Nonparametric Estimation

Evgeny Burnaev

Skoltech, Moscow, Russia

Outline

Problem Statement

- $S_m = \{\mathbf{x}_1, \dots, \mathbf{x}_m\} \sim F$ is a given sample, $\mathbf{x} \in \mathbb{R}^1$
- $F(\mathbf{x})$ is an absolutely continuous CDF with an unknown density $p(\mathbf{x})$
- We estimate $p(\mathbf{x})$ in the point $\mathbf{x}$, i.e. construct $\hat{p}_m(\mathbf{x}) = \hat{p}_m(\mathbf{x}|S_m)$
- Earlier we assume that

$$p \in \{p(\mathbf{x}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}, \Theta \subset \mathbb{R}^N,$$

  i.e. we use some parametric family
- Now we do not use such assumption $\Rightarrow$ Nonparametric Estimation

Losses/Risk

- We estimate $\hat{p}_m(\mathbf{x}_0)$ for some $\mathbf{x}_0$
- We consider quadratic loss function:

Definition

Mean Squared Error:

$$MSE(\hat{p}_m, p; \mathbf{x}_0) = \mathbb{E}_p[(\hat{p}_m(\mathbf{x}_0) - p(\mathbf{x}_0))^2]$$

- If we construct $\hat{p}_m(\mathbf{x})\,\forall \mathbf{x} \in \mathbb{R}^1$, then we use

Definition

Mean Integrated Squared Error:

$$MISE(\hat{p}_m, p) = \mathbb{E}_p\left[\int_{\mathbb{R}} (\hat{p}_m(\mathbf{x}) - p(\mathbf{x}))^2 d\mathbf{x}\right]$$

Bias-variance decomposition

Definition

Bias: $\mathrm{bias}(\mathbf{x}_0) = \mathbb{E}_p \hat{p}_m(\mathbf{x}_0) - p(\mathbf{x}_0)$

We get the following decomposition

Lemma

$$MSE(\hat{p}_m, p, \mathbf{x}_0) = \mathrm{bias}^2(\mathbf{x}_0) + \mathbb{V}_p(\hat{p}_m(\mathbf{x}_0)) =$$
$$= [\mathbb{E}_p \hat{p}_m(\mathbf{x}_0) - p(\mathbf{x}_0)]^2 + \mathbb{E}_p[\hat{p}_m(\mathbf{x}_0) - \mathbb{E}_p \hat{p}_m(\mathbf{x}_0)]^2$$

Lemma

$$MISE(\hat{p}_m, p) = \int_{\mathbb{R}} \mathrm{bias}^2(\mathbf{x}) d\mathbf{x} + \int_{\mathbb{R}} \mathbb{V}_p(\hat{p}_m(\mathbf{x})) dx$$

We will use these statements when constructing optimal density estimates

Histogram

- The simplest way to estimate density is to construct a histogram
- Let us consider an interval $[a, b) \ni \{\mathbf{x}_1, \ldots \mathbf{x}_m\}$ and divide it into N equal bins Δ_i of size $h = \frac{b-a}{N}$:

$$\Delta_i = [a + ih, a + (i+1)h), i = 0, 1, \ldots, N-1$$

- Let ν_i be a number of data points, belonging to Δ_i

Definition

$$\hat{p}_m(\mathbf{x}) = \begin{cases} \frac{\nu_0}{mh}, & \mathbf{x} \in \Delta_0, \\ \ldots & \\ \frac{\nu_{N-1}}{mh}, & \mathbf{x} \in \Delta_{N-1}; \end{cases} = \frac{1}{mh} \sum_{i=0}^{N-1} \nu_i \mathbb{I}\{\mathbf{x} \in \Delta_i\}$$

For $\mathbf{x} \in \Delta_i$ and small h:

$$\mathbb{E}_p \hat{p}_m(\mathbf{x}) = \frac{\mathbb{E}\nu_j}{mh} = \frac{\int_{\Delta_j} p(\mathbf{z}) d\mathbf{z}}{h} \approx \frac{p(\mathbf{x})h}{h} = p(\mathbf{x})$$

Smoothing Selection: Bias I

- Let us consider approaches to select h (smoothing parameter)
- Let us consider $\mathbf{x}_0 \in \Delta_j$:

$$
\begin{aligned}
\mathrm{bias}(\mathbf{x}_0) &= \mathbb{E}_p \hat{p}_m(\mathbf{x}_0) - p(\mathbf{x}_0) = \frac{1}{h}\int_{\Delta_j} p(\mathbf{x}) d\mathbf{x} - \frac{1}{h}\int_{\Delta_j} p(\mathbf{x}_0) d\mathbf{x} = \\
&= \frac{1}{h}\int_{\Delta_j} (p(\mathbf{x}) - p(\mathbf{x}_0)) d\mathbf{x} \approx \frac{1}{h}\int_{\Delta_j} p'(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) d\mathbf{x} \approx \\
&\approx p'(\mathbf{x}_0)\left[a + \left(j + \frac{1}{2}\right) h - \mathbf{x}_0\right]
\end{aligned}
$$

Smoothing Selection: Bias II

- $$\int_a^b \mathrm{bias}^2(\mathbf{x})d\mathbf{x} = \sum_{j=0}^{N-1}\int_{\Delta_j}\mathrm{bias}^2(\mathbf{x})d\mathbf{x} \approx$$
$$\approx \sum_{j=0}^{N-1}\int_{\Delta_j}[p'(\mathbf{x})]^2[a+(j+\frac{1}{2})h-\mathbf{x}]^2 d\mathbf{x} \approx$$
$$\approx \sum_{j=0}^{N-1}[p'(a+(j+\frac{1}{2})h)]^2\int_{\Delta_j}(a+(j+\frac{1}{2})h-\mathbf{x})^2 d\mathbf{x}$$
$$= \sum_{j=0}^{N-1}[p'(a+(j+\frac{1}{2})h)]^2\left(-\frac{(a+(j+\frac{1}{2})h-\mathbf{x})^3}{3}\right)\Bigg|_{\Delta_j} \approx$$
$$\approx \left(\int_a^b [p'(\mathbf{x})]^2 d\mathbf{x}\right)\frac{h^2}{12}$$

Smoothing Selection: Variance

$$\nu_j \sim \mathrm{Binom}\left(\int_{\Delta_j} p(\mathbf{x})d\mathbf{x}, m\right)$$

$$\mathbb{V}_p(\hat{p}_m(\mathbf{x}_0)) = \mathbb{V}_p\left(\frac{\nu_j}{mh}\right) = \frac{1}{(mh)^2}\mathbb{V}_p(\nu_j) =$$

$$= \frac{1}{(mh)^2} m \int_{\Delta_j} p(\mathbf{x})d\mathbf{x}\left(1 - \int_{\Delta_j} p(\mathbf{x})d\mathbf{x}\right) \approx \frac{1}{mh^2}\int_{\Delta_j} p(\mathbf{x})d\mathbf{x}$$

$$\int_a^b \mathbb{V}_p\left(\hat{p}_m(\mathbf{x})\right)d\mathbf{x} \approx \sum_{j=0}^{N-1}\left(\frac{1}{mh^2}\int_{\Delta_j} p(\mathbf{x})d\mathbf{x}\right)h =$$

$$= \frac{1}{mh}\int_a^b p(\mathbf{x})d\mathbf{x} = \frac{1}{mh}$$

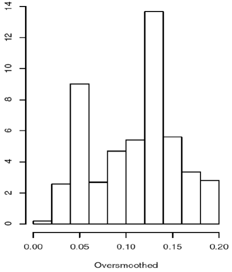
Smoothing Selection: MISE

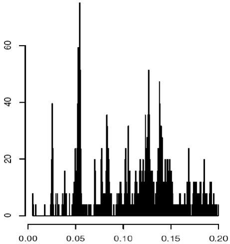
- Thus we get that

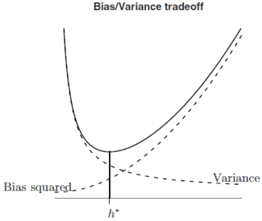
$$MISE(\hat{p}_m, p) \approx \left(\int_{\mathbb{R}} [p'(\mathbf{x})]^2 d\mathbf{x} \right) \frac{h^2}{12} + \frac{1}{nh}$$

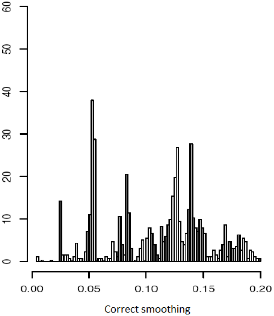
- The bigger h we use the bigger bias and smaller variance we get, and vice versa: Bias-Variance Tradeoff
- Too big $h =$ oversmoothing, too small $h =$ undersmoothing

Smoothing Selection: Example









Optimal Smoothing

- Value h, which $\approx$ minimizes $MISE$, is equal to

$$h^* = \frac{1}{m^{\frac{1}{3}}} \left(\frac{6}{\int_{\mathbb{R}} [p'(\mathbf{x})]^2 d\mathbf{x}} \right)^{\frac{1}{3}}$$

- Then we get that

$$MISE(\hat{p}_m, p) \approx \frac{C}{m^{\frac{2}{3}}},\ \textit{where } C = \left(\frac{3}{4}\right)^{\frac{2}{3}} \left(\int_{\mathbb{R}} [p'(\mathbf{x})]^2 d\mathbf{x} \right)^{\frac{1}{3}}$$

- Thus for a histogram with optimal h, we get that

$$MISE = O(m^{-\frac{2}{3}})$$

Smoothing Selection: Empirical Risk

- In practice it is not possible to calculate h^* since h^* depends on an unknown density
- Thus we should estimate $MISE$ and minimize it w.r.t. to h
- Since

$$\int_{\mathbb{R}} (\hat{p}_m(\mathbf{x}) - p(\mathbf{x}))^2 d\mathbf{x} = \int_{\mathbb{R}} \hat{p}_m(\mathbf{x})^2 d\mathbf{x} - 2\int_{\mathbb{R}} \hat{p}_m(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} + \int_{\mathbb{R}} p(\mathbf{x})^2 d\mathbf{x},$$

then we can only miminize

$$\mathcal{J}(h) = \int_{\mathbb{R}} \hat{p}_m(\mathbf{x})^2 d\mathbf{x} - 2\int_{\mathbb{R}} \hat{p}_m(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

Smoothgin Selection: Cross-Validation

Definition

Cross-Validation for Risk Estimation:

$$\hat{\mathcal{J}}(h) = \int_{\mathbb{R}} [\hat{p}_m(\mathbf{x})]^2 d\mathbf{x} - \frac{2}{m} \sum_{i=1}^{m} \hat{p}_{(-i)}(\mathbf{x}_i),$$

where $\hat{p}_{(-i)}(\mathbf{x})$ is a histogram estimated using the sample with i-th observation excluded

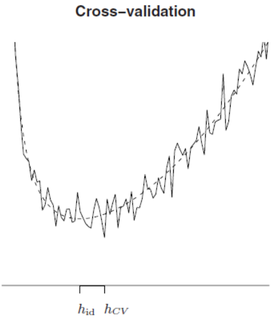
Theorem

$$\mathbb{E}\hat{\mathcal{J}}(h) \approx \mathbb{E}\mathcal{J}(h)$$

$$\hat{\mathcal{J}}(h) = \frac{2}{(m-1)h} - \frac{m+1}{(m-1)h} \sum_{i=1}^{N} \left(\frac{\nu_j}{m}\right)^2$$

Typical Risk behaviour

Typical behavior of $\hat{\mathcal{J}}(h)$:



Thus instead of uknown $MISE$ we can optimize $\hat{\mathcal{J}}(h)$ and find optimal h_{cv} which is sufficiently close to $h_{id} = h^*$

Confidence Tube: Definition

- Let us construct confidence intervals for $p(\mathbf{x})$
- For this we will use a histogram $\hat{p}_m(\mathbf{x})$, defined above
- Let us define

$$\overline{p}_m(\mathbf{x}) = \mathbb{E}\hat{p}_m(\mathbf{x}) = \frac{\int_{\Delta_j} p(\mathbf{z}) d\mathbf{z}}{h}, \ \mathbf{x} \in \Delta_j$$

In fact, $\overline{p}_m(\mathbf{x})$ is a histogram-averaging of $p(\mathbf{x})$

Definition

A pair of functions $(p_-(\mathbf{x}), p_+(\mathbf{x}))$ is a $1-\alpha$ confidence tube if

$$\mathbb{P}_p(p_-(\mathbf{x}) \leq \overline{p}_m(\mathbf{x}) \leq p_+(\mathbf{x}) \ \ \forall \mathbf{x}) \geq 1-\alpha$$

Confidence Tube: Properties

Theorem

Let $N = N(m)$ *is a number of bins in a histogram* $\hat{p}_m$*, such that* $N(m) \to \infty$ *and* $\frac{N(m)\log(m)}{m} \to \infty$ *for* $m \to \infty$
Let us define

$$p_-(\mathbf{x}) = (\max\{\sqrt{\hat{p}_m(\mathbf{x})} - C, 0\})^2, p_+(\mathbf{x}) = (\sqrt{\hat{p}_m(\mathbf{x})} + C)^2,$$

where $C = \frac{z_{\frac{\alpha}{2N}}}{2}\sqrt{\frac{N}{m(b-a)}}$
Then $(p_-(\mathbf{x}), p_+(\mathbf{x}))$ *is an* $1-\alpha$ *confidence tube*

Confidence Tube: Proof I

Proof:

☐ From the central limit theorem we get that

$$\frac{\nu_j}{m} = \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\{\mathbf{x}_i \in \Delta_j\}$$
$$\sim \mathcal{N}\left(\int_{\Delta_j} p(\mathbf{x})d\mathbf{x}, \frac{\int_{\Delta_j} p(\mathbf{x})d\mathbf{x}(1 - \int_{\Delta_j} p(\mathbf{x})d\mathbf{x})}{m}\right)$$

Using delta-method we get that $\sqrt{\frac{\nu_j}{m}} \sim \mathcal{N}\left(\sqrt{\int_{\Delta_j} p(\mathbf{x})d\mathbf{x}}, \frac{1}{4m}\right)$.

Moreover, we can prove that $\sqrt{\frac{\nu_j}{m}}$ are approximately independent.

Then $2\sqrt{m}\left(\sqrt{\frac{\nu_j}{m}} - \sqrt{\int_{\Delta_j} p(\mathbf{x})d\mathbf{x}}\right) \approx \xi_j$, where
$\xi_0, \ldots, \xi_{N-1} \sim \mathcal{N}(0, 1)$

Confidence Tube: Proof II

Let us set $A = \{p_-(\mathbf{x}) \le \overline{p}(\mathbf{x}) \le p_+(\mathbf{x}) \ \forall \mathbf{x}\} =$

$$= \{\sqrt{\hat{p}_m(\mathbf{x})} - C \le \sqrt{\overline{p}(\mathbf{x})} \le \sqrt{\hat{p}_m(\mathbf{x})} + C \ \ \forall \mathbf{x} \ \} =$$

$$= \{\max_{\mathbf{x}} |\sqrt{\hat{p}_m(\mathbf{x})} - \sqrt{\overline{p}(\mathbf{x})}| \le C\}$$

$$\mathbb{P}(A^c) = \mathbb{P}\{\max_{\mathbf{x}} |\sqrt{\hat{p}_m(\mathbf{x})} - \sqrt{\overline{p}(\mathbf{x})}| > C\}$$

$$= \mathbb{P}\left\{\max_{j=\overline{0,N-1}} \left|\sqrt{\frac{\nu_j}{mh}} - \sqrt{\frac{\int_{\Delta_j} p(\mathbf{x}) d\mathbf{x} h}{h}}\right| > C\right\}$$

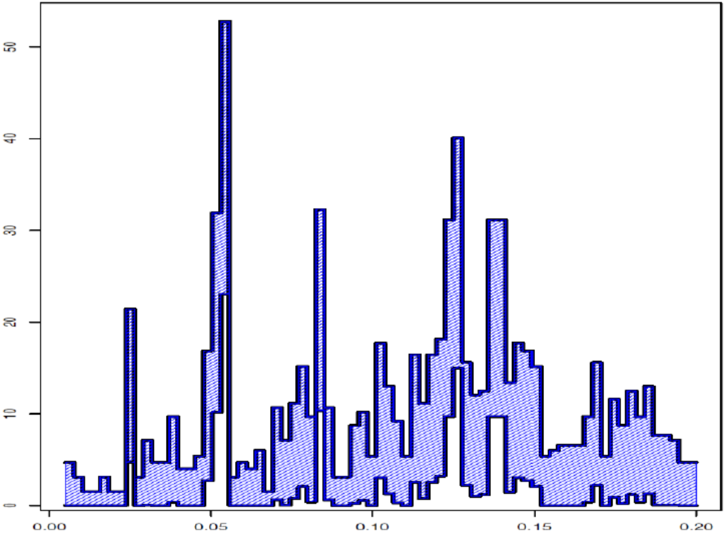
$$\approx \mathbb{P}\left\{\max_{j=\overline{0,N-1}} \frac{|\xi_j|}{2\sqrt{mh}} > \frac{z_{\frac{\alpha}{2n}}}{2}\sqrt{\frac{N}{m(b-a)}}\right\}$$

$$= \mathbb{P}\{\max_{j=\overline{0,N-1}} |\xi_j| > z_{\frac{\alpha}{2N}}\} \le \sum_{j=0}^{N-1} \mathbb{P}\{|\xi_j| > z_{\frac{\alpha}{2N}}\} =$$

$$\sum_{j=0}^{N-1} \frac{\alpha}{N} = \alpha,$$

i.e. for such $p_-(\mathbf{x}), p_+(\mathbf{x})$ conditions on confidence tube are correct. ■

Confidence Tube: Example



1. Nonparametric Density Estimation
2. Kernel Density Estimation
3. Nonparametric regression

Kernels

KDE allows to get smoother estimate (compared to histogram based ones) with faster convergence rates

Definition

Kernel is a function K, such that

$$K(\mathbf{x}) \geq 0, \int_{\mathbb{R}} K(\mathbf{x}) d\mathbf{x} = 1, \int_{\mathbb{R}} \mathbf{x} K(\mathbf{x}) d\mathbf{x} = 0, \sigma_K^2 \equiv \int_{\mathbb{R}} \mathbf{x}^2 K(\mathbf{x}) d\mathbf{x}$$

Examples

- $K(x) = \frac{1}{2}\mathbb{I}\{|x| < 1\}$ — rectangular kernel
- $K(x) = (1 - |x|)\mathbb{I}\{|x| < 1\}$ — triangle kernel
- $K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ — Gaussian kernel
- $K(x) = \frac{3}{4}(1 - x^2)\mathbb{I}\{|x| < 1\}$ — Epanechnikov kernel

In the sequel we will consider only smooth kernels

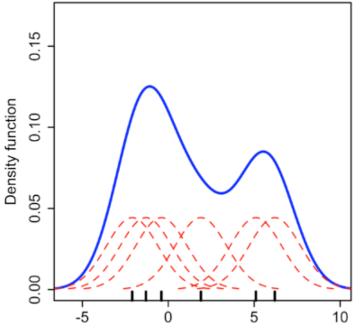
Kernel Density Estimation: Definition

Definition

KDE has the form

$$\hat{p}_m(\mathbf{x}) = \frac{1}{mh} \sum_{i=1}^{m} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right),$$

h is a kernel width



KDE: MISE

A shape of K influence quality of estimate not so significant compared to a value of h

Theorem

$$MISE(\hat{p}_m, p) \approx \frac{1}{4}\sigma_K^4 h^4 \int_{\mathbb{R}} (p''(\mathbf{x}))^2 d\mathbf{x} + \frac{1}{mh} \int_{\mathbb{R}} (K(\mathbf{x}))^2 d\mathbf{x}$$

For $h = h^*$ we get minimum of the risk

$$h^* = \left(\frac{1}{m} \frac{\int_{\mathbb{R}} (K(\mathbf{x}))^2 d\mathbf{x}}{\left(\int_{\mathbb{R}} \mathbf{x}^2 K(\mathbf{x}) d\mathbf{x}\right)^2 \left(\int_{\mathbb{R}} (p''(\mathbf{x}))^2 d\mathbf{x}\right)} \right)^{\frac{1}{5}}$$

For $h = h^*$ we get that $MISE(\hat{p}_m, p) = O\left(\frac{1}{m^{\frac{4}{5}}}\right)$

MISE: Proof I

Proof: let us use a bias-variance decomposition

□

- $\text{bias}(\mathbf{x}) = \mathbb{E}_p \hat{p}_m(\mathbf{x}) - p(\mathbf{x}) =$
  $\int_{\mathbb{R}} (\frac{1}{mh} \sum_{i=1}^{m} K(\frac{\mathbf{x}-\mathbf{x}_i}{h})) p(\mathbf{x}_1) \ldots p(\mathbf{x}_m) d\mathbf{x}_1 \ldots d\mathbf{x}_m -$
  $\frac{1}{m} \sum_{i=1}^{m} \int_{\mathbb{R}} K(\mathbf{z}) p(\mathbf{z}) d\mathbf{z} \approx$
  $\int_{\mathbb{R}} K(\mathbf{z}) [-p'(\mathbf{x}) \mathbf{z} h + p''(\mathbf{x}) \frac{(\mathbf{z}h)^2}{2}] d\mathbf{z} = \frac{1}{2} \sigma_K^2 h^2 p''(\mathbf{x})$
- $\int_{\mathbb{R}} (\text{bias}(\mathbf{x}))^2 d\mathbf{x} = \frac{1}{4} \sigma_K^4 h^4 \int_{\mathbb{R}} [p''(\mathbf{x})]^2 d\mathbf{x}$

MISE: Proof II

We get that

- $\int_{\mathbb{R}} \mathbb{V}_p(\hat{p}_m(\mathbf{x}))d\mathbf{x} = \int_{\mathbb{R}} \mathbb{V}_p[\frac{1}{mh}\sum_{i=1}^m K(\frac{\mathbf{x}-\mathbf{x}_i}{h})]d\mathbf{x} =$
  $\frac{1}{(mh)^2}\sum_{i=1}^m \int_{\mathbb{R}} \mathbb{V}_p(K(\frac{\mathbf{x}-\mathbf{x}_i}{h}))d\mathbf{x} \leq$
  $\frac{1}{(mh)^2}\sum_{i=1}^m \int_{\mathbb{R}} \mathbb{E}_p K(\frac{\mathbf{x}-\mathbf{x}_i}{h})^2 d\mathbf{x} =$
  $\frac{1}{(mh)^2}\sum_{i=1}^m \int_{\mathbb{R}} \int_{\mathbb{R}} K(\frac{\mathbf{x}-\mathbf{x}_i}{h})^2 p(\mathbf{x}_i)d\mathbf{x}_i d\mathbf{x} =$
  $\frac{1}{(mh)^2}\sum_{i=1}^m \int_{\mathbb{R}} p(\mathbf{x}_i)\int_{\mathbb{R}} K(\frac{\mathbf{x}-\mathbf{x}_i}{h})^2 d\mathbf{x}d\mathbf{x}_i =$
  $\frac{1}{(mh)^2}\sum_{i=1}^m \int_{\mathbb{R}} p(\mathbf{x}_i)d\mathbf{x}_i h\int_{\mathbb{R}} K^2(\mathbf{z})d\mathbf{z} = \frac{1}{mh}\int_{\mathbb{R}} K^2(\mathbf{z})d\mathbf{z}$

Thus we see that for some h^* we get a minimum of $MISE(\hat{p}_m, p)$
■

Kernel width selection: comments

- For h^* and $\hat{p}_m$ we get that $MISE = O(m^{-\frac{4}{5}})$, that is KDE is better than histogram estimate
- It can be proved that under some rather general conditions it is not possible to find a convergence speed better than $m^{\frac{4}{5}}$
- As it is with a histogram, for big h we get oversmoothing, and for small h – we get undersmoothing (due to bias)

KDE: Cross-Validation

Risk function is equal to

$$\hat{J}(h) = \int_{\mathbb{R}} \hat{p}_m^2(\mathbf{x}) d\mathbf{x} - \frac{2}{m} \sum_{i=1}^{m} \hat{p}_{(-i)}(\mathbf{x}_i)$$

Theorem

For any $h > 0$ we get that $\mathbb{E}[\hat{J}(h)] \approx \mathbb{E}[J(h)]$. Moreover,

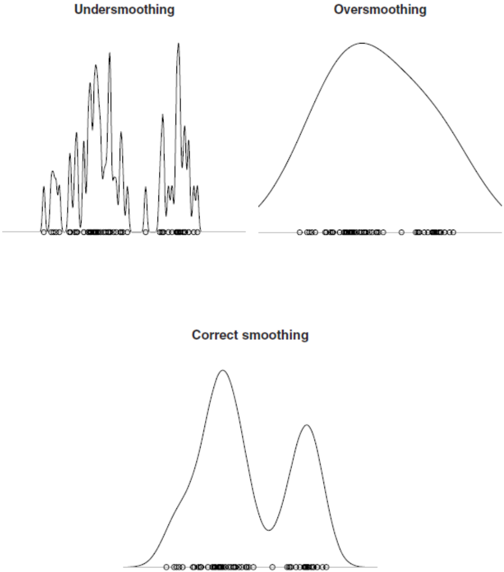
$$\hat{J}(h) \approx \frac{1}{mh^2} \sum_{i,j} K^* \left(\frac{\mathbf{x}_i - \mathbf{x}_j}{h} \right) + \frac{2}{mh} K(0),$$

where

$$K^*(\mathbf{x}) = K^{(2)}(\mathbf{x}) - 2K(\mathbf{x}),\ K^{(2)}(\mathbf{z}) = \int K(\mathbf{z} - \mathbf{x}) K(\mathbf{x}) d\mathbf{x}$$

E.g., if $K = \mathcal{N}(0,1)$, then $K^{(2)} = \mathcal{N}(0,2)$

Kernel width selection: Example



Confidence interval for averaged density I

- Let us set $\overline{p}_m(\mathbf{x}) = \mathbb{E}\hat{p}_m(\mathbf{x}) = \int_{\mathbb{R}} \frac{1}{h} K(\frac{\mathbf{x}-\mathbf{z}}{h}) p(\mathbf{z}) d\mathbf{z}$
- Let us assume that $\operatorname{supp}(p) \subset [a, b)$
- Then we can define $(1-\alpha)$-confidence tube

$$p_{-}(\mathbf{x}) = \hat{p}_m(\mathbf{x}) - \frac{z_\alpha}{\sqrt{m}} s(\mathbf{x}), p_{+}(\mathbf{x}) = \hat{p}_m(\mathbf{x}) + \frac{z_\alpha}{\sqrt{m}} s(\mathbf{x}),$$

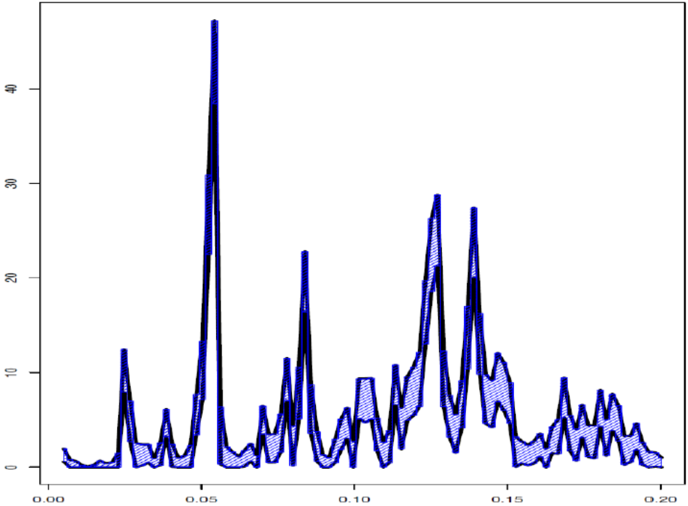
where

— $s^2(\mathbf{x}) = \frac{1}{m-1} \sum_{i=1}^{m} [Y_i(\mathbf{x}) - \overline{Y}_m(\mathbf{x})]^2, Y_i(\mathbf{x}) = \frac{1}{h} K(\frac{\mathbf{x}-\mathbf{x}_i}{h})$,

— $z_\alpha = \Phi^{-1}\left(\frac{1+(1-\alpha)^{\frac{w}{b-a}}}{2}\right)$, $\Phi(\cdot)$ is a function of a standard normal distribution

— w is an effective kernel width (for a Gaussian kernel smoothing $w = 3h$)

Confidence interval for averaged density II



Multidimensional KDE I

Let us consider a multidimensional case, i.e. $\mathbf{x} = [x_1, \ldots, x_N]^{\mathrm{T}}$ is a point in $\mathbb{R}^N$. Thus $i-$th observation is an N-dimensional vector:

$$\mathbf{x}_i = [x_i^1, \ldots x_i^N]^{\mathrm{T}}$$

Let $h = [h_1, \ldots, h_N]^{\mathrm{T}}$ be a vector of kernel widths
Then

$$\hat{p}_m(\mathbf{x}) = \frac{1}{m h_1 \cdot \ldots \cdot h_N} \sum_{i=1}^{m} \left[\prod_{j=1}^{N} K\left(\frac{x^j - x_i^j}{h_j} \right) \right]$$

Multidimensional KDE II

For such estimate the risk is equal to

$$
\begin{aligned}
MISE(\hat{p}_m, p) \approx & \frac{1}{4}\sigma_K^4 \Big[\sum_{j=1}^{N} h_j^4 \int_{\mathbb{R}^N} p_{jj}^2(\mathbf{x}) d\mathbf{x} \\
& + \sum_{j \neq k} h_j^2 h_k^2 \int_{\mathbb{R}^N} p_{jj}(\mathbf{x}) p_{kk}^2(\mathbf{x}) d\mathbf{x} \Big] \\
& + \frac{\left(\int_{\mathbb{R}^N} K^2(\mathbf{x}) d\mathbf{x} \right)^N}{m h_1 \cdot \ldots \cdot h_N},
\end{aligned}
$$

where

$$
p_{jj}(\mathbf{x}) = \frac{\partial^2 p(\mathbf{x})}{\partial x_j^2}
$$

Optimal kernel width is equal to $h_i^* \approx m^{-\frac{1}{4+N}}$

The risk has the form

$$
MISE(\hat{p}_m, p) = O(m^{-\frac{4}{4+N}})
$$

Curse of Dimensionality

- Optimal rate of convergence is $O(n^{-\frac{4}{4+N}})$: if N increases convergence rate decreases
- Let us consider a table with values of m necessary to get the mean squared estimation error in $\mathbf{x}_0 = 0$ less than 0.1 depending on N for a multidimensional normal density and optimal kernel width:

N	1	2	3	4	5	6	7	8	9
m	4	19	67	223	768	2790	10700	43700	187000

- Here N is a dimension, m is a sample size

Nonparametric regression: Definition

- Let us consider m observations:
  $S_m = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$, generated from a joint density $p(\mathbf{x}, y)$
- These observations are generated by the model

$$y_i = r(\mathbf{x}_i) + \varepsilon_i,$$

  where ε_i is an i.i.d white noise, $\mathbb{E}\varepsilon_i = 0$, $\mathbb{V}(\varepsilon_i) = \sigma^2$
- We should estimate a regression function

$$r(\mathbf{x}) = \mathbb{E}(y|\mathbf{x}) = \int_{\mathbb{R}} yp(y|\mathbf{x})dy = \frac{\int_{\mathbb{R}} yp(\mathbf{x}, y)dy}{\int_{\mathbb{R}} p(\mathbf{x}, y)dy} = \frac{\int_{\mathbb{R}} yp(\mathbf{x}, y)dy}{p(\mathbf{x})}$$

Nadaraya-Watson estimate I

Definition

Let $\hat{p}_m(\mathbf{x})$ and $\hat{p}_m(\mathbf{x}, y)$ be kernel density estimates, obtained using samples $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ and $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$ correspondingly, and the kernel K. Then if $\hat{p}_m(\mathbf{x}) \neq 0$, then

$$\hat{r}_m^{NW}(\mathbf{x}) = \frac{\int_{\mathbb{R}} y \hat{p}_m(\mathbf{x}, y) dy}{\hat{p}_m(\mathbf{x})}$$

We can notice that Nadaraya-Watson estimate can be used also in case when $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ are some fixed and deterministic values, e.g. $\mathbf{x}_i = \frac{i}{m}$

Nadaraya-Watson estimate II

We use Nadaraya-Watson estimate for $r(\mathbf{x})$:

Definition

Nadaraya-Watson estimate has the form

$$\hat{r}_m^{NW}(\mathbf{x}) = \sum_{i=1}^{m} w_i(\mathbf{x}) y_i,$$

where

$$w_i = \frac{K(\frac{\mathbf{x}-\mathbf{x}_i}{h})}{\sum_{j=1}^{m} K(\frac{\mathbf{x}-\mathbf{x}_j}{h})},$$

and K is a some kernel function

Thus, the estimate is a weighted sum of y_i, where any point, close to $\mathbf{x}$, has a big weight

Nonparametric regression: MISE

- Let us consider the risk and optimal kernel width

Theorem

$$MISE(\hat{r}_m^{NW}, r) \approx \frac{h^4}{4}\left(\int_{\mathbb{R}} \mathbf{x}^2 K^2(\mathbf{x}) d\mathbf{x}\right)^4 \int \left(r''(\mathbf{x}) + 2r'(\mathbf{x})\frac{p'(\mathbf{x})}{p(\mathbf{x})}\right)^2 d\mathbf{x} + \frac{1}{h}\int_{\mathbb{R}} \frac{\sigma^2 \int_{\mathbb{R}} K^2(\mathbf{x}) d\mathbf{x}}{mp(\mathbf{x})} d\mathbf{x}$$

- Optimal kernel width has the form $h^* = \mathrm{const} m^{-\frac{1}{5}}$
- Then the risk is

$$MISE(\hat{r}_m^{NW}, r) = O(m^{-\frac{4}{5}})$$

Optimal Width

- Again we can not calculate h^* in practice, since it depends on unknown values of $r(\mathbf{x})$, $p(\mathbf{x})$
- Then we should minimize the risk estimate w.r.t. h

$$\hat{\mathcal{J}}(h) = \sum_{i=1}^{m} (y_i - \hat{r}^{NW}_{(-i)}(\mathbf{x}_i))^2,$$

where $\hat{r}^{NW}_{(-i)}$ is a Nadaraya-Watson estimate, constructed using the sample, from which the observation $(\mathbf{x}_i, y_i)$ is excluded

Theorem

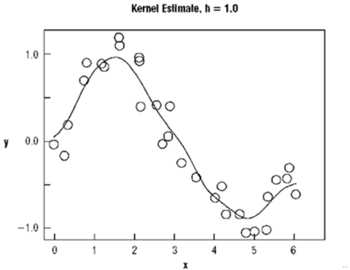
$$\hat{\mathcal{J}}(h) = \sum_{i=1}^{m} \left(y_i - \hat{r}^{NW}_{(-i)}(\mathbf{x}_i)\right)^2 \frac{1}{\left(1 - \frac{K(0)}{\sum_{j=1}^{m} K\left(\frac{\mathbf{x}_i - \mathbf{x}_j}{h}\right)}\right)^2}$$

Smoothing

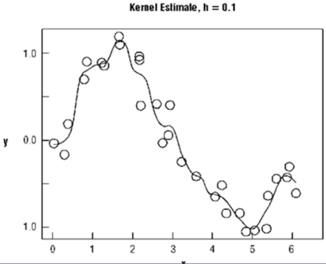
As with a histogram estimate and KDE we have a bias-variance trade-off:

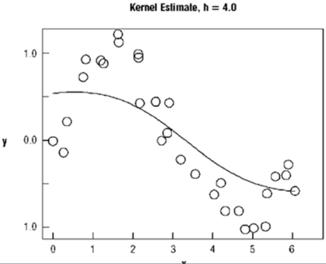
- for big h we have oversmoothing (the estimate is too smooth), and
- for small h we have undersmoothing (the estimate is to wiggly)

Nonparametric regression: Example



Correct smoothing





Confidence tube for regression I

- Let us construct a confidence tube
- First, let us estimate σ^2. Let $\mathbf{x}_i$ are ordered in increasing order. If $r(x)$ is smooth, we get that $r(\mathbf{x}_{i+1}) - r(\mathbf{x}_i) \approx 0$
- Then

$$y_{i+1} - y_i = [r(\mathbf{x}_{i+1}) - r(\mathbf{x}_i)] - [r(\mathbf{x}_i) + \varepsilon_i] \approx \varepsilon_{i+1} - \varepsilon_i$$

$$\mathbb{V}(y_{i+1} - y_i) \approx \mathbb{V}(\varepsilon_{i+1} - \varepsilon_i) = \mathbb{V}(\varepsilon_{i+1}) + \mathbb{V}(\varepsilon_i) = 2\sigma^2$$

$$\hat{\sigma}^2 = \frac{1}{2(m-1)} \sum_{i=1}^{m-1} (y_{i+1} - y_i)^2$$

- We construct a confidence tube for a smoothed version $\overline{r}_m(x) = \mathbb{E}(\hat{r}_m^{NW}(\mathbf{x}))$ of a real regression r

Confidence tube for regression II

Approximate $(1-\alpha)$ confidence interval for $\overline{r}_m(\mathbf{x})$ has the form

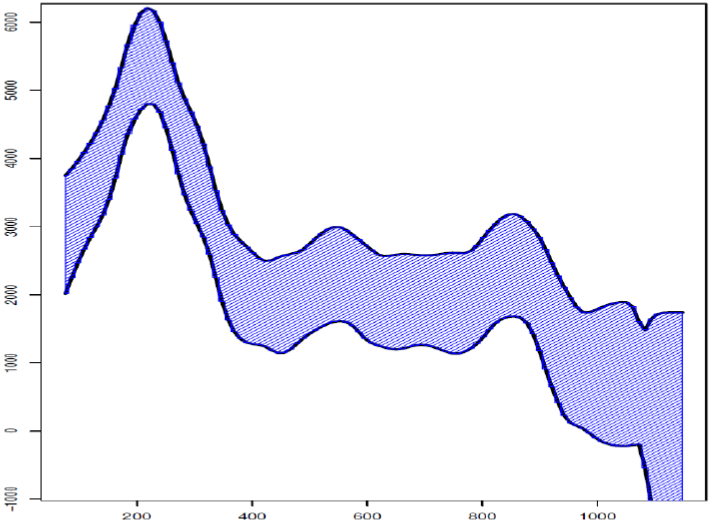
$$r_-(\mathbf{x}) = \hat{r}_m^{NW}(\mathbf{x}) - z_\alpha \hat{\sigma} \sqrt{\sum_{i=1}^m w_i^2(\mathbf{x})}$$

$$r_+(\mathbf{x}) = \hat{r}_m^{NW}(\mathbf{x}) + z_\alpha \hat{\sigma} \sqrt{\sum_{i=1}^m w_i^2(\mathbf{x})},$$

where

- $\hat{\sigma}$, and w_i are defined above,
- $z_\alpha = \Phi^{-1}\left(\frac{1+(1-\alpha)^{\frac{w}{b-a}}}{2}\right)$,
- Φ is a cumulative distribution function of a standard normal distribution
- w is an effective kernel width, $\mathbf{x}_1, \ldots, \mathbf{x}_m \in [a, b)$

Confidence tube for regression III



Confidence tube for regression: Comments I

- Constructed confidence tube (cf. with histogram and KDE) does not provide exact confidence intervals for the regression, but for its smoothed version
- E.g. confidence tube for KDE is in fact a confidence tube for a density, being equal to a smoothed (with the same kernel) initial density
- We can not construct a confidence interval for the initial density due to the following reasons:
  — Let $\hat{p}_m(\mathbf{x})$ is an estimate of $p(\mathbf{x})$. Let us define $\mathbb{E}\hat{p}_m(\mathbf{x}) = \overline{p}(\mathbf{x}), \mathbb{V}(\hat{p}_m(\mathbf{x})) = S_m(\mathbf{x})$, then

$$\frac{\hat{p}_m(\mathbf{x}) - p_m(\mathbf{x})}{S_m(\mathbf{x})} = \frac{\hat{p}_m(\mathbf{x}) - \overline{p}_m(\mathbf{x})}{S_m(\mathbf{x})} + \frac{\overline{p}_m(\mathbf{x}) - p_m(\mathbf{x})}{S_m(\mathbf{x})}$$

Confidence tube for regression: Comments II

- Thanks to CLT the first summand converges to a standard normal distribution, using which we can construct a confidence interval
- The second term is equal to the ratio of a bias to a standard deviation. In a parametric case usually bias is significantly smaller than standard deviation. I.e. the second term converges to zero when $m \to \infty$
- In a nonparametric case optimal smoothing leads to a balance of a bias and a standard deviation. Therefore the second term may not tend to zero even for big sample sizes. Due to this effect the confidence tube will not be centered w.r.t. the ground truth density

Structural Nonparametric Regression I

- If $\mathbf{x} = [x_1, \ldots, x_N]^{\mathrm{T}}$, then due to curse of dimensionality it is not reasonable to generalize NW estimate in the same way as KDE to the multidimensional case
- Instead we can consider an additive model, e.g.
  - $y = \sum_{j=1}^{N} r_j(x^j) + \alpha + \varepsilon$
    or
  - $y = \sum_{j=1}^{N} r_j(x^j) + \sum_{j<k} r_{jk}(x^j, x^k) + \alpha + \varepsilon$

Structural Nonparametric Regression II

Preparation of the first additive model

Algorithm (Backfitting)

Initialization: $\hat{\alpha} = \overline{y}_m; \hat{r}_1, \dots \hat{r}_N$

Until $\hat{r}_1, \dots, \hat{r}_N$ stabilize

- For all $j = 1, \dots, N$:
    1. Calculate $\widetilde{\varepsilon}_i = y_i - \hat{\alpha} - \sum_{k \neq j} \hat{r}_k(x_i^k), i = 1, \dots, m$
    2. Construct $\hat{r}_j(x^j)$ as a regression function of $\widetilde{\varepsilon}_i$ on j-th component x^j (i.e. as observations we use $\{(x_1^j, \widetilde{\varepsilon}_1), \dots, (x_m^j, \widetilde{\varepsilon}_m)\}$
    3. Set $\hat{r}_j := \hat{r}_j - \frac{1}{m} \sum_{i=1}^m \hat{r}_j(x_i^j)$