EPFL | MGT-418 : Convex Optimization | Project 2
Quantile Regression

Charles Vuichard, Erick Maraz, Gloria Dal Santo

November 2020

Exercise 1

Firstly, we assume that the statement that has to be proven is true for all values of $\boldsymbol{w}$, and not only for $\boldsymbol{w}^*$.
The objective function of problem (1) is the following:

$$f_0(b) = \sum_{i=1}^{m} \tau \cdot \max\{0, \boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i\} + (1-\tau) \cdot \max\{0, y_i - \boldsymbol{w}^\top \boldsymbol{x}_i - b\}$$

Since $\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i$, $\forall i = 1, ..., m$ is linear in b, the max functions are convex. Then, $f_0(b)$ is a linear combination of convex function, which implies that $f_0(b)$ is convex in b. Moreover $f_0(b)$ is a piecewise linear function in b.

If $\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i \leq 0\,, \forall i = 1, ..., m$ (i.e. all the samples are above the hyperplane $\mathcal{H} = \boldsymbol{w}^\top \boldsymbol{x} + b$) we have that

$$\sum_{i=1}^{m} L(r_i(\boldsymbol{w}, b)) = \sum_{i=1}^{m} -(1-\tau)(\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i) \tag{1}$$

which means that each term of $f_0(b)$ has a slope of $-(1-\tau)$. The objective function then will have a negative slope of $-m(1-\tau)$. As we increase b there will be some samples that go below $\mathcal{H}$, changing the slope of $f_0(b)$. For instance, if there is one sample $(\boldsymbol{x}_i, y_i)$ such that $\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i \geq 0$ then the slope of $f_0(b)$ will be $-(m-1)(1-\tau) + \tau$ and has a kink when $\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i = 0$. On the other hand, if $\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i \geq 0$, $\forall i = 1, .., m$ we have that

$$\sum_{i=1}^{m} L(r_i(\boldsymbol{w}, b)) = \sum_{i=1}^{m} \tau(\boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i) \tag{2}$$

and the objective function will have a positive slope $m\tau$. As we decrease b the slope becomes less positive since the number of samples above the hyperplane increases.
From the observations above and recalling that for convex functions any local minimum is also a global minimum, we can deduct that the optimal value of b will be such that the slope of $f_0(b)$ has a value between $m\tau$ and $-m(1-\tau)$.
Going back to eq.(1), assuming we are increasing b, at some point we will observe a change in the sign of the slope, and we will obtain

$$\tau(m - \mathrm{K}(b)) - (1-\tau)\mathrm{K}(b) \geq 0 \tag{3}$$

where $\mathrm{K}(b)$ is the number of samples above $\mathcal{H}$. In other words, the number of samples that are below or on $\mathcal{H}$ becomes greater or equal to the samples that are above $\mathcal{H}$. This condition is satisfied for $\mathrm{K}(b) \geq m\tau$. Substituting $\mathrm{K}(b)$ in the first term of eq.(3) we have that there must

be least $m(1-\tau)$ samples that lie on or below $\mathcal{H}$ for the inequality to hold.
With the same reasoning, if we consider that at the beginning we are in condition described by eq.(2) and we decrease b, at some point we will have

$$\tau \mathrm{W}(b) - (1-\tau)(m - \mathrm{W}(b)) \leq 0 \tag{4}$$

where $\mathrm{W}(b)$ represents the number of samples below the hyperplane $\mathcal{H}$ and it satisfies eq.(4) when $\mathrm{W}(b) \leq m(1-\tau)$. Substituting $\mathrm{W}(b)$ in the second term of eq.(4) we have that there must be least $m\tau$ samples that lie on or above $\mathcal{H}$ for the inequality to hold.
If the constrains on $\mathrm{W}(b)$ and $\mathrm{K}(b)$ are satisfied then the optimal value of b would be referred to the condition in which the two slopes cancel out, that is when $\mathrm{K}(b) = m\tau$ and $\mathrm{W}(b) = (1-m)\tau$.
The optimal value b^* may be not unique, indeed it is usually the case that as soon as $\mathrm{K}(b) = m\tau$ (or $\mathrm{W}(b) = (1-m)\tau$, if we arrive from a positive slope) is satisfied, the next samples that cross the hyperplane is at some distance from the previous one.
We can put a further constrain on the results that we have obtained applying the ceil function: since the number of samples is an integer number we have that for any b^* at least $\lceil \tau m \rceil$ training samples lie on or above $\mathcal{H}$ and at least $\lceil m(1-\tau \rceil$ training samples lie on or below $\mathcal{H}$
The considerations just made do not depend on the choice of $\boldsymbol{w}$, indeed it can be chosen a non optimal value and the statements would still be verified.

Exercise 2

The τ-quantile is defined as:

$$\inf_{\boldsymbol{\eta}} \{\eta / |\{i = 1, ..., m | r_i(\boldsymbol{w}^*, b^*) \leq \eta\}| \geq \tau m\} \tag{5}$$

So, η is defined as :

$$\eta / |\{i = 1, ..., m | r_i(\boldsymbol{w}^*, b^*) \leq \eta\}| \geq \tau m \tag{6}$$

From question 1, we now that :

- At least $\lceil \tau m \rceil$ training samples lie on or above the hyperplane $\mathcal{H}$
- At least $\lceil \tau(1-m) \rceil$ training samples lie on or below $\mathcal{H}$

$r_i(\boldsymbol{w}^*, b^*) = \eta = 0$ means that the sample lie on $\mathcal{H}$.
Therefore, $r_i(\boldsymbol{w}^*, b^*) \leq 0$ refers to training samples lying on or above the hyperplane $\mathcal{H}$.
From this results, together with eq.(6), we can write:

$$\begin{array}{rcll} |\{i = 1, ..., m | r_i(\boldsymbol{w}^*, b^*) \leq 0\}| & \geq & \lceil \tau m \rceil & \geq \tau m \\ |\{i = 1, ..., m | - r_i(\boldsymbol{w}^*, b^*) \leq 0\}| & \geq & \lceil \tau(1-m) \rceil & \geq \tau(1-m) \end{array}$$

Hence, 0 is a feasible solution for eq.(5). What if it is also the optimal one? Since b^* is the largest among all the optimal values of b, then there is a training sample lying on $\mathcal{H}$.
If we try to decrease the solution $\eta = 0$, then $|\{i = 1, ..., m | r_i(\boldsymbol{w}^*, b^*) \leq 0\}| = \lceil \tau m \rceil - 1$. $\lceil \tau m \rceil - 1 \leq \tau m$, leads to a unfeasible solution.
We can conclude that 0 is the τ-quantile of the uniform distribution.

Exercise 3

The pinball loss function is defined as:

$$L(z) = \tau[z]_+ + (1-\tau)[z]_- \tag{7}$$

where $[z]_+ = \max\{0, z\}$ and $[z]_- = \max\{0, -z\}$.
To express $L(z)$ as a function of $|z|$ and z, first we note that:

$$\begin{aligned} |z| &= [z]_+ + [z]_- \\ z &= [z]_+ - [z]_- \end{aligned}$$

From which we can derive the following equations by solving for $[z]_+$ and $[z]_-$

$$[z]_+ = \frac{|z| + z}{2} \tag{8}$$

$$[z]_- = \frac{|z| - z}{2} \tag{9}$$

Finally, replacing (8) and (9) into eq.(7) results in:

$$L(z) = \frac{1}{2}|z| + \left(\tau - \frac{1}{2}\right) z \tag{10}$$

To reformulate the problem

$$\underset{\boldsymbol{w}\in\mathbb{R}^n,\, b\in\mathbb{R}}{\text{minimize}} \sum_{i=1}^{m} L(r_i(\boldsymbol{w}, b))$$

where $r_i(\boldsymbol{w}, b) = \boldsymbol{w}^T\boldsymbol{x}_i + b - y_i$, as a linear program, we define the epigraphical variable t_i to get rid of the absolute value term in (10):

$$|\boldsymbol{w}^T\boldsymbol{x}_i + b - y_i| \quad \Longrightarrow \quad -t_i \leq \boldsymbol{w}^T\boldsymbol{x}_i + b - y_i \leq t_i \quad \forall i = 1, ..., m$$

The reformulated program is:

$$\begin{aligned} \underset{\boldsymbol{w}\in\mathbb{R}^n,\, b\in\mathbb{R},\, \boldsymbol{t}\in\mathbb{R}^m_+}{\text{minimize}} \quad & \sum_{i=1}^{m} \frac{1}{2}t_i + \left(\tau - \frac{1}{2}\right)\left(\boldsymbol{w}^T\boldsymbol{x}_i + b - y_i\right) \\ \text{subject to} \quad & \boldsymbol{w}^T\boldsymbol{x}_i + b - y_i \leq t_i \qquad \forall i = 1, ..., m \\ & -\boldsymbol{w}^T\boldsymbol{x}_i - b + y_i \leq t_i \qquad \forall i = 1, ..., m \end{aligned}$$

Exercise 4

Fig.(1) illustrates the results obtained with the Matlab script. In the dataset there are two outliers at coordinates $(-9.5, 15)$ and $(9, -5)$. It can be noticed that the squared loss is more sensitive to these outliers than the quantile regression since it is more tilted towards them. This makes the quantile regression a more accurate classifier for similar type of datasets.
The effect of τ is that of penalizing more the samples that lie above or below the hyperplane $\mathcal{H} = \omega^\top x + b$. For example, for values of τ less than 0.5 the pinball loss penalizes more the samples that are above $\mathcal{H}$. In this case, to minimize $L(r_i(\omega, b))$, ω and b must be chosen such that $\mathcal{H}$ is closer to the samples that lie on the top of the graph. The case for $\tau = 0.5$ gives a fair penalization and the resulting hyperplane splits the data set into two planes equally populated.

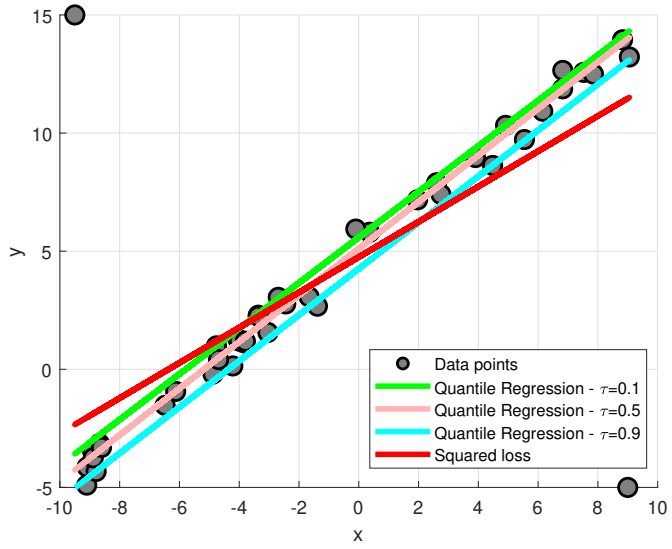

Figure 1: Quantile and Square Regressions

Exercise 5

In Fig.(2), the cumulative distribution function of the uniform distribution over the prediction errors $r_i(\boldsymbol{w}, b)$ for $\tau = 0.1, 0.5, 0.9$ is plotted. For every plot, it can be seen that 0 is the τ-quantile of the uniform distribution (points: (0,0.1) (0,0.5) (0,0.9)).

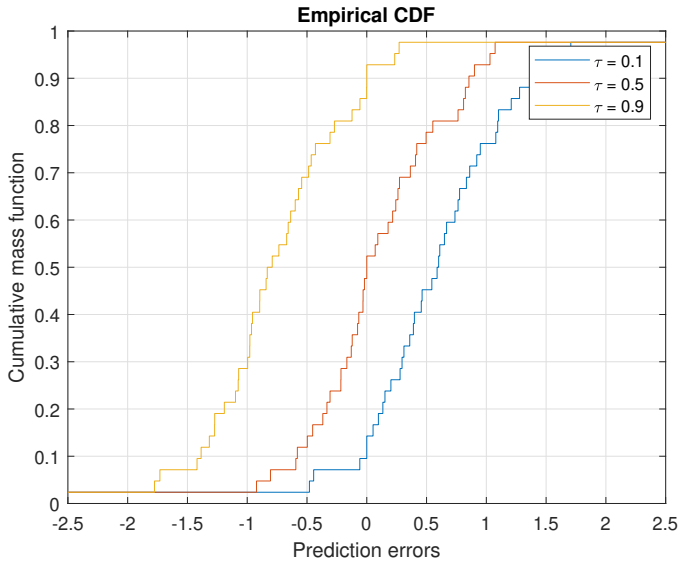

Figure 2: Empirical quantile estimation

Exercise 6

Question (a)

We have the problem:

$$\begin{aligned} \underset{\boldsymbol{w}\in\mathbb{R}^n,\, b\in\mathbb{R},\, \boldsymbol{t}\in\mathbb{R}^m_+}{\text{minimize}} \quad & \sum_{i=1}^{m}\left(\frac{1}{2}t_i + \left(\tau - \frac{1}{2}\right)\left(\boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b - y_i\right)\right) + \frac{\rho}{2}\|\boldsymbol{w}\|_2^2 \\ \text{subject to} \quad & \boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b - y_i \leq t_i \qquad \forall i = 1, ..., m \\ & -\boldsymbol{w}^T\phi(\boldsymbol{x}_i) - b + y_i \leq t_i \qquad \forall i = 1, ..., m \end{aligned} \tag{11}$$

The Lagrangian for the problem above is:

$$\begin{aligned} \mathcal{L}(\boldsymbol{w}, b, \boldsymbol{t}, \boldsymbol{\lambda}^+, \boldsymbol{\lambda}^-) = & \sum_{i=1}^{m}\frac{1}{2}t_i + \left(\tau - \frac{1}{2}\right)\left(\boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b - y_i\right) \\ & + \sum_{i=1}^{m}\lambda_i^+(\boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b - y_i - t_i) \\ & + \sum_{i=1}^{m}\lambda_i^-(-\boldsymbol{w}^T\phi(\boldsymbol{x}_i) - b + y_i - t_i) \\ & + \frac{\rho}{2}\|\boldsymbol{w}\|_2^2 \end{aligned} \tag{12}$$

Question (b)

Reformulating eq.(12) we can obtain the following equation:

$$\begin{aligned} \mathcal{L}(\boldsymbol{w}, b, \boldsymbol{t}, \boldsymbol{\lambda}^+, \boldsymbol{\lambda}^-) = & \sum_{i=1}^{m}\left(\frac{1}{2} - \lambda_i^+ - \lambda_i^-\right)t_i \\ & + \sum_{i=1}^{m}\left(\tau - \frac{1}{2} + \lambda_i^+ - \lambda_i^-\right)\boldsymbol{w}^T\phi(\boldsymbol{x}_i)) \\ & - \sum_{i=1}^{m}\left(\tau - \frac{1}{2} + \lambda_i^+ - \lambda_i^-\right)y_i \\ & + \sum_{i=1}^{m}\left(\tau - \frac{1}{2} + \lambda_i^+ - \lambda_i^-\right)b \\ & + \frac{\rho}{2}\|\boldsymbol{w}\|_2^2 \end{aligned}$$

To evaluate the dual objective, we note that

- the infimum over $\boldsymbol{w}$ satifies

$$\nabla_{\boldsymbol{w}}\left(\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{t}, \boldsymbol{\lambda}^+, \boldsymbol{\lambda}^-)\right) = 0 \quad \Longrightarrow \quad \boldsymbol{w}^* = -\frac{1}{\rho}\sum_{i=1}^{m}\left(\tau - \frac{1}{2} + \lambda_i^+ - \lambda_i^-\right)\phi(\boldsymbol{x}_i) \tag{13}$$

- the infimum over b is finite iff

$$\sum_{i=1}^{m}(\lambda_i^+ - \lambda_i^-) = m\left(\frac{1}{2} - \tau\right) \tag{14}$$

- the infimum over t_i is finite iff

$$\frac{1}{2} = \lambda_i^+ + \lambda_i^- \quad \Longrightarrow \quad \frac{1}{2} \geq \lambda_i^+ + \lambda_i^- \quad \forall i = 1, ..., m \tag{15}$$

  Since $\lambda_i^- \geq 0$ and $\lambda_i^+ \geq 0$, the condition in the t_i term: $\frac{1}{2} \geq -\lambda_i^+ - \lambda_i^- \ \ \forall i = 1, ..., m$ is always satisfied.

From (13), we write :

$$\begin{aligned}
&\sum_{i=1}^{m} \left(\tau - \frac{1}{2} + \lambda_i^+ - \lambda_i^- \right) \phi(\boldsymbol{x}_i)^T \boldsymbol{w} + \frac{\rho}{2} \|\boldsymbol{w}\|_2^2 \\
&= \sum_{i=1}^{m} \left(\tau - \frac{1}{2} + \lambda_i^+ - \lambda_i^- \right) \phi(\boldsymbol{x}_i)^T \left(-\sum_{j=1}^{m} \left(\tau - \frac{1}{2} + \lambda_j^+ - \lambda_j^- \right) \frac{\phi(\boldsymbol{x}_j)}{\rho} \right) \\
&+ \frac{\rho}{2} \left(-\sum_{i=1}^{m} \left(\tau - \frac{1}{2} + \lambda_i^+ - \lambda_i^- \right) \frac{\phi(\boldsymbol{x}_i)}{\rho} \right)^T \left(-\sum_{j=1}^{m} \left(\tau - \frac{1}{2} + \lambda_j^+ - \lambda_j^- \right) \frac{\phi(\boldsymbol{x}_j)}{\rho} \right) \\
&= \frac{\left(-1 + \frac{1}{2}\right)}{\rho} \left(\sum_{i=1}^{m} \left(\tau - \frac{1}{2} + \lambda_i^+ - \lambda_i^- \right) \phi(\boldsymbol{x}_i) \right)^T \left(\sum_{j=1}^{m} \left(\tau - \frac{1}{2} + \lambda_j^+ - \lambda_j^- \right) \phi(\boldsymbol{x}_j) \right) \\
&= -\frac{1}{2\rho} \sum_{i=1}^{m} \left(\tau - \frac{1}{2} + \lambda_i^+ - \lambda_i^- \right) \left(\sum_{j=1}^{m} \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j) \left(\tau - \frac{1}{2} + \lambda_j^+ - \lambda_j^- \right) \right)
\end{aligned}$$

Using the result above, (14), (15) and the Lagrangian, we define the dual objective :

$$g(\boldsymbol{\lambda}^+, \boldsymbol{\lambda}^-) = -\sum_{i=1}^{m} \left(\tau - \frac{1}{2} + \lambda_i^+ - \lambda_i^- \right) \left(y_i + \frac{1}{2\rho} \sum_{j=1}^{m} K_{ij} \left(\tau - \frac{1}{2} + \lambda_j^+ - \lambda_j^- \right) \right)$$

And the dual problem of (11):

$$\begin{aligned}
\underset{\lambda^+ \in \mathbb{R}_+^n,\, \lambda^- \in \mathbb{R}_+^n}{\text{maximize}} \quad & -\sum_{i=1}^{m} \left(\tau - \frac{1}{2} + \lambda_i^+ - \lambda_i^- \right) \left(y_i + \frac{1}{2\rho} \sum_{j=1}^{m} K_{ij} \left(\tau - \frac{1}{2} + \lambda_j^+ - \lambda_j^- \right) \right) \\
\text{subject to} \quad & \sum_{i=1}^{m} (\lambda_i^+ - \lambda_i^-) = m \left(\frac{1}{2} - \tau \right) \\
& \lambda_i^+ + \lambda_i^- \leq \frac{1}{2} \quad \forall i = 1, ..., m
\end{aligned}$$

Question (c)

The problem in eq.(11) is convex as its objective function is convex (a sum of a squared norm and a sum of affine functions) and its feasible set is a polyhedron.
Because the given problem is convex and satisfies Slater's condition (*e.g.*, $w = 0, b = 0, t_i = 2|y_i| \ \forall i = 1, ..., m$ is strictly feasible), the KKT conditions are both sufficient and necessary for optimality.

KKT conditions :

$$
\begin{aligned}
&\text{(PF):} && \boldsymbol{w}^{*T}\phi(\boldsymbol{x}_i) + b^* - y_i \leq t_i^* && \forall i = 1, ..., m \\
& && -\boldsymbol{w}^{*T}\phi(\boldsymbol{x}_i) - b^* + y_i \leq t_i^* && \forall i = 1, ..., m \\
&\text{(DF):} && (\lambda_i^-)^* \geq 0 && \forall i = 1, ..., m \\
& && (\lambda_i^+)^* \geq 0 && \forall i = 1, ..., m \\
&\text{(CS):} && (\lambda_i^-)^*(\boldsymbol{w}^{*T}\phi(\boldsymbol{x}_i) + b^* - y_i - t_i^*) = 0 && \forall i = 1, ..., m \\
& && (\lambda_i^+)^*(-\boldsymbol{w}^{*T}\phi(\boldsymbol{x}_i) - b^* + y_i - t_i^*) = 0 && \forall i = 1, ..., m
\end{aligned}
$$

(ST): (13), (14), (15)

Note that the optimal primal and dual decision variables must satisfy the KKT conditions. Therefore, (ST) (13) :

$$\boldsymbol{w}^* = \frac{1}{\rho}\sum_{i=1}^{m}\left(\frac{1}{2} - \tau - (\lambda_i^+)^* + (\lambda_i^-)^*\right)\phi(\boldsymbol{x}_i)$$

Let $k = \{i \,|\, i \in \{1, ..., m\}$ and $1/2 > (\lambda_i^+)^* + (\lambda_i^-)^*\}$. From (CS) and (13):

$$\boldsymbol{w}^{*T}\phi(\boldsymbol{x}_k) + b^* - y_k - t_k^* = 0 \quad \forall k \tag{16}$$

$$-\boldsymbol{w}^{*T}\phi(\boldsymbol{x}_k) - b^* + y_k - t_k^* = 0 \quad \forall k \tag{17}$$

Subtracting equation (17) from (16):

$$2\,\phi(\boldsymbol{x}_k)^T\boldsymbol{w}^* + 2b^* - 2y_k = 0 \quad \forall k$$

And solving for b^*:

$$b^* = y_k + \phi(\boldsymbol{x}_k)^T\frac{1}{\rho}\sum_{i=1}^{m}(\frac{1}{2} - \tau - \lambda_i^{+*} + \lambda_i^{-*})\phi(\boldsymbol{x}_i) \quad \forall k$$

$$b^* = y_k + \frac{1}{\rho}\sum_{i=1}^{m}K_{ki}\left(\frac{1}{2} - \tau - \lambda_i^{+*} + \lambda_i^{-*}\right) \quad \forall k$$

Question (d)

From the Fig.(3), we can easily verify that the kernelized quantile regression is more accurate for the given set of data than the original quantile regression without kernelization. Indeed, the one without kernelization, tries to fit a non linearly distributed datapoints with a linear function. With the Gaussian kernel we can fit the data with a boundary of higher degree and thus reducing the error that the original quantile produces.

Exercise 7

Figure 4 compares the actual electricity prices with the ones predicted through quantile regression for $\tau = 0.1, 0.5, 0.9$. The data points that have been used to estimate each electricity price at a specific hour are the prices of the 10 hours before it (in the given dataset these correspond to the last ten elements of each row of $\mathbf{x}$). It can be noticed that the prices predicted with $\tau = 0.1$ result in a overestimation of the actual prices. On the other hand, for $\tau = 0.9$ the predicted prices result to be lower that the actual ones. A trade off between these two results is given by $\tau = 0.5$, for which we have a prediction that lies around the median.
If we interpret τ as a percentage, $100 \cdot \tau$ % of the actual plot is above the predicted τ-plots. And $100 \cdot (1 - \tau)$ % of the actual plot is below the predicted τ-plots. Moreover, when the actual plot crosses a predicted τ-plot, this means that the prediction is exact for that point $(r_i(w, b) = 0)$

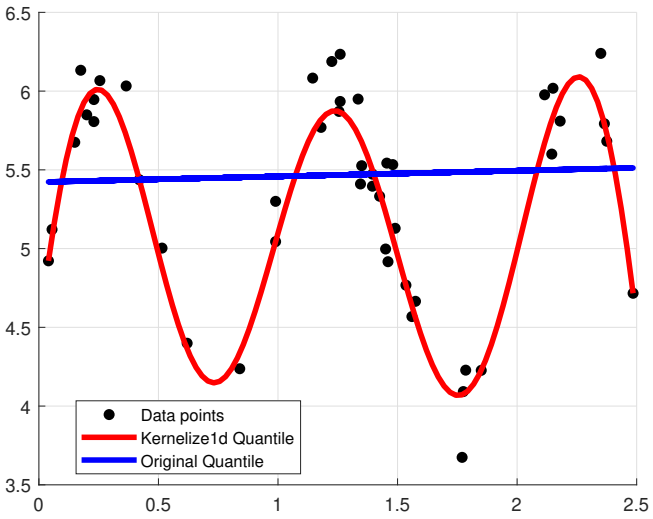

Figure 3: Gaussian kernel Regression

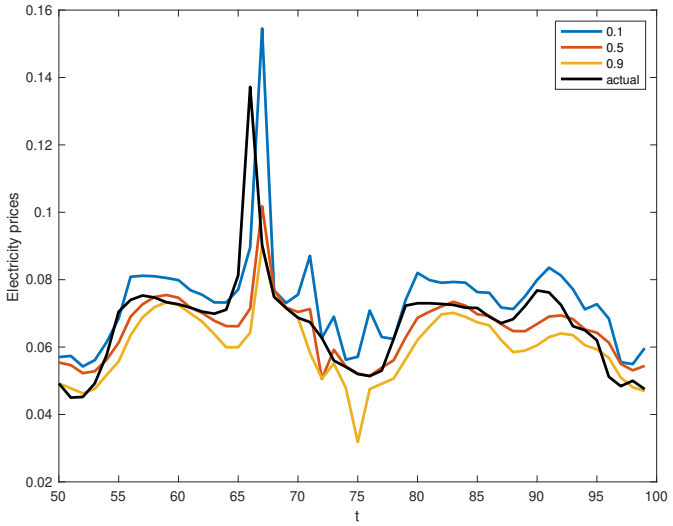

Figure 4: Electricity Price Prediction