Local averaging methods

Vincent Divol

In this chapter, we investigate a class of predictors, called local averaging methods. Those methods are defined by computing a weighted average of the different outputs $\mathbf{y_i}$ from a sample of n observations $(\mathbf{x_1}, \mathbf{y_1}), \dots, (\mathbf{x_n}, \mathbf{y_n})$. As such, those methods are simple to compute and to interpret. However, they are best suited to low-dimensional setting as they suffer from the curse of dimensionality.

1 The regression problem

Let $(\mathbf{x_1}, \mathbf{y_1}), \dots, (\mathbf{x_n}, \mathbf{y_n})$ be a training sample with distribution P. We focus here on regression on the cube $[0,1]^d$: the set of inputs is $\mathcal{X} = [0,1]^d$, the set of outputs is $\mathcal{Y} = \mathbb{R}$, and we use the squared loss $\ell(y, y') = (y - y')^2$. Before studying local averaging methods, let us recall some basic facts on regression. We proved several weeks ago that the Bayes predictor for the squared loss is given by $f_P^\star(x) = \mathbb{E}_P[\mathbf{y}|\mathbf{x} = x]$, the conditional expectation of $\mathbf{y}$ given that $\mathbf{x} = x$. We can always write $\mathbf{y}$ as

$$\mathbf{y} = f_P^\star(\mathbf{x}) + \mathbf{e} \tag{1}$$

where $\mathbf{e}$ is defined as $\mathbf{e} = \mathbf{y} - f_P^\star(\mathbf{x})$. By construction, $\mathbb{E}[\mathbf{e}|\mathbf{x}] = 0$. We may therefore think of $\mathbf{y}$ as being obtained by corrupting $f_P^\star(\mathbf{x})$ by some random centered noise $\mathbf{e}$. Note however that the distribution of the noise $\mathbf{e}$ may depend on $\mathbf{x}$.

Example 1.1. Each input $\mathbf{x}$ represents a street in a city (the city being represented by a square $[0,1]^2$), and $\mathbf{y}$ represents the CO2 concentration at $\mathbf{x}$. The output $\mathbf{y}$ will vary depending on when the CO2 concentration is measured. In this setting, $f_P^\star(x)$ represents the average CO2 concentration at the street x. The distribution of the noise $\mathbf{e}$ may vary depending on $\mathbf{x}$: for example,

some streets x in the city may have higher variations of CO2 concentration than others, so that $\mathbb{E}[\mathbf{e}^2|\mathbf{x} = x]$ will be larger for those streets.

The Bayes risk $\mathcal{R}_P^\star$ is equal to

$$\mathcal{R}_P^\star = \mathbb{E}_P[(f_P^\star(\mathbf{x}) - \mathbf{y})^2] = \mathbb{E}_P[\mathbf{e}^2]. \tag{2}$$

Fix a function $f : \mathcal{X} \to \mathbb{R}$. Let us compute $\mathcal{R}_P(f) = \mathbb{E}_P[(f(\mathbf{x}) - \mathbf{y})^2]$. To do so, we first compute $\mathbb{E}_P[(f(\mathbf{x}) - \mathbf{y})^2|\mathbf{x}]$:

$$\begin{aligned}
\mathbb{E}_P[(f(\mathbf{x})-\mathbf{y})^2|\mathbf{x}] &= \mathbb{E}_P[(f(\mathbf{x}) - f_P^\star(\mathbf{x}) - \mathbf{e})^2|\mathbf{x}] \\
&= \mathbb{E}_P[(f(\mathbf{x}) - f_P^\star(\mathbf{x}))^2|\mathbf{x}] + 2\mathbb{E}[(f(\mathbf{x}) - f_P^\star(\mathbf{x}))\mathbf{e}|\mathbf{x}] + \mathbb{E}[\mathbf{e}^2|\mathbf{x}] \\
&= (f(\mathbf{x}) - f_P^\star(\mathbf{x}))^2 + 2(f(\mathbf{x}) - f_P^\star(\mathbf{x}))\mathbb{E}[\mathbf{e}|\mathbf{x}] + \mathbb{E}[\mathbf{e}^2|\mathbf{x}] \\
&= (f(\mathbf{x}) - f_P^\star(\mathbf{x}))^2 + \mathbb{E}[\mathbf{e}^2|\mathbf{x}],
\end{aligned}$$

where we use that $\mathbb{E}[\mathbf{e}|\mathbf{x}] = 0$. By the law of total expectation,

$$\begin{aligned}
\mathcal{R}_P(f) &= \mathbb{E}_P[\mathbb{E}_P[(f(\mathbf{x}) - \mathbf{y})^2|\mathbf{x}]] \\
&= \mathbb{E}[(f(\mathbf{x}) - f_P^\star(\mathbf{x}))^2] + \mathbb{E}[\mathbb{E}[\mathbf{e}^2|\mathbf{x}]] \\
&= \mathbb{E}[(f(\mathbf{x}) - f_P^\star(\mathbf{x}))^2] + \mathcal{R}_P^\star.
\end{aligned}$$

Therefore, the excess of risk of f is equal to

$$\mathcal{R}_P(f) - \mathcal{R}_P^\star = \mathbb{E}_P[(f(\mathbf{x}) - f_P^\star(\mathbf{x}))^2] = \int_{[0,1]^d} (f(x) - f_P^\star(x))^2 \mathrm{d}P_{\mathbf{x}}(x). \tag{3}$$

Two information are relevant to understand this model: properties of the noise $\mathbf{e}$ and regularity of the Bayes predictor $f_P^\star$. If $f_P^\star$ is a smooth function (for example Lipschitz continuous) and the noise $\mathbf{e}$ is small, then we expect $f_P^\star(x)$ to be similar to $\mathbf{y_i}$ for $\mathbf{x_i}$ close to x. This yields to the following heuritic.

Heuristic. *Given an input x, the predictor $\hat{f}(x)$ should be similar to the outputs $\mathbf{y_i}$ for $\mathbf{x_i}$ close to x.*

We introduce a large class of simple predictors that satisfy this heuristic. Let $w_1(x), \dots, w_n(x)$ be weights with $\sum_{i=1}^n w_i(x) = 1$ and define

$$\hat{f}_w(x) = \sum_{i=1}^n w_i(x)\mathbf{y_i}. \tag{4}$$

The weights $w_i(x)$ depend on the inputs $\mathbf{x_1}, \dots, \mathbf{x_n}$. According to the heuristic, the weights $w_i(x)$ should be high if x is close to $\mathbf{x_i}$, and low otherwise.

Let us write $\mathbf{e_i} = \mathbf{y_i} - f_p^\star(\mathbf{x_i})$. We make the following assumptions on the model.

(A1) the Bayes predictor $f_P^\star : [0,1]^d \to \mathbb{R}$ is α-Lipschitz continuous, that is, for all $x, x' \in [0,1]^d$,

$$|f_P^\star(x) - f_P^\star(x')| \leq \alpha \|x - x'\|. \tag{5}$$

(A2) the Bayes predictor $f_P^\star$ is bounded by $\beta > 0$: for all $x \in [0,1]^d$, $|f_P^\star(x)| \leq \beta$.

(A3) the error $\mathbf{e}$ is bounded: $|\mathbf{e}| \leq \sigma$ for some $\sigma > 0$.

Under this set of assumptions, we can obtain a general decomposition result. Let $x \in [0,1]^d$. We have

$$\begin{aligned}
|\hat{f}_w(x) - f_P^\star(x)| &= \left| \sum_{i=1}^n w_i(x)(f_P^\star(\mathbf{x_i}) + \mathbf{e_i}) - f_P^\star(x) \right| \\
&\leq \left| \sum_{i=1}^n w_i(x)(f_P^\star(\mathbf{x_i}) - f_P^\star(x)) \right| + \left| \sum_{i=1}^n w_i(x)\mathbf{e_i} \right| \\
&\leq \alpha \sum_{i=1}^n |w_i(x)| \|\mathbf{x_i} - x\| + \left| \sum_{i=1}^n w_i(x)\mathbf{e_i} \right|. \qquad (6)
\end{aligned}$$

We refer to the first term in this decomposition as the **approximation error** $\mathrm{App}(x)$: it measures how the local average estimator is able to approximate the Bayes predictor at the point x. The second term measures the inherent noise present in the model, and we call it the fluctuation error at x, denoted by $\mathrm{Fluc}(x)$. Using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we obtain

$$(\hat{f}_w(x) - f_P^\star(x))^2 \leq 2\mathrm{App}(x)^2 + 2\mathrm{Fluc}(x)^2. \tag{7}$$

Let us see how this general decomposition can be used to bound the excess of risk for different weighting schemes.

2 Partition estimators

A partition of a set $\mathcal{X}$ is a collection $\mathcal{A} = (A_j)_{j=1,\dots,J}$ of subsets of $\mathcal{X}$ that are pairwise disjoint (that is $A_j \cap A_{j'} = \emptyset$ if $j \neq j'$) and such that $\bigcup_{j=1}^J A_j = \mathcal{X}$.

Definition 2.1 (Partition estimator)**.** *Consider* $(\mathbf{x_1}, \mathbf{y_1}), \ldots, (\mathbf{x_n}, \mathbf{y_n})$ *a training sample of size* n *from a distribution* P*, with inputs* $\mathbf{x_i} \in [0,1]^d$ *and outputs* $\mathbf{y_i} \in \mathbb{R}$*. Let* $\mathcal{A}$ *be a partition of* $[0,1]^d$*. For* $x \in \mathcal{X}$*, we let* $A(x)$ *be the the element* A_j *of the partition such that* $x \in A_j$*. We define the weights* $w_i : [0,1]^d \to \mathbb{R}$ *associated with the partition* $\mathcal{A}$ *by*

$$w_i(x) := \frac{\mathbf{1}\{\mathbf{x_i} \in A(x)\}}{\sum_{i'=1}^n \mathbf{1}\{\mathbf{x_{i'}} \in A(x)\}}. \tag{8}$$

If $\sum_{i'=1}^n \mathbf{1}\{\mathbf{x_{i'}} \in A(x)\} = 0$*, then, by convention, we let* $w_i(x) = 0$*. The partition estimator* $\hat{f}_{\mathcal{A}}$ *associated with the partition* $\mathcal{A}$ *is the local average estimator with weights* w_i*. The predictor* $\hat{f}_{\mathcal{A}}$ *is also called a regressogram.*

The predictor $\hat{f}_{\mathcal{A}}$ has a very simple structure. For $j = 1, \ldots, J$, let I_j be the set of indexes i such that $\mathbf{x_i} \in A_j$, and let $\mathbf{n_j}$ be the size of I_j. If $\mathbf{n_j} = 0$, then $\hat{f}_w(x) = 0$ for $x \in A_j$. Otherwise, if $\mathbf{n_j} > 0$ and $x \in A_j$, the predictor $\hat{f}_{\mathcal{A}}(x)$ is equal to

$$\hat{f}_{\mathcal{A}}x) = \sum_{i=1}^n w_i(x)\mathbf{y_i} = \frac{\sum_{i=1}^n \mathbf{1}\{\mathbf{x_i} \in A_j\}\mathbf{y_i}}{\sum_{i'=1}^n \mathbf{1}\{\mathbf{x_{i'}} \in A_j\}} = \frac{1}{\mathbf{n_j}} \sum_{i \in I_j} \mathbf{y_i}.$$

To put it otherwise, the prediction $\hat{f}_w$ is constant on each set A_j, equal to the average of the outputs $\mathbf{y_i}$ such that the corresponding input $\mathbf{x_i}$ belongs to A_j.

Example 2.2. Let $\mathcal{X} = [0,1]^d$ and let $L > 0$ be an integer. For $1 \leq j_1, \ldots, j_d \leq L$, let $\vec{j} = (j_1, \ldots, j_d)$ and

$$A_{\vec{j}} = \left[\frac{j_1 - 1}{L}, \frac{j_1}{L}\right) \times \cdot \left[\frac{j_d - 1}{L}, \frac{j_d}{L}\right). \tag{9}$$

The cubes $A_{\vec{j}}$ for $1 \leq j_1, \ldots, j_d \leq L$ define a partition $\mathcal{A}_L$ of $\mathcal{X}$ into a grid of cubes of side length $1/L$. The predictor $\hat{f}_{\mathcal{A}_L} =: \hat{f}_L$ associated with the cube partition is constant on each of these cubes. For $d = 1$, this is simply a histogram.

The remainder of this section is dedicated to analyzing the cube partition estimator. We denote by $\hat{f}_L$ the partition estimator with partition $\mathcal{A}_L$. Let us first give a short summary of the proof strategy. We know that $\hat{f}_L(x)$ is equal to the average of the outputs $\mathbf{y_i}$ for $\mathbf{x_i}$ being in the same cube as x.

As $\mathbf{y_i} = f_P^\star(\mathbf{x_i}) + \mathbf{e_i}$, and as $\mathbf{x_i}$ is at distance $1/L$ from $\mathbf{x}$, the output $\mathbf{y_i}$ is at distance $\alpha/L + |\mathbf{e_i}|$ from $f_P^\star(x)$ (see Figure 1). When we average the different outputs $\mathbf{y_i}$, the different error terms $\mathbf{e_i}$ will cancel out on average, so that we get an error of order $\alpha/L + \sigma/\sqrt{\mathbf{n_{\tilde{j}}}}$. The conclusion is obtained by controlling $\mathbf{n_{\tilde{j}}}$, which follows a binomial random variable.

Let us now turn to the rigorous mathematical analysis. Fix an index $\vec{j}$, and assume for now that $\mathbf{n_{\tilde{j}}} > 0$. For $x \in A_{\vec{j}}$, it holds that

$$\begin{aligned} |\mathrm{App}(x)| &\leq \alpha^2 \left(\sum_{i=1}^{n} |w_i(x)| \|\mathbf{x_i} - x\| \right)^2 \\ &\leq \alpha^2 \left(\frac{1}{\mathbf{n_{\tilde{j}}}} \sum_{i \in I_{\vec{j}}} \|\mathbf{x_i} - x\| \right)^2 \leq \alpha^2 d L^{-2}, \end{aligned} \tag{10}$$

where the last inequality comes from that $\|x - \mathbf{x_i}\| \leq \sqrt{d}/L$ when x and $\mathbf{x_i}$ belong to the same cube $A_{\vec{j}}$.

The fluctuation term is equal to

$$\mathrm{Fluc}(x) = \frac{1}{\mathbf{n_{\tilde{j}}}} \sum_{i \in I_{\vec{j}}} \mathbf{e_i}. \tag{11}$$

Conditionally on $I_{\vec{j}}$, the random variables $(\mathbf{e_i})_{i \in I_{\vec{j}}}$ are independent and identically distributed. Therefore, the conditional expectation of the fluctuation error with respect to the training sample is equal to

$$\mathbb{E}\left[\mathrm{Fluc}(x)^2 \middle| I_{\vec{j}} \right] = \frac{1}{\mathbf{n_{\tilde{j}}}^2} \sum_{i \in I_{\vec{j}}} |\mathbf{e_i}|^2 \leq \frac{\sigma^2}{\mathbf{n_{\tilde{j}}}}. \tag{12}$$

From (7), we obtain

$$\mathbb{E}[(\hat{f}_L(x) - f_P^\star(x))^2 \mathbf{1}\{\mathbf{n_{\tilde{j}}} > 0\}] \leq 2\frac{\alpha^2 d}{L^2} + 2\mathbb{E}[\mathbf{1}\{\mathbf{n_{\tilde{j}}} > 0\} \frac{\sigma^2}{\mathbf{n_{\tilde{j}}}}]. \tag{13}$$

It remains to control $\mathbb{E}[\mathbf{1}\{\mathbf{n_{\tilde{j}}} > 0\} \mathbf{n_{\tilde{j}}}^{-1}]$. Note that $\mathbf{n_{\tilde{j}}}$ follows a binomial random variable of parameters n and $p_{\vec{j}} := P(\mathbf{x} \in A_{\vec{j}})$. Indeed, $\mathbf{n_{\tilde{j}}}$ is the sum over all observations of independent Bernoulli random variables, equal to 1 if the observation is in $A_{\vec{j}}$, and 0 otherwise.

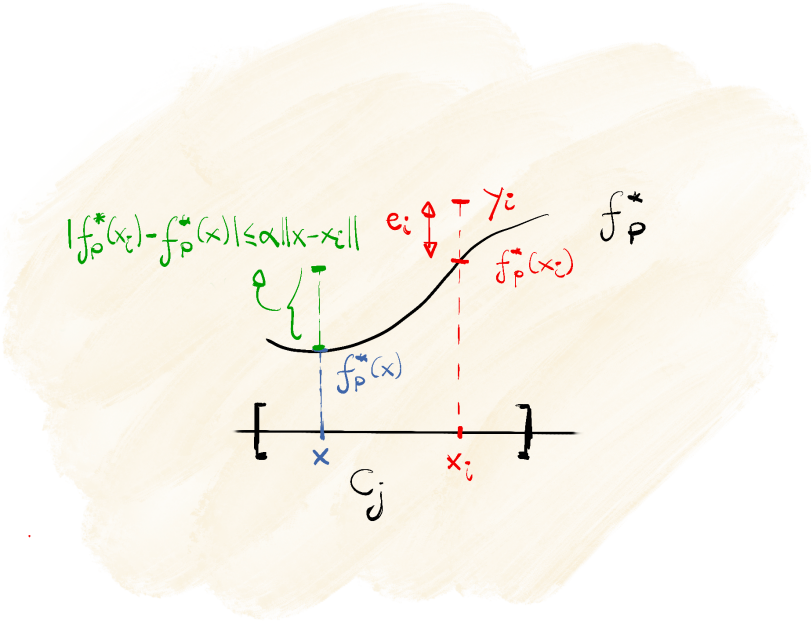

Figure 1: Decomposition of the distance between $\mathbf{y_i}$ and $f_P^\star(x)$ into the stochastic error term $\mathbf{e_i}$ and the distance between $f_P^\star(x)$ and $f_P^\star(\mathbf{x_i})$, which is bounded thanks to the Lipschitz property of $f_P^\star$.

Lemma 2.3. *Let* $\mathbf{N}$ *be a binomial random variable of parameter* n *and* p*. Then,*

$$\mathbb{E}[\mathbf{1}\{\mathbf{N} > 0\}\mathbf{N}^{-1}] \leq \frac{2}{pn}. \tag{14}$$

Proof. We recall the formula $\frac{1}{k+1}\binom{n}{k} = \frac{1}{n}\binom{n+1}{k+1}$. The formula for the density of a binomial random variable implies that

$$\begin{aligned}
\mathbb{E}[\mathbf{1}\{\mathbf{N} > 0\}\mathbf{N}^{-1}] &= \sum_{k=1}^{n} \binom{n}{k} p^k (1-p)^{n-k} \frac{1}{k} \\
&\leq \sum_{k=1}^{n} \binom{n}{k} p^k (1-p)^{n-k} \frac{2}{k+1} \\
&\leq \frac{2}{n+1} \sum_{k=1}^{n} \binom{n+1}{k+1} p^k (1-p)^{n-k} \\
&\leq \frac{2}{n+1} \sum_{l=2}^{n+1} \binom{n+1}{l} p^{l-1} (1-p)^{n-l+1} \\
&\leq \frac{2(1-p)}{p(n+1)} \sum_{l=2}^{n+1} \binom{n+1}{l} p^{l} (1-p)^{n-l} \\
&\leq \frac{2(1-p)}{p(n+1)} \leq \frac{2}{pn}.
\end{aligned}$$

□

Using the lemma and (13) yields

$$\mathbb{E}[(\hat{f}_L(x) - f_P^\star(x))^2 \mathbf{1}\{\mathbf{n}_{\tilde{\mathbf{j}}} > 0\}] \leq 2\frac{\alpha^2 d}{L^2} + \frac{4\sigma^2}{p_{\tilde{j}} n}. \tag{15}$$

When $\mathbf{n}_{\tilde{\mathbf{j}}} = 0$, then $\hat{f}_L(x) = 0$ by convention. In that case, we obtain

$$\begin{aligned}
\mathbb{E}[(\hat{f}_L(x) - f_P^\star(x))^2 \mathbf{1}\{\mathbf{n}_{\tilde{\mathbf{j}}} = 0\}] = f_P^\star(x)^2 \mathbb{P}(\mathbf{n}_{\tilde{\mathbf{j}}} = 0) &= f_P^\star(x)^2 (1 - p_{\tilde{j}})^n \\
&\leq \beta^2 \exp(-n p_{\tilde{j}}),
\end{aligned} \tag{16}$$

where we use Assumption (A2) and the formula for the probability of a binomial random variable being equal to 0. Putting the two estimates together yields

$$\mathbb{E}[(\hat{f}_L(x) - f_P^\star(x))^2] \leq 2\frac{\alpha^2 d}{L^2} + \frac{4\sigma^2}{p_{\tilde{j}} n} + \beta^2 \exp(-n p_{\tilde{j}}). \tag{17}$$

Recall from (3) that the excess of risk of $\hat{f}_L$ is equal to

$$\mathcal{R}_P(\hat{f}_L) - \mathcal{R}_P(f_P^\star) = \int_{[0,1]^d} (\hat{f}_L(x) - f_P^\star(x))^2 \mathrm{d}P_{\mathbf{x}}(x).$$

We obtain the following bound on the expected excess of risk (where expectation represents expectation with respect to the training sample):

$$\begin{aligned}
\mathbb{E}[\mathcal{R}_P(\hat{f}_L) - \mathcal{R}_P(f_P^\star)] &= \int_{[0,1]^d} \mathbb{E}[f_L(x) - f_P^\star(x))^2] \mathrm{d}P_{\mathbf{x}}(x) \\
&= \sum_{\vec{j}} \int_{A_{\vec{j}}} \mathbb{E}[f_L(x) - f_P^\star(x))^2] \mathrm{d}P_{\mathbf{x}}(x) \\
&\leq \sum_{\vec{j}} \int_{A_{\vec{j}}} (2\frac{\alpha^2 d}{L^2} + \frac{4\sigma^2}{p_{\vec{j}} n} + \beta^2 \exp(-n p_{\vec{j}})) \mathrm{d}P_{\mathbf{x}}(x) \\
&\leq \sum_{\vec{j}} p_{\vec{j}} (2\frac{\alpha^2 d}{L^2} + \frac{4\sigma^2}{p_{\vec{j}} n} + \beta^2 \exp(-n p_{\vec{j}})) \\
&\leq 2\frac{\alpha^2 d}{L^2} + \frac{4\sigma^2 L^d}{n} + \beta^2 \sum_{\vec{j}} p_{\vec{j}} \exp(-n p_{\vec{j}})),
\end{aligned}$$

where we use at the last line that there are exactly L^d indexes $\vec{j}$. To conclude, we need to bound the last term in the above equation. One can check that this sum is maximized in the case where all the probabilities $p_{\vec{j}}$ are equal: this sum is therefore smaller than $\exp(-nL^{-d})$.

Theorem 2.4 (Excess of risk of the cube partition estimator)**.** *Assume that conditions (A1)-(A3) hold. Then, the cube partition estimator $\hat{f}_L$ with side length $1/L$ satisfies*

$$\mathbb{E}[\mathcal{R}_P(\hat{f}_L) - \mathcal{R}_P(f_P^\star)] \leq 2\frac{\alpha^2 d}{L^2} + \frac{4\sigma^2 L^d}{n} + \beta^2 \exp(-nL^{-d}). \tag{18}$$

In particular, if $L = cn^{1/(d+2)}$ for some constant c, we obtain a bound of the form

$$\mathbb{E}[\mathcal{R}_P(\hat{f}_L) - \mathcal{R}_P(f_P^\star)] \leq Cn^{-2/(d+2)} \tag{19}$$

for some other constant C.

What should we take away from the above theorem? First, a good news: the partition estimator is consistent, as the excess of risk converges to 0. However, the rate of convergence gets increasingly slow when the number of features d increases. We say that **partition estimators suffer from the curse of dimensionality**. For example, for $d = 18$, the rate of convergence is equal to $n^{-0.1}$, which is only equal to 0.1 even for a number of observations equal to $n = 10^{10}$. This suggests that partition estimators should only be used in low-dimensional settings.

Example 2.5. In this example, we are exploring whether there is a relation between the oil price and the volume of oil sold at a given day at the Brent Complex, a physically and financially traded oil market based around the North Sea of Northwest Europe. The pairs $(\mathbf{x_1}, \mathbf{y_1}), \ldots, (\mathbf{x_1}, \mathbf{y_1})$ represent an oil price ($\mathbf{x}$ value) and a volume sold ($\mathbf{y}$ value). The dataset was downloaded from Kaggle[1]. In this example $d = 1$ and there are $n = 2859$ observations. Theorem 2.4 suggests that we should choose L of order $n^{1/3} \simeq 15$ when designing a partition estimator. This is what is done in Figure 2. We also plot the test error (obtained by randomly splitting the dataset in a training set and a testing set) as a function of L. We see that the minimum of the test error is obtained for L roughly of order 50: the theorem only gives an order of magnitude of what should be a good value of L, and nothing more precise. Moreover, we encounter once again two well-known phenomena: underfitting for L too small, and overfitting for L too large. In practice, L should be selected through cross-validation.

3 Nadaraya-Watson estimators

The partition estimator of the previous section can be summarized in one sentence: the prediction $\hat{f}_L(x)$ is equal to the average of the outputs $\mathbf{y_i}$ corresponding to the inputs $\mathbf{x_i}$ being in the same cube as x. In this section, we investigate a variation on this same idea. We choose as a prediction at the point x the average of the outputs $\mathbf{y_i}$ such that $\mathbf{x_i}$ is at distance less than h from x, where $h > 0$ is a fixed parameter. This is equivalent to defining a local averaging estimator with weights

$$w_i(x) = \frac{\mathbf{1}\{\|x - \mathbf{x_i}\| \leq h\}}{\sum_{i'=1}^{n} \mathbf{1}\{\|x - \mathbf{x_{i'}}\| \leq h\}}.$$

[1] See https://www.kaggle.com/datasets/psycon/historical-brent-oil-price-from-2000-to-202204.

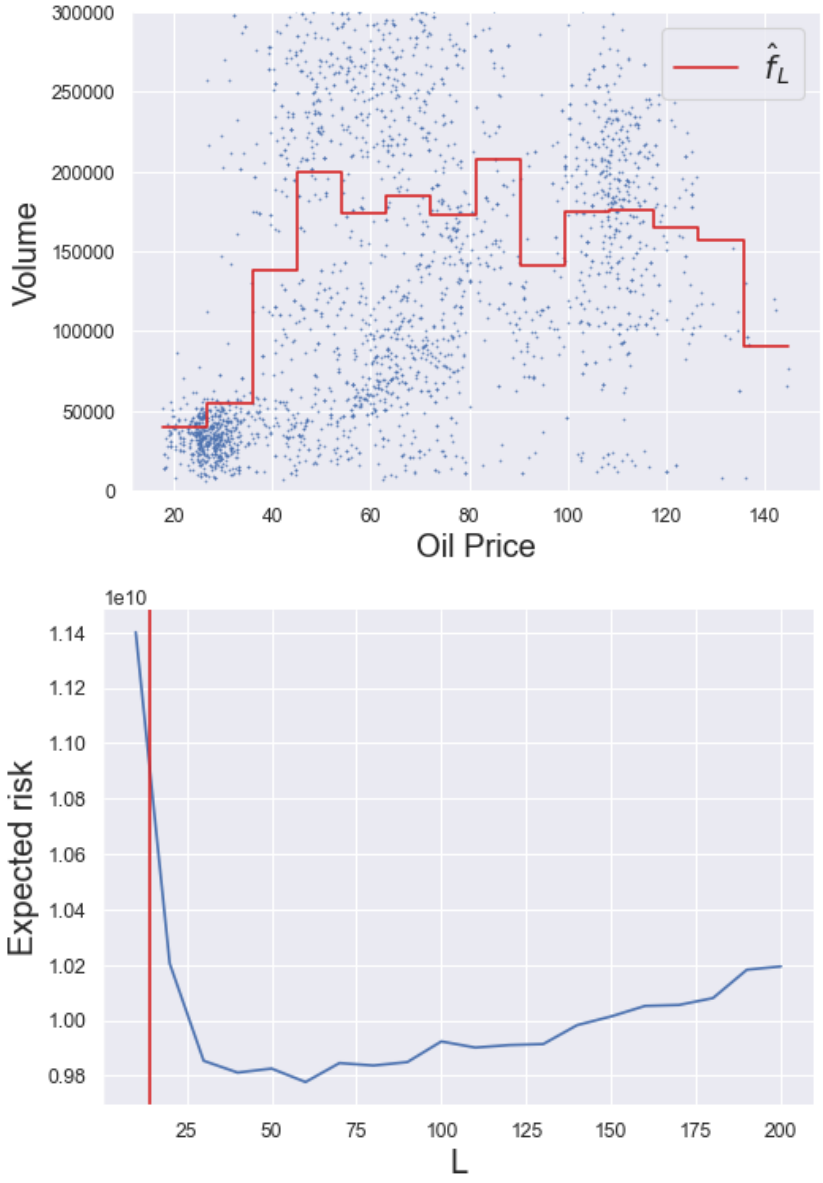

Figure 2: Top: prediction $\hat{f}_{L_0}$ for $L_0 = n^{1/3}$. Bottom: Expected risk for different values of L. The vertical line indicates L_0. The minimum excess of risk is attained for L roughly equal to $3L_0$.

This can be generalized to other weighting schemes.

Definition 3.1. *Consider* $(\mathbf{x_1}, \mathbf{y_1}), \dots, (\mathbf{x_n}, \mathbf{y_n})$ *a training sample of size* n *from a distribution* P*, with inputs* $\mathbf{x_i} \in [0,1]^d$ *and outputs* $\mathbf{y_i} \in \mathbb{R}$*. Let* $K : \mathbb{R}^d \to \mathbb{R}$ *be a function with* $\int K = 1$ *and let* $h > 0$*. Let* K_h *be the function defined by* $K_h(x) = h^{-d}K(x/h)$ *for* $x \in \mathbb{R}^d$*. The Nadaraya-Watson estimator* $\hat{f}_h^{\mathrm{NW}}$ *with kernel* K_h *is defined as the local averaging estimator with weights at* $x \in [0,1]^d$ *equal to*

$$w_i(x) := \frac{K_h(x - \mathbf{x_i})}{\sum_{i'=1}^n K_h(x - \mathbf{x_{i'}})}. \tag{20}$$

The word "kernel" in the above definition is the one that is commonly used by statisticians. Note however that the local averaging method is **not** a kernel method and that the two should not be confused.

The analysis of the Nadaraya-Watson estimator is more complex than the one of the partition estimator, and we refer the interested reader to [**?**, Chapter 1.5]. Let us here only mention that under assumptions similar to assumptions (A1)-(A3), it is possible to show that the Nadaraya-Watson estimator $\hat{f}_h$ satisfies

$$\mathbb{E}[\mathcal{R}_P(\hat{f}_h^{\mathrm{NW}}) - \mathcal{R}_P(f_P^\star)] \leq Cn^{-2/(d+2)}, \tag{21}$$

where h is of order $n^{-1/(d+2)}$ and C is a constant depending on the parameters of the model. Therefore, the Nadaraya-Watson estimator attains the same rate of convergence as the partition estimator and also suffers from the curse of dimensionality. This rate can be improved should the Bayes predictor $f_P^\star$ be k-times differentiable. In this case, one can build a Nadaraya-Watson estimator attaining a rate of convergence of order $n^{-2k/(d+2k)}$.

Example 3.2. A simple choice of kernel is given by the gaussian kernel defined by $K(u) = 1/(2\pi)^{d/2} \exp(-\|u\|^2/2)$ for $u \in \mathbb{R}^d$. We implement the Nadaraya-Watson estimator on the same dataset as in Example 2.5, for the gaussian kernel with different choices of bandwidths h. Once again, the performance of the estimator will crucially depend on h (see Figure 3), a parameter which should be selected thanks to cross-validation to avoid both underfitting and overfitting.

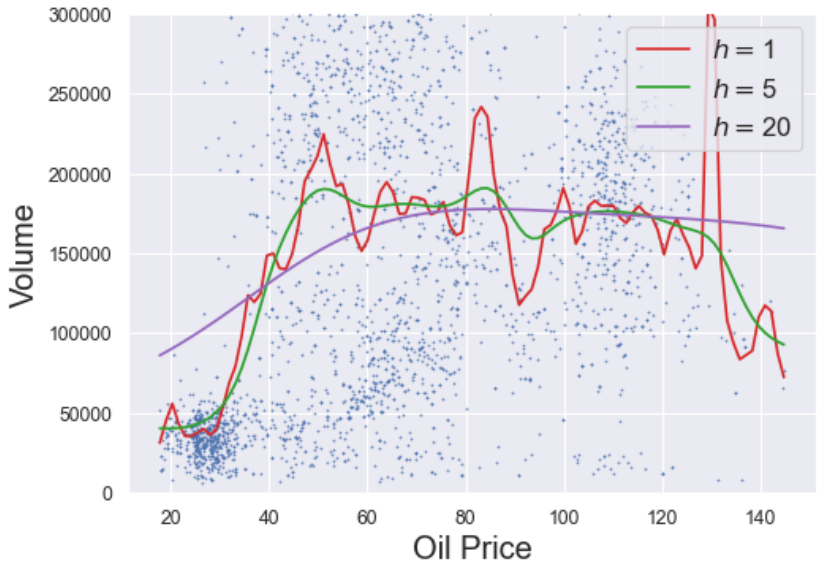

Figure 3: Nadaraya-Watson predictor $\hat{f}_h^{\mathrm{NW}}$ for different values of h on the oil dataset.

4 Nearest-neighbor methods

Here is a very simple idea to make a prediction $\hat{f}(x)$ at $x \in [0,1]^d$: look at the point $\mathbf{x_i}$ the closest to x, and choose $\hat{f}(x) = \mathbf{y_i}$. Such a prediction is called the 1-nearest-neighbor estimator. A variation of this scheme is the k-nearest-neighbor (or k-NN) estimator, which is defined by averaging the outputs $\mathbf{y_i}$ corresponding the k inputs that are the closest from x.

Definition 4.1. *Consider $(\mathbf{x_1}, \mathbf{y_1}), \ldots, (\mathbf{x_n}, \mathbf{y_n})$ a training sample of size n from a distribution P, with inputs $\mathbf{x_i} \in [0,1]^d$ and outputs $\mathbf{y_i} \in \mathbb{R}$. Let $k \geq 1$ be an integer. For $x \in [0,1]^d$, we order the inputs $\mathbf{x_i}$ according to their distance to x:*

$$\|x - \mathbf{x_{i_1}}(x)\| \leq \|x - \mathbf{x_{i_2}}(x)\| \leq \cdots \leq \|x - \mathbf{x_{i_n}}(x)\|. \tag{22}$$

We let $I_k(x) = \{\mathbf{i_1}(x), \ldots, \mathbf{i_k}(x)\}$ and define the weights

$$w_{i,k}(x) = \begin{cases} \frac{1}{k} & \text{if } i \in I_k(x) \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

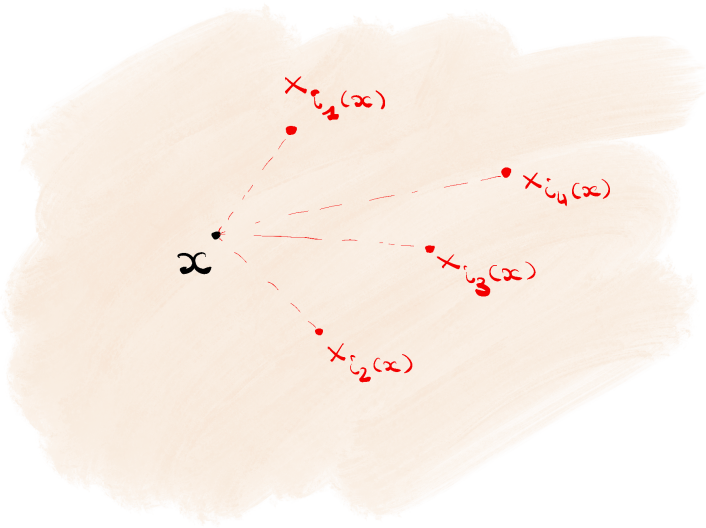

Figure 4: Definition of the indexes $\mathbf{i_1}(x), \ldots, \mathbf{i_4}(x)$.

The k-NN estimator $\hat{f}_k^{\mathrm{NN}}$ is the local averaging estimator associated with the weights $w_{i,k}$.

The k-NN estimator at a point x is equal to

$$\hat{f}_k^{\mathrm{NN}}(x) = \frac{1}{k} \sum_{i \in I_k(x)} \mathbf{y_i}, \tag{24}$$

that is we average the outputs of the k nearest inputs from x. The approximation error is equal to

$$\mathrm{App}(x) := \alpha \sum_{i=1}^{n} |w_i(x)| \|\mathbf{x_i} - x\| = \frac{\alpha}{k} \sum_{i \in I_k(x)} \|\mathbf{x_i} - x\|, \tag{25}$$

that is the average distance between x and its k-nearest neighbors. The fluctuation error is given by

$$\mathrm{Fluc}(x) := \sum_{i=1}^{n} w_i(x) \mathbf{e_i} = \frac{1}{k} \sum_{i \in I_k(x)} \mathbf{e_i}. \tag{26}$$

Conditionally on $I_k(x)$, this is a sum of i.i.d. random variables bounded by σ^2. We thus obtain as in Section 2 that

$$\mathbb{E}[\mathrm{Fluc}(x)^2] \leq \frac{\sigma^2}{k}. \tag{27}$$

The main part of the analysis of the k-NN estimator consists in controlling the distance $\|x - \mathbf{x_{i_k(x)}}\|$ between a point x and its kth nearest neighbor, allowing us to bound the approximation error $\mathrm{App}(x)$. Let us first consider the case $k = 1$. To make our life easier, we will assume that the distribution $P_{\mathbf{x}}$ of the inputs $\mathbf{x_i}$ has a lower bounded density on the cube.

(A4) The distribution $P_{\mathbf{x}}$ has a density p on $[0,1]^d$. Furthermore, there exists a constant $p_{\min} > 0$ such that $p(x) \geq p_{\min}$ for every $x \in [0,1]^d$.

Condition (A4) ensures that the inputs $\mathbf{x_i}$s cover all regions of the cube, and that none is missed out (which would be the case if the density p is zero on that region).

Lemma 4.2. *Assume that condition (A4) holds and let $x \in [0,1]^d$. Let ω_d be the volume of the unit ball in $\mathbb{R}^d$. Then, for every $t \geq 0$,*

$$\mathbb{P}(\|x - \mathbf{x_{i_1(x)}}\| \geq t) \leq \exp(-\omega_d 2^{-d} p_{\min} n t^d). \tag{28}$$

Proof. The condition $\|x - \mathbf{x_{i_1(x)}}\| \geq t$ is satisfied if and only if the ball $B(x,t)$ centered at x of radius t does not intersect $\{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$. The number $\mathbf{N}$ of inputs $\mathbf{x_i}$ that fall in the ball $B(x,t)$ follows a binomial random variable of parameter n and $P(B(x,t))$. Therefore,

$$\mathbb{P}(\|x - \mathbf{x_{i_1(x)}}\| \geq t) = (1 - P(B(x,t))^n \leq \exp(-nP(B(x,t))). \tag{29}$$

The probability $P(B(x,t))$ is lower bounded by

$$\int_{[0,1]^d} \mathbf{1}\{u \in B(x,t)\} p(u) \mathrm{d}u \geq p_{\min} \int_{[0,1]^d} \mathbf{1}\{u \in B(x,t)\} \mathrm{d}u \geq p_{\min} \frac{\omega_d}{2^d} t^d.$$

Indeed, at least a fraction of $1/2^d$ of the ball $B(x,t)$ intersects the cube $[0,1]^d$ (the worst case being attained for x being a corner of the cube). □

Going from a bound on the tail probability to a bound on the second moment is possible thanks to the next lemma.

Lemma 4.3. *Let* $\mathbf{z}$ *be a nonnegative random variable. Then*

$$\mathbb{E}[\mathbf{z}^2] = 2\int_0^{+\infty} u\mathbb{P}(\mathbf{z} \geq u)\mathrm{d}u. \tag{30}$$

Proof. We have

$$\mathbb{E}[\mathbf{z}^2] = \mathbb{E}[\int_0^{+\infty} \mathbf{1}\{\mathbf{z}^2 \geq t\}\mathrm{d}t] = \int_0^{+\infty} \mathbb{P}(\mathbf{z}^2 \geq t)\mathrm{d}t.$$

The change of variable $t = u^2$ gives the result. □

Applying this lemma yields that It holds that

$$\begin{aligned}\mathbb{E}[\|x - \mathbf{x}_{\mathbf{i_1(x)}}\|^2] &= 2\int_0^{+\infty} u\mathbb{P}(\|x - \mathbf{x}_{\mathbf{i_1(x)}}\| \geq u)\mathrm{d}u \\ &\leq 2\int_0^{+\infty} u\exp(-\omega_d 2^{-d} p_{\min} n u^d)\mathrm{d}u.\end{aligned}$$

This last integral can be computed through the change of variables $v = \omega_d 2^{-d} p_{\min} n u^d$ and by recognizing the expression of the Gamma function[2].

Lemma 4.4. *Assume that condition (A4) holds and let* $x \in [0,1]^d$*. Then, it holds that*

$$\mathbb{E}[\|x - \mathbf{x}_{\mathbf{i_1(x)}}\|^2] \leq \frac{\gamma}{n^{2/d}}, \tag{31}$$

where $\gamma = \frac{8\Gamma(2/d)}{d(\omega_d p_{\min})^{2/d}}$.

We consider now the case $k > 1$. In this case, the approximation error satisfies

$$\begin{aligned}\mathbb{E}[\mathrm{App}(x)^2] &\leq \alpha^2 \mathbb{E}\left[\left(\frac{1}{k}\sum_{i\in I_k} \|\mathbf{x_i} - x\|\right)^2\right] \\ &\leq \frac{\alpha^2}{k}\mathbb{E}\left[\sum_{i\in I_k} \|\mathbf{x_i} - x\|^2\right] \text{ by Jensen inequality.}\end{aligned} \tag{32}$$

The sum of squared distances is bounded thanks to an elementary (but elegant) idea: for any set J of k indexes, we have

$$\sum_{i\in I_k} \|\mathbf{x_i} - x\|^2 \leq \sum_{j\in J} \|\mathbf{x_j} - x\|^2. \tag{33}$$

[2]See https://en.wikipedia.org/wiki/Gamma_function.

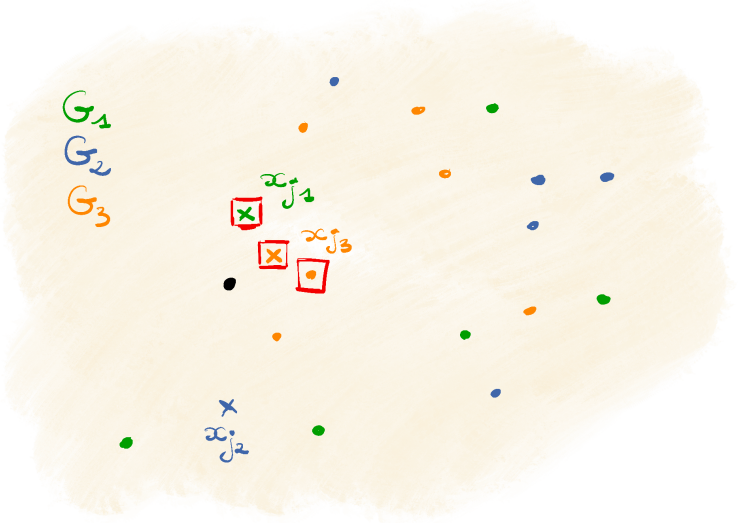

Figure 5: The red squares indicate the 3 nearest neighbors from the black dot x. Each color represents a group G_l of observations, whereas the crossed point is the nearest neighbor $\mathbf{x_{j_l}}$ to x in that group. The set of points $\{\mathbf{x_{j_1}}, \ldots, \mathbf{x_{j_k}}\}$ is always farther from x on average than the set of k-nearest neighbors.

Indeed, if we pick some index j_0 not in I_k in our set J, then the sum of the squared distances over indexes in J can always be decreased by replacing j_0 by one of the indexes of I_k that is not in J. The set J is built by splitting the set of observations $\mathbf{x_1}, \ldots, \mathbf{x_n}$ in k different groups of size roughly n/k. For sake of simplicity, we will assume that n/k is an integer and let $G_l = \{\mathbf{x_{n(l-1)/k+1}}, \ldots, \mathbf{x_{nl/k}}\}$ for $l = 1, \ldots, k$, that is G_1 contains the first n/k observations, G_2 the next n/k observations, and so on. We let j_l be the index of the nearest neighbor of x in the set G_l. See also Figure 5. Then, $\|x - \mathbf{x_{j_l}}\|^2$ is the squared distance between a point x and its nearest neighbor from a sample of n/k observations with distribution $P_{\mathbf{x}}$. According to Lemma 4.4, we have

$$\mathbb{E}[\|x - \mathbf{x_{j_l}}\|^2] \leq \frac{\gamma}{(n/k)^{2/d}}.$$

We define $J = \{j_1, \ldots, j_k\}$. Equation (33) then yields

$$\begin{aligned} \mathbb{E}[\sum_{i \in I_k} \|\mathbf{x_i} - x\|^2] &\leq \mathbb{E}[\sum_{j \in J} \|\mathbf{x_j} - x\|^2] \\ &\leq \sum_{l=1}^{k} \mathbb{E}[\|x - \mathbf{x_{j_l}}\|^2] \leq k\gamma \left(\frac{k}{n}\right)^{2/d}. \end{aligned} \tag{34}$$

Putting together (27), (32) and this last equation yields the following theorem.

Theorem 4.5 (Excess of risk of the k-nearest neighbor estimator)**.** *Assume that conditions (A1), (A2) and (A4) hold. Then, the k-nearest neighbor estimator $\hat{f}_k^{\mathrm{NN}}$ satisfies*

$$\mathbb{E}[\mathcal{R}_P(\hat{f}_k^{\mathrm{NN}}) - \mathcal{R}_P(f_P^\star)] \leq 2\alpha^2\gamma \left(\frac{k}{n}\right)^{2/d} + 2\frac{\sigma^2}{k}. \tag{35}$$

In particular, if $k = cn^{2/(d+2)}$ for some constant c, we obtain a bound of the form

$$\mathbb{E}[\mathcal{R}_P(\hat{f}_k^{\mathrm{NN}}) - \mathcal{R}_P(f_P^\star)] \leq Cn^{-2/(d+2)} \tag{36}$$

for some larger constant C.

For an optimal choice of k, the excess of risk of the k-NN estimator is of the same order $n^{-2/(d+2)}$ as the excess of risk of the partition estimator of Section 2. In particular, the k-NN estimator also suffers from the curse

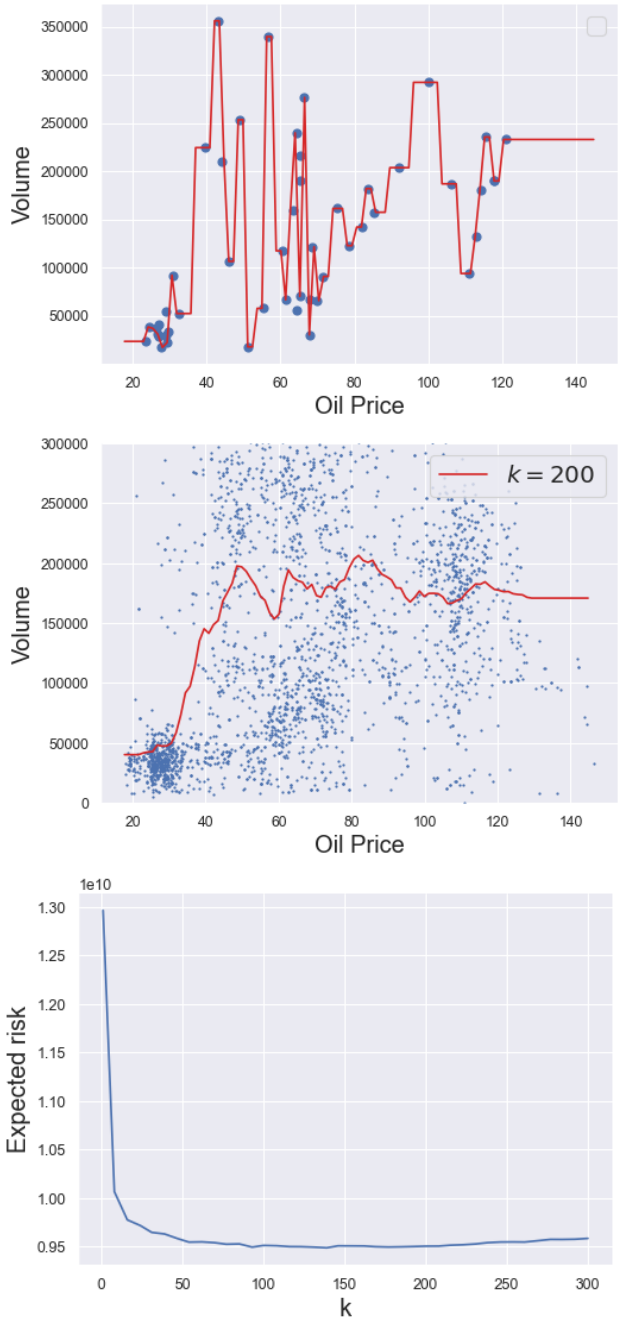

Figure 6: Top: the 1-NN estimator on a subsample of size $n = 50$. Middle: the k-NN estimator on the full dataset for the theoretical value $k = n^{2/3} \simeq 200$. Bottom: Expected risk for different values of k.

of dimensionality. One can actually prove that, in a certain sense, the curse of dimensionality is unavoidable if we only make assumptions (A1)-(A4) on the Bayes estimator $f_P^\star$. More structural assumptions on the function $f_P^\star$ are needed to obtain better rates of convergence in high dimension $d \gg 1$.

Example 4.6. Eventually, we apply the k-NN estimator to the oil dataset. First, for visualization purposes, we plot the k-NN estimator for $k = 1$ on a subset of $n = 50$ observations, see Figure 6. Theorem 4.5 predicts that a choice of k of order $n^{2/3}$ is optimal for such a problem: in our example, this gives a value of $k \simeq 200$, and the corresponding k-NN estimator is displayed in Figure 6. We then split the set of observations into a train set and a test set, while recording the excess of risk on the test set of $\hat{f}_k^{\mathrm{NN}}$ for different values of k. It appears that $k = 50$ is enough to obtain a small excess of risk. The theorem only gives a rough order of magnitude of what k should be and not a precise value. Cross-validation should be implemented to select the parameter k in practice.