, \sigma) \to \max_{\mu,\sigma}.$$

- We find the next estimation $\hat{\sigma}_n$ as follows:

$$\frac{dl(\mu, \sigma)}{d\mu} = \sum_{i=1}^{n} \left[-\frac{1}{\sigma} + \frac{|x_i - \mu|}{\sigma^2} \right] = 0,$$

$$\hat{\sigma}_n = \frac{1}{n} \sum_{i=1}^{n} |x_i - \mu|.$$

- We find the next estimation $\hat{\mu}_n$ as follows. The likelihood function

$$\mathcal{L}(\mu, \sigma) = \frac{1}{(2\sigma)^n} \exp\left(-\frac{1}{\sigma} \sum_{i=1}^{n} |x_i - \mu| \right)$$

is supposed to be maximized $\Leftrightarrow$ when $\sum\limits_{i=1}^{n} |x_i - \mu|$ is supposed to be minimized.

Theorem 2. *Let m be the sample median, then then next inequality*

$$\sum_{i=1}^{n} |x_i - \mu| \geqslant \sum_{i=1}^{n} |x_i - m|$$

holds for every value of μ

Proof. Consider two cases.

1. Let n be even. Let m be any real number between the two middle values so that we have

$$x_{(1)} \leqslant \ldots \leqslant x_{(\frac{n}{2})} \leqslant \mu \leqslant x_{(\frac{n}{2}+1)} \leqslant \ldots \leqslant x_{(n)}.$$

For $\left\{ i = 1, \ldots, \dfrac{n}{2} \right\}$ the triangle inequality gives us

$$|x_{(i)} - m| + |m - x_{(n-i+1)}| = |x_{(n-i+1)} - x_{(i)}| \leqslant |x_{(i)} - \mu| + |\mu - x_{(n-i+1)}|.$$

Hence,

$$\sum_{i=1}^{\frac{n}{2}} |x_{(i)} - m| + |m - x_{(n-i+1)}| \leqslant \sum_{i=1}^{\frac{n}{2}} |x_{(i)} - \mu| + |\mu - x_{(n-i+1)}|$$

and

$$\sum_{i=1}^{\frac{n}{2}} |m - x_{(n-i+1)}| = \sum_{i=\frac{n}{2}+1}^{n} |m - x_{(i)}|,$$

$$\sum_{i=1}^{\frac{n}{2}} |\mu - x_{(n-i+1)}| = \sum_{i=\frac{n}{2}+1}^{n} |\mu - x_{(i)}|.$$

As a result we have

$$\sum_{i=1}^{n} |m - x_{(i)}| \leqslant \sum_{i=1}^{n} |\mu - x_{(i)}|.$$

2. Let n be odd. Then $m = x_{(\frac{n+1}{2})}$ is the sample median:

$$x_{(1)} \leqslant \ldots \leqslant x_{(\frac{n+1}{2})} \leqslant \ldots \leqslant x_{(n)}.$$

For $i = \left\{1, \ldots, \dfrac{n+1}{2}\right\}$ the triangle inequality gives us

$|x_{(i)} - x_{(\frac{n+1}{2})}| + |x_{(\frac{n+1}{2})} - x_{(n-i+1)}| = |x_{(n-i+1)} - x_{(i)}| \leqslant |x_{(i)} - \mu| + |\mu - x_{(n-i+1)}|.$

Hence,

$$\sum_{i=1}^{\frac{n+1}{2}} |x_{(i)} - x_{(\frac{n+1}{2})}| + |x_{(\frac{n+1}{2})} - x_{(n-i+1)}| \leqslant \sum_{i=1}^{\frac{n+1}{2}} |x_{(i)} - \mu| + |\mu - x_{(n-i+1)}|$$

and

$$\sum_{i=1}^{\frac{n+1}{2}} |x_{(\frac{n+1}{2})} - x_{(n-i+1)}| = \sum_{i=\frac{n+1}{2}}^{n} |x_{(\frac{n+1}{2})} - x_{(i)}|,$$

$$\sum_{i=1}^{\frac{n+1}{2}} |\mu - x_{(n-i+1)}| = \sum_{i=\frac{n+1}{2}}^{n} |\mu - x_{(i)}|.$$

As a result we have

$$\sum_{i=1}^{n} |x_{(\frac{n+1}{2})} - x_{(i)}| \leqslant \sum_{i=1}^{n} |\mu - x_{(i)}|.$$

□

The estimation parameters are

$$\hat{\mu}_n = median\{x_1, \ldots, x_n\},$$

$$\hat{\sigma}_n = \frac{1}{n}\sum_{i=1}^{n} |x_i - \hat{\mu}_n|.$$

3.1.2 Problem 2 (Method of moments)

Consider $X_1, \ldots, X_n$ — i.i.d such that $X_i \sim Laplace(\mu, \sigma), \forall i = 1, \ldots, n$. Then we need to solve the next system

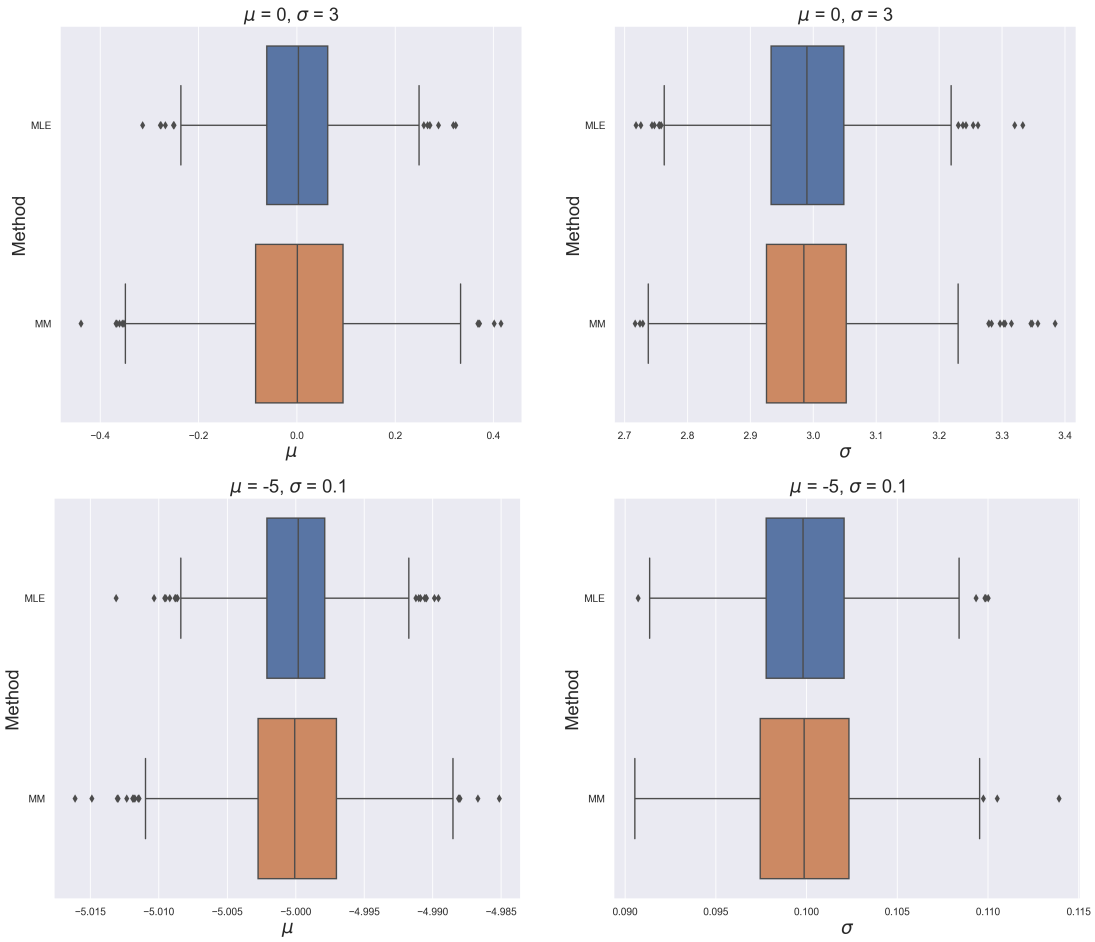
$$\begin{cases} \mu = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i, \\ \mu^2 + 2\sigma^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i^2, \end{cases}$$

from which we derive that

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$\hat{\sigma}_n = \sqrt{\frac{1}{2n}\sum_{i=1}^{n} x_i^2 - \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2}.$$

3.2 Numerical solution



We can see that MLE method is better a little bit than MM method.

4 Task 4

4.1 Theoretical solution

4.1.1 Problem 1

Let's calculate the density function of ξ^2 as $f_{\xi^2}(x) = \dfrac{dF_{\xi^2}(x)}{dx}$. Firstly,

$$F_{\xi^2}(x) = \mathbb{P}(\xi^2 \leqslant x) = \mathbb{P}(|\xi| \leqslant \sqrt{x}) = \int\limits_{-\sqrt{x}}^{\sqrt{x}} f_\xi(t)dt =$$

$$f_{\xi^2}(x) = \frac{f_\xi(\sqrt{x}) + f_\xi(-\sqrt{x})}{2\sqrt{x}} = \{\xi \sim \mathcal{N}(0,\theta)\} = \frac{f_\xi(\sqrt{x})}{\sqrt{x}}.$$

Now we can decompose the density function $f_{\xi^2}(x)$ as follows:

$$f_{\xi^2}(x) = \frac{1}{\sqrt{x}}\frac{1}{\sqrt{2\pi\theta}}\exp\left(-\frac{x}{2\theta}\right) = h(x)\cdot\exp\left(vx - g(v)\right))$$

where

$$h(x) = \frac{1}{\sqrt{2\pi x}},$$

$$v = -\frac{1}{2\theta},$$

$$g(v) = \log\left(\sqrt{-\frac{1}{2v}}\right) = -\frac{1}{2}\log(-2v).$$

Thus, $P_{\sigma^2} = \{Law(\xi^2),\ \xi \sim \mathcal{N}(0,\theta)\}$ is an exponential family.

4.1.2 Problem 2

Consider $X \sim P_\theta$. Simple derivation gives us

$$\mathbb{E}[X] = g'(v) = -\frac{1}{2v} = \theta,$$

$$\mathrm{Var}[X] = g''(v) = \frac{1}{2v^2} = 2\theta^2.$$

MLE method. Construct the log-likelihood function

$$l(\theta) = \log p(x_1,\dots,x_n;\theta) = \sum_{i=1}^{n}\log p(x_i;\theta)$$

which is supposed to be maximized by θ. Using optimality condition we get

$$\sum_{i=1}^{n}\frac{d}{d\theta}\left[-\frac{1}{2}\log(2\pi x_i) - \frac{1}{2}\log(\theta) - \frac{x_i}{2\theta}\right] = 0,$$

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

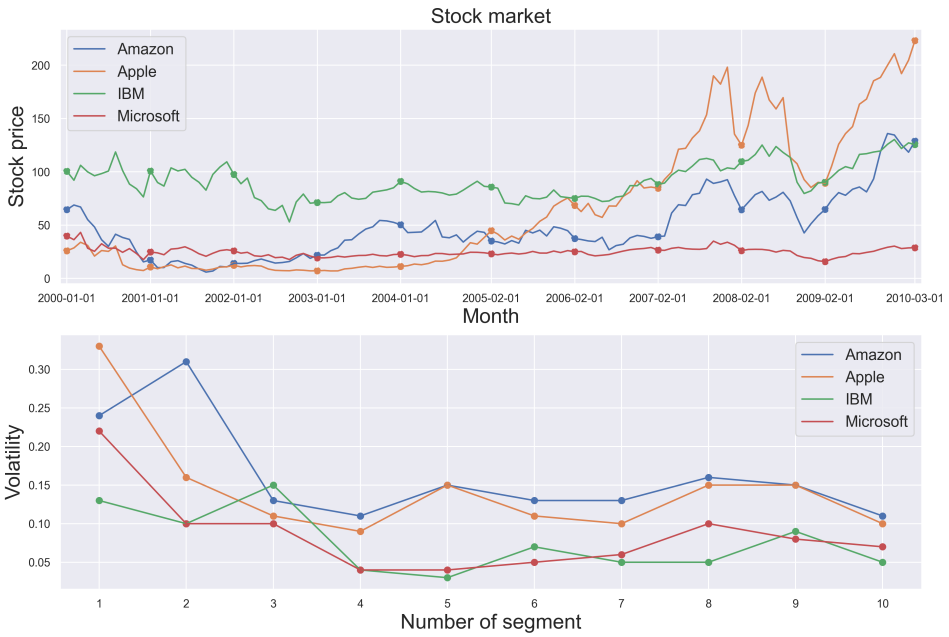
Method of moments. Using the result above we derive the first moment

$$\theta = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

from which we get

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

4.2 Numerical solutions



We split the top graph with points and now we have 10 time segments. Each point on bottom graph represents volatility parameter on corresponding interval. As we can see the volatility of Amazon and Apple companies higher than the volatility of IBM and Microsoft.

5 Appendix

- Code can be found: here;
- The data for **Task 4** was taken from here.