Ozonmasters

Statistics

Home work

Kozhemyak Vitaly

October 18, 2020

Contents

1 Task 1

1.1 Theoretical solution

The next functional in kernel K :

$$J(K) = \left(\int\limits_{-\infty}^{+\infty} K^2(x)dx\right)\left(\int\limits_{-\infty}^{+\infty} x^2 K(x)dx\right)^{1/2}$$

is called *kernel efficiency.*

1. Consider $K_1(x) = \mathbb{I}[x \in [-1/2, 1/2]]$. Then

$$J(K_1) = \left(\int\limits_{-1/2}^{+1/2} dx\right)\left(\int\limits_{-1/2}^{+1/2} x^2 dx\right)^{1/2} = \left(\frac{1}{12}\right)^{1/2}.$$

2. Consider $K_2(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. Then

$$J(K_2) = \left(\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} e^{-x^2} dx\right)\left(\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} x^2 e^{-\frac{x^2}{2}} dx\right)^{1/2} = \frac{1}{2\sqrt{\pi}}.$$

3. Consider $K_{ep}(x) = \frac{3}{4}(1 - x^2)\mathbb{I}[|x| \leqslant 1]$. Then

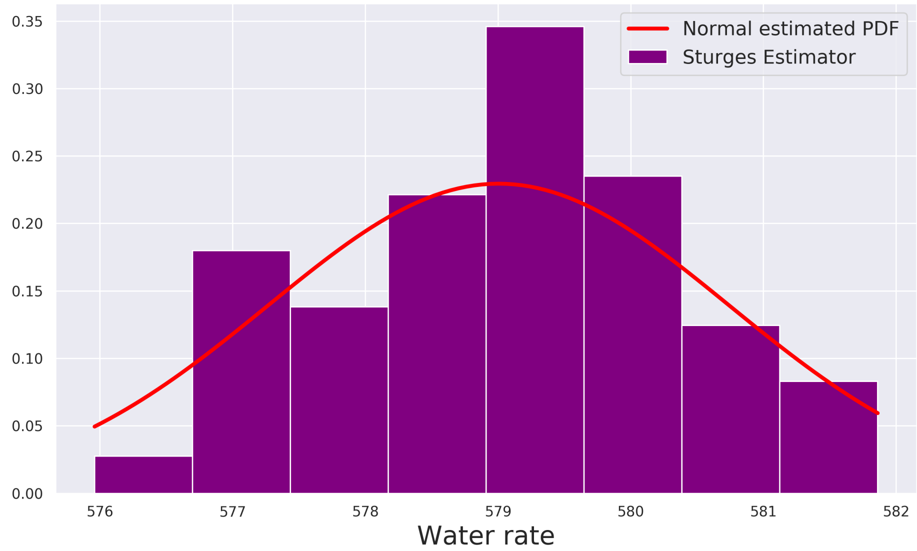
$$J(K_{ep}) = \left(\frac{9}{16}\int\limits_{-1}^{1} (1 - x^2)^2 dx\right)\left(\frac{3}{4}\int\limits_{-1}^{1} x^2(1 - x^2)dx\right)^{1/2} = \frac{3}{5\sqrt{5}}.$$

Let's calculate efficiency relative to the Epanechnikov kernel

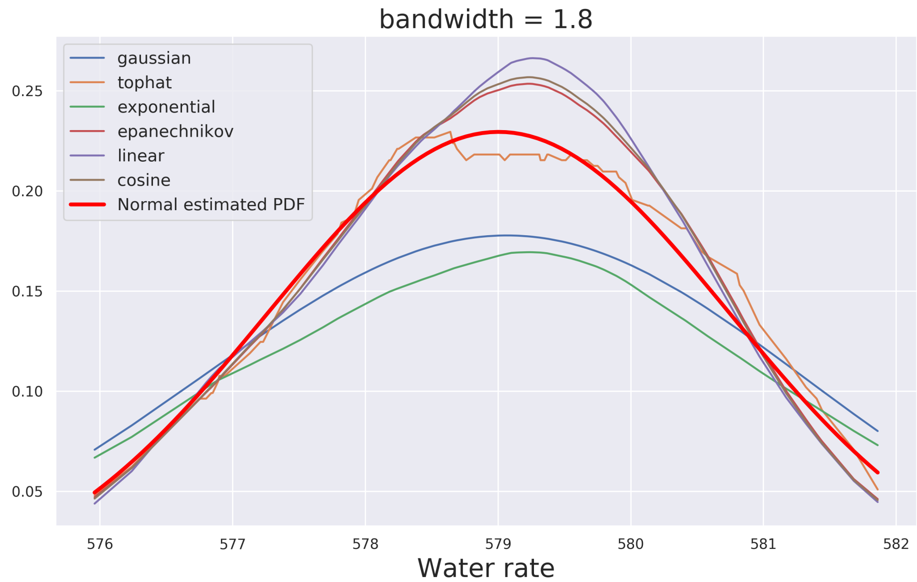
- For $K_1(x)$ it's $\frac{J(K_{ep})}{J(K_1)} \cdot 100\% = 92.9\%$;
- For $K_2(x)$ it's $\frac{J(K_{ep})}{J(K_2)} \cdot 100\% = 95.1\%$.

1.2 Numerical solution

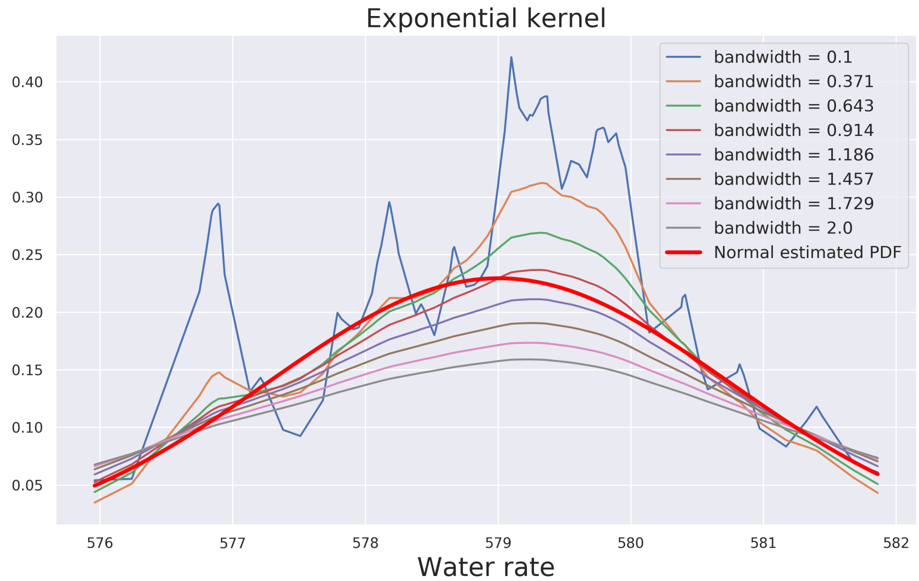
The next picture shows us the water rate distribution using Sturges estimator.



To show the bias-variance trade off you should visit the next link.
The next picture shows the different kernel estimators



Now we are going to choose the best exponential density estimator w.r.t. bandwidth.



We can see that the optimal bandwidth ≈ 1.

2 Task 2

2.1 Theoretical solution

Suppose we have random variable X drawn from distribution with the density function

$$p(x) = \frac{1}{2}\mathcal{N}(x; 0, 1) + \frac{1}{4}\mathcal{E}(x + 2; 1) + \frac{1}{4}\mathcal{E}(-x + 2; 1). \tag{1}$$

Expected value. According to the definition of expected value

$$\mathbb{E}[X] = \int\limits_{-\infty}^{+\infty} xp(x)dx = \frac{1}{2}\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} xe^{-\frac{1}{2}x^2}dx + \frac{1}{4}\int\limits_{-2}^{+\infty} xe^{-(x+2)}dx + \frac{1}{4}\int\limits_{-\infty}^{2} xe^{-(-x+2)}dx = 0.$$

Variance. According to the definition of variance

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mathbb{E}[X^2] =$$

$$= \int\limits_{-\infty}^{+\infty} x^2p(x)dx = \frac{1}{2}\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} x^2e^{-\frac{1}{2}x^2}dx + \frac{1}{4}\int\limits_{-2}^{+\infty} x^2e^{-(x+2)}dx + \frac{1}{4}\int\limits_{-\infty}^{2} x^2e^{-(-x+2)}dx =$$

$$= \frac{1}{2} + 1 + 1 = \frac{5}{2}.$$

Characteristic function. According to the definition of characteristic function

$$\varphi_X(t) = \mathbb{E}\, e^{itX} =$$

$$= \int_{-\infty}^{+\infty} e^{itx} p(x) dx = \frac{1}{2}\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{itx} e^{-\frac{1}{2}x^2} dx + \frac{1}{4} \int_{-2}^{+\infty} e^{itx} e^{-(x+2)} dx + \frac{1}{4} \int_{-\infty}^{2} e^{itx} e^{-(-x+2)} dx =$$

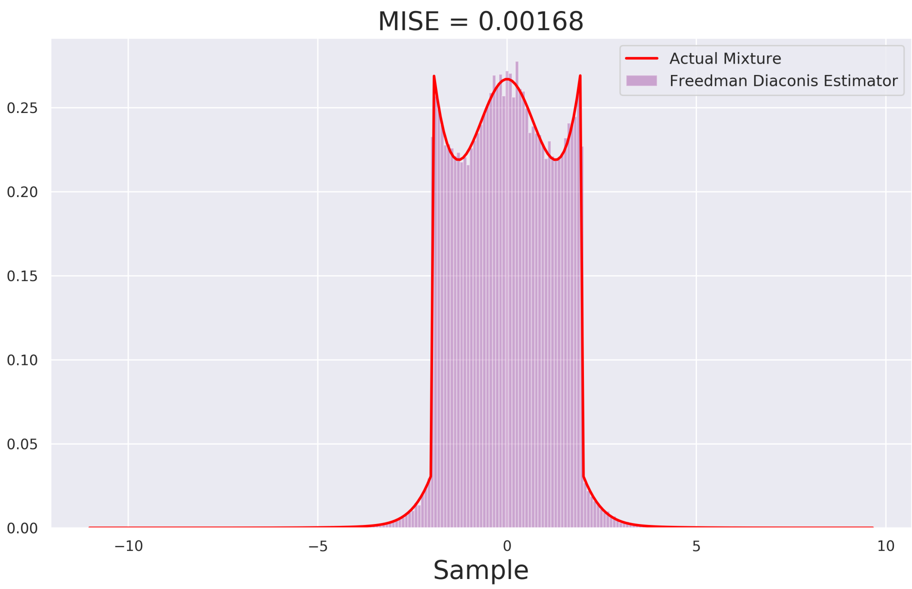
$$= \frac{1}{2} e^{-\frac{1}{2}t^2} + \frac{1}{4} e^{-2} \left[\int_{-2}^{+\infty} \cos(tx) e^{-x} dx + i \int_{-2}^{+\infty} \sin(tx) e^{-x} dx \right] -$$

$$- \frac{1}{4} e^{-2} \left[\int_{-\infty}^{2} \cos(tx) e^{x} dx + i \int_{-\infty}^{2} \sin(tx) e^{x} dx \right] = \frac{1}{2} e^{-\frac{1}{2}t^2} + \frac{i}{2} e^{-2} \int_{-2}^{+\infty} \sin(tx) e^{-x} dx =$$

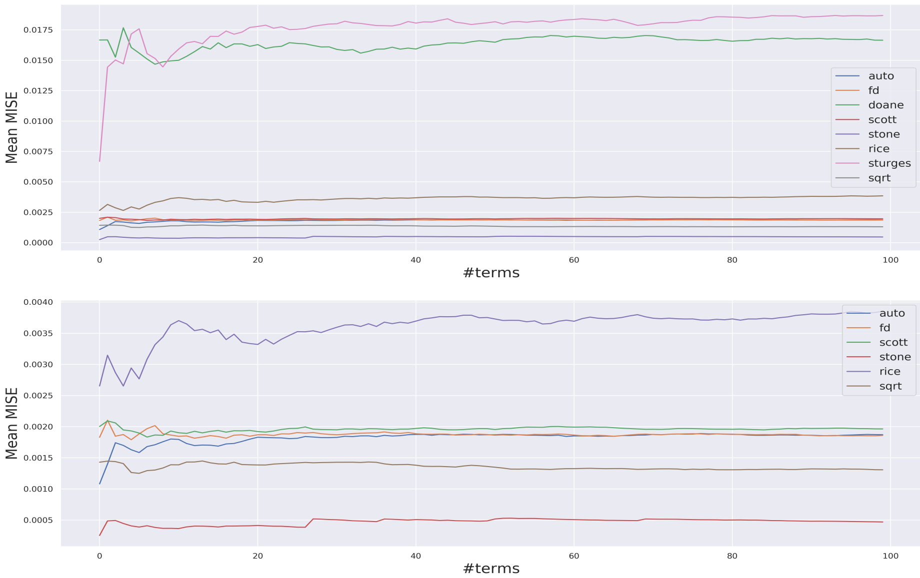
$$= \frac{1}{2} e^{-\frac{1}{2}t^2} + \frac{i}{2} e^{-2} \frac{t}{t^2 + 1}$$

2.2 Numerical solution

The following picture shows the histogram of the actual mixture (1) using Freedman-Diaconis rule with number of samples $n = 100000$

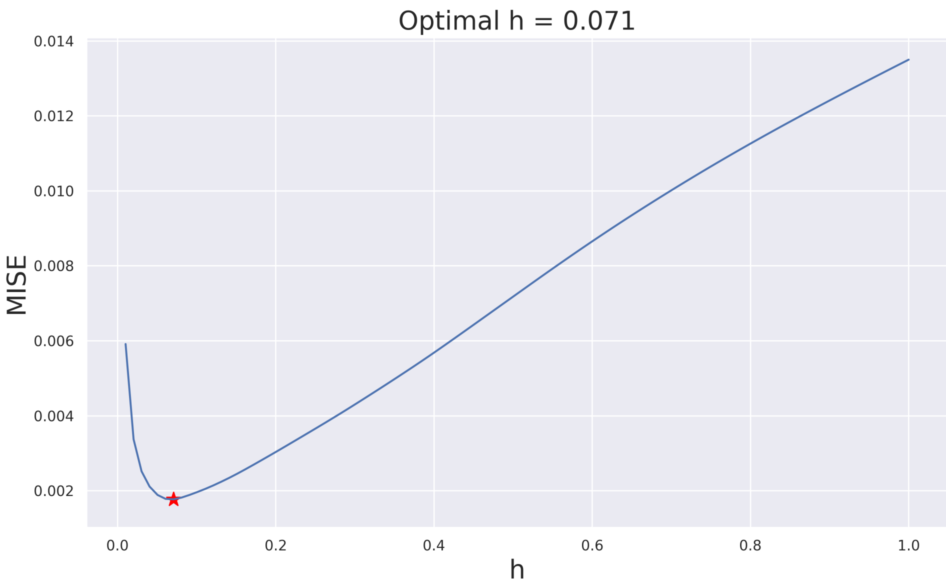


The following picture shows quality of different bin edge estimators

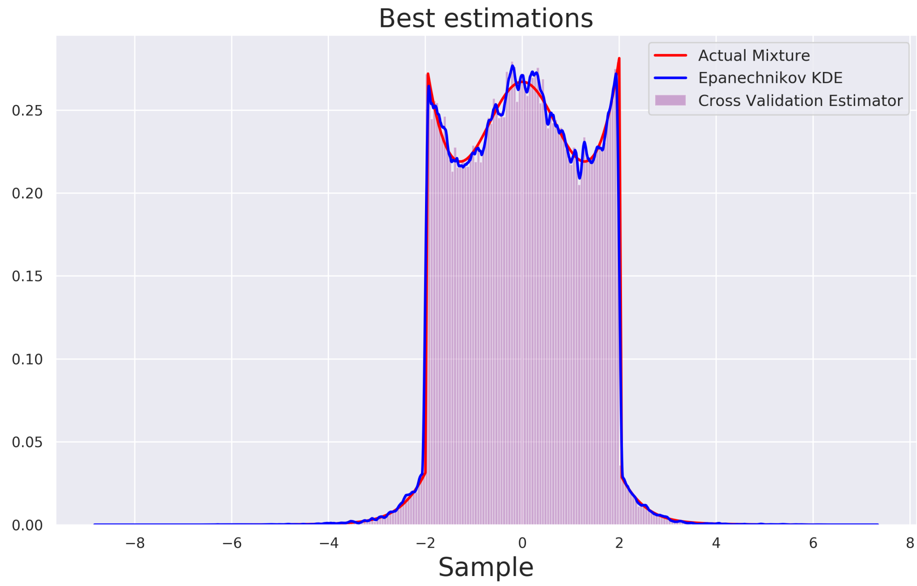


We can see that **Cross validation estimator (leave-one-out) a.k.a "stone"** gives us the best quality in sense of MISE. Use this article to get an acquaintance with other methods.

The following picture shows the optimal bandwidth



Now we have the finally result



3 Task 3

3.1 Theoretical solution

Consider the mixture model. Suppose that the observations X_i come from the next Gaussian mixture model

$$\mathbb{P}(X_i = x) = p(x) = \sum_{k=1}^{K} \pi_k p(x; \mu_k, \sigma_k^2)$$

with K mixture components. We want to maximize the likelihood function

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = p(x_1, \ldots, x_n) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k p(x_i; \mu_k, \sigma_k^2) \to \max_\theta$$

where $\theta = \{\mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2, \pi_1, \ldots, \pi_K\}$.

Consider the log-likelihood function

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k p(x_i; \mu_k, \sigma_k^2) \right) \to \max_\theta$$

which is supposed to be maximized by θ. It's hard problem to maximize $\log(\sum \ldots)$. But we can remember that we have the latent variables $Y_1, \ldots, Y_n$, so we can rewrite the likelihood function as follows

$$p(x_1,\dots,x_n)=\prod_{i=1}^{n}\sum_{k=1}^{K}\pi_k p(x_i;\mu_k,\sigma_k^2)=\prod_{i=1}^{n}\prod_{k=1}^{K}\pi_k^{\mathbb{I}[Y_i=k|X_i=x_i]}p(x_i;\mu_k,\sigma_k^2)^{\mathbb{I}[Y_i=k|X_i=x_i]},$$

where $Y_i\in\{1,\dots,K\}$ represents the mixture component for X_i and $\mathbb{P}(X_i=x_i|Y_i=k)$ is the mixture component.

Since

$$\mathcal{L}(\theta)=\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{I}(Y_i=k|X_i=x_i)(\log\pi_k+\log p(x_i;\mu_k,\sigma_k^2))$$

is a random variable, so expected value should be applied

$$\mathbb{E}_{Y_i}\mathcal{L}(\theta)=\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{P}(Y_i=k|X_i=x_i)(\log\pi_k+\log p(x_i;\mu_k,\sigma_k^2)).$$

Notice also that

$$\gamma_k(Y_i):=\mathbb{P}(Y_i=k|X_i=x_i)=\frac{\mathbb{P}(X_i=x_i|Y_i=k)\mathbb{P}(Y_i=k)}{\mathbb{P}(X_i=x_i)}=\frac{\pi_k p(x_i;\mu_k,\sigma_k^2)}{\sum\limits_{k=1}^{K}\pi_k p(x_i;\mu_k,\sigma_k^2)}.$$

Now we can differentiate $\mathbb{E}_{Y_i}\mathcal{L}(\theta)$ by μ_k,σ_k^2 and get

$$\hat{\mu}_k=\frac{\sum\limits_{i=1}^{n}\gamma_k(Y_i)x_i}{\sum\limits_{i=1}^{n}\gamma_k(Y_i)},$$

$$\hat{\sigma}_k^2=\frac{\sum\limits_{i=1}^{n}\gamma_k(Y_i)(x_i-\mu_k)^2}{\sum\limits_{i=1}^{n}\gamma_k(Y_i)}.$$

To find π_k we need to consider the next problem

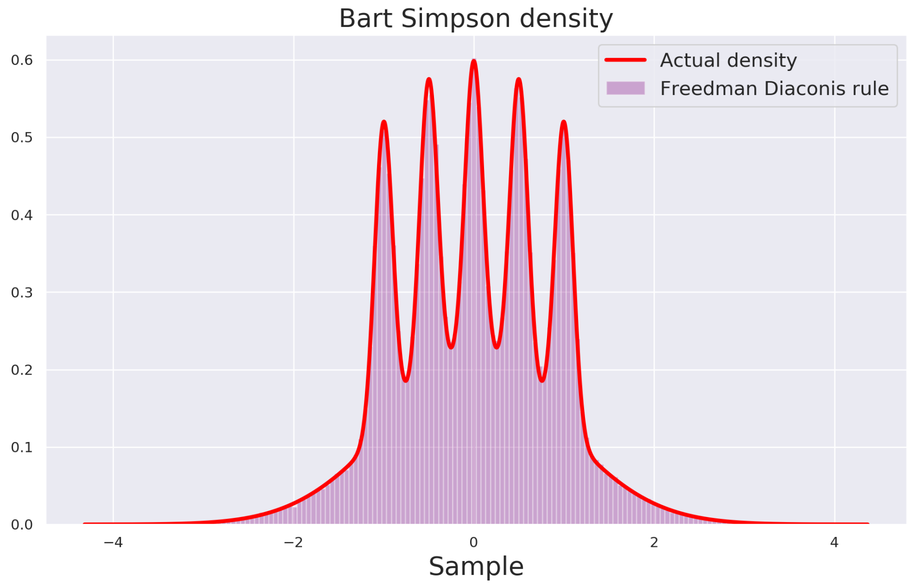
$$\begin{cases}\sum\limits_{i=1}^{n}\sum\limits_{k=1}^{K}\gamma_k(Y_i)\log\pi_k\to\max\limits_{\pi_k\geqslant 0},\\ \sum\limits_{k=1}^{K}\pi_k=1.\end{cases}$$

The KKT theorem gives us the solution

$$\hat{\pi}_k=\frac{1}{n}\sum_{i=1}^{n}\gamma_k(Y_i).$$

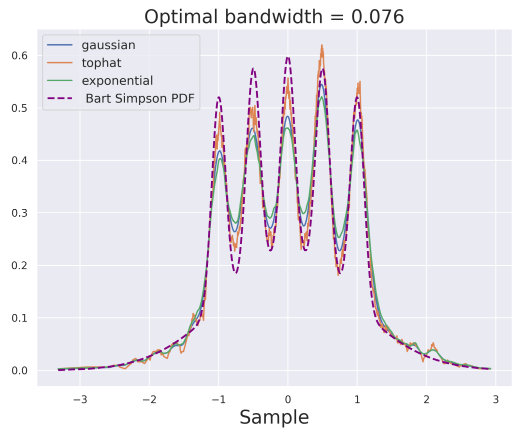
3.2 Numerical solution

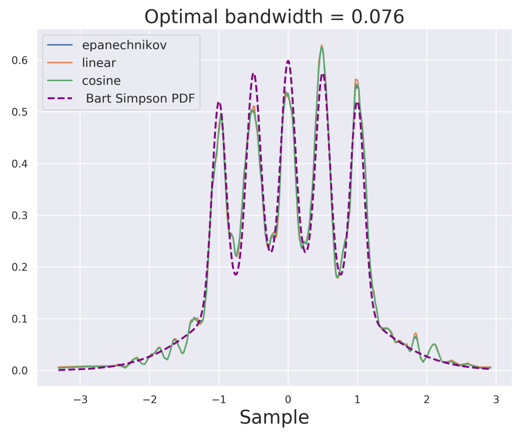
Generate $X_1, \dots, X_n \sim p_{BS}(x)$ and plot histogram using Freedman Diaconis rule.



To demonstrate how EM-algorithm is fitted w.r.t. different number of components visit this link.

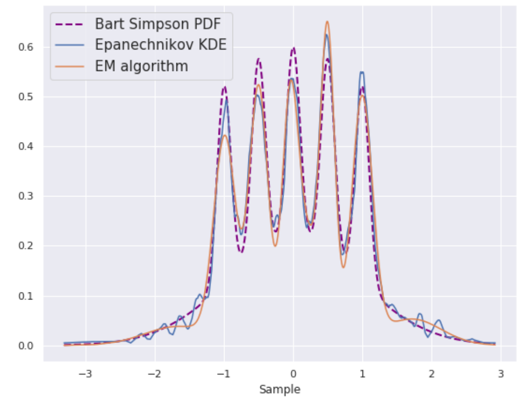
To choose optimal `bandwidth` Cross Validation with maximize likelihood is used.

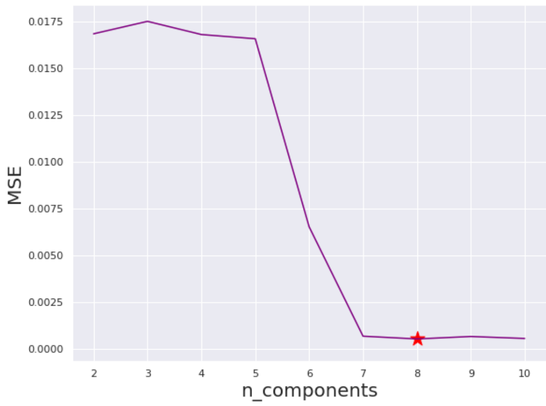




To determinate number of components for EM-algorithm the Epanechnikov KDE was chosen. We are going to minimize the next functional

$$\frac{1}{J}\sum_{i=1}^{J}(\hat{p}_n^{EM}(x_i) - \hat{p}_n^{K}(x_i))^2 \to \min_{n\_components}.$$





4 Task 4

To show that the minimum of the functional

$$J(K) = \left(\int\limits_{-\infty}^{+\infty} K^2(x)dx\right)^{4/5} \cdot \left(\int\limits_{-\infty}^{+\infty} x^2 K(x)dx\right)^{2/5} \to \min_{K\in\mathcal{C}}$$

is attained on

$$K^*(x) = \frac{3}{4}(1-x^2)\cdot \mathbb{I}[|x|\leqslant 1]$$

where

$$\mathcal{C} = \left\{K : K(x) = K(-x) \geqslant 0, \int\limits_{-\infty}^{+\infty} K(x)dx = 1\right\}, \tag{2}$$

we are going to use the calculus of variations.

Suppose $K^*(x)$ is a minimizer of the functional. Let $\varphi(x)$ be a continuous function such that $\varphi \in \mathcal{C}$. Consider the variation $\delta K^*(x) = \dfrac{K^*(x)+s\varphi(x)}{1+s} \equiv K(x)$ that is also in the admissible set, e.g. $K \in \mathcal{C}$.

Denote

$$g(s) = \left(\int\limits_{-\infty}^{+\infty}\left(\frac{K^*(x)+s\varphi(x)}{1+s}\right)^2 dx\right)^{4/5} \left(\int\limits_{-\infty}^{+\infty} x^2\left(\frac{K^*(x)+s\varphi(x)}{1+s}\right)dx\right)^{2/5}.$$

Hence $g(s)$ obtains a local minimum as $s=0$ we have $g'(0)=0$. Differentiate w.r.t

s we have

$$g'(0) = \frac{4}{5}\left(\int\limits_{-\infty}^{+\infty}(K^*(x))^2dx\right)^{-1/5}\left(2\int\limits_{-\infty}^{+\infty}K^*(x)(\varphi(x)-K^*(x))dx\right)\left(\int\limits_{-\infty}^{+\infty}x^2K^*(x)dx\right)^{2/5}+$$

$$+\frac{2}{5}\left(\int\limits_{-\infty}^{+\infty}(K^*(x))^2dx\right)^{4/5}\left(\int\limits_{-\infty}^{+\infty}x^2(\varphi(x)-K^*(x))dx\right)^{-3/5}\left(\int\limits_{-\infty}^{+\infty}x^2\varphi(x)dx\right)=0.$$

Some calculus give us

$$4\left(\int\limits_{-\infty}^{+\infty}K^*(x)(\varphi(x)-K^*(x))dx\right)\left(\int\limits_{-\infty}^{+\infty}x^2K^*(x)dx\right)+$$

$$+\left(\int\limits_{-\infty}^{+\infty}(K^*(x))^2dx\right)\left(\int\limits_{-\infty}^{+\infty}x^2(\varphi(x)-K^*(x))dx\right)=0,$$

$$\int\limits_{-\infty}^{+\infty}\varphi(x)\left[4K^*(x)\left(\int\limits_{-\infty}^{+\infty}t^2K^*(t)dt\right)+x^2\left(\int\limits_{-\infty}^{+\infty}(K^*(t))^2dt\right)-\right.$$

$$\left.-5\left(\int\limits_{-\infty}^{+\infty}t^2K^*(t)dt\right)\left(\int\limits_{-\infty}^{+\infty}(K^*(t))^2dt\right)\right]dx=0.$$

This equation holds for $\forall\varphi\in\mathcal{C}$, hence, $\forall x\in\mathbb{R}$

$$4K^*(x)\left(\int\limits_{-\infty}^{+\infty}t^2K^*(t)dt\right)+x^2\left(\int\limits_{-\infty}^{+\infty}(K^*(t))^2dt\right)-5\left(\int\limits_{-\infty}^{+\infty}t^2K^*(t)dt\right)\left(\int\limits_{-\infty}^{+\infty}(K^*(t))^2dt\right)=0.$$

Rewrite the last equation as following

$$K^*(x) = -\frac{\left(\int\limits_{-\infty}^{+\infty}(K^*(t))^2dt\right)}{4\left(\int\limits_{-\infty}^{+\infty}t^2K^*(t)dt\right)}x^2+\frac{5}{4}\left(\int\limits_{-\infty}^{+\infty}(K^*(t))^2dt\right).$$

It gives us the next tip $K^*(x) = -ax^2+c$ where $a\geqslant 0, b\geqslant 0$ are parameters which can be found. But for convergence of the integrals we must modify $K^*(x)$ as following

$$K^*(x) = (-ax^2+c)\mathbb{I}[|x|\leqslant\alpha],\ \forall\alpha>0.$$

Using $\int\limits_{-\infty}^{+\infty} K^*(x)dx = 1$ we solve the next system of equations

$$a = \frac{\left(\int\limits_{-\infty}^{+\infty} (K^*(t))^2 dt\right)}{4\left(\int\limits_{-\infty}^{+\infty} t^2 K^*(t) dt\right)}, \ c = \left(\int\limits_{-\infty}^{+\infty} (K^*(t))^2 dt\right).$$

As a result we have

$$a = \frac{3}{4\alpha^3}, \ c = \frac{3}{4\alpha}.$$

We can choose $\alpha = 1$ and finish the proof.