Nonparametric Estimation

Maxim Panov

Skoltech

November, 2021

Outline

Outline

Nonparametric Density Estimation
- Problem Statement
- Histograms for Density Estimation
- Kernel Density Estimation

Nonparametric Regression
- Nadaraya-Watson Estimator
- Confidence Band for Regression Function

Nonparametric Density Estimation: Problem Statement

Let $X_1, \dots, X_n \sim F$, where F is a CDF, corresponding to an unknown density p.

The goal is to estimate p at a point x, i.e. construct

$$\widehat{p}_n(x) = \widehat{p}_n(x; X_1, \dots, X_n).$$

Earlier, in similar tasks we assumed p to be a member of some parametric family:

$$p \in \{p(x; \theta),\ \theta \in \Theta\}, \Theta \subset \mathbb{R}^d.$$

Now we don't have this restriction $\Rightarrow$ nonparametric estimation.

Nonparametric Density Estimation: Loss and Risk

Let $\widehat{p}_n(x_0)$ be an estimate of the density at the point x_0.
If we consider the **squared loss function**, we can introduce the following quantity:

Definition

Mean Squared Error:

$$MSE(\widehat{p}_n, p; x_0) = \mathbb{E}_p\left[\left(\widehat{p}_n(x_0) - p(x_0)\right)^2\right].$$

If we have an estimate $\widehat{p}_n(x)$ at every point $x \in \mathbb{R}$, then

Definition

Mean Integrated Squared Error:

$$MISE(\widehat{p}_n, p) = \mathbb{E}_p\left[\int \left(\widehat{p}_n(x) - p(x)\right)^2 dx\right].$$

Bias-variance Decomposition

Definition

Bias: $bias(x_0) = \mathbb{E}_p \widehat{p}_n(x_0) - p(x_0)$.

The following decomposition of the error holds:

Lemma

$$MSE(\widehat{p}_n, p, x_0) = bias^2(x_0) + \mathbb{V}_p \widehat{p}_n(x_0) =$$
$$= \left[\mathbb{E}_p \widehat{p}_n(x_0) - p(x_0)\right]^2 + \mathbb{E}_p \left[\widehat{p}_n(x_0) - \mathbb{E}_p \widehat{p}_n(x_0)\right]^2.$$

Lemma

$$MISE(\widehat{p}_n, p) = \int bias^2(x)dx + \int \mathbb{V}_p \widehat{p}_n(x)dx.$$

We will use these lemmas later to construct “optimal” density estimates.

Histogram

Perhaps the simplest way to estimate density is to build a **histogram**.

- Take an interval $[a, b) \ni X_1, \ldots, X_n$.
- Divide it into N equal parts Δ_i of length $h = \frac{b-a}{N}$:

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

- Let ν_i be the number of samples that fell into Δ_i.

Definition

$$\widehat{p}_n(x) = \begin{cases} \frac{\nu_0}{nh}, & x \in \Delta_0, \\ \ldots & \\ \frac{\nu_{N-1}}{nh}, & x \in \Delta_{N-1} \end{cases} = \frac{1}{nh} \sum_{i=0}^{N-1} \nu_i \mathbb{I}\{x \in \Delta_i\}.$$

For $x \in \Delta_j$ and small h: $\mathbb{E}_p \widehat{p}_n(x) = \frac{\mathbb{E}\nu_j}{nh} = \frac{\int_{\Delta_j} p(u)du}{h} \approx \frac{p(x)h}{h} = p(x)$.

Histogram: Choosing Smoothness Parameter

We have introduced a hyperparameter h, which controls "smoothness" of the histogram.

How to choose it?

Lets perform calculations for $x_0 \in \Delta_j$:

$$bias(x_0) = \mathbb{E}_p \widehat{p}_n(x_0) - p(x_0) = \frac{1}{h} \int\limits_{\Delta_j} p(x) dx - \frac{1}{h} \int\limits_{\Delta_j} p(x_0) dx =$$

$$= \frac{1}{h} \int\limits_{\Delta_j} (p(x) - p(x_0)) dx \approx \frac{1}{h} \int\limits_{\Delta_j} p'(x_0)(x - x_0) dx \approx$$

$$\approx p'(x_0)[a + (j + \frac{1}{2})h - x_0].$$

Histogram: Choosing Smoothness Parameter

$$\int_a^b bias^2(x_0)dx_0 = \sum_{j=0}^{N-1} \int_{\Delta_j} bias^2(x_0)dx_0 \approx \sum_{j=0}^{N-1} \int_{\Delta_j} [p'(x_0)]^2 [a + (j + \frac{1}{2})h - x_0]^2 dx_0$$

$$\approx \sum_{j=0}^{N-1} [p'(a + (j + \frac{1}{2})h)]^2 \int_{\Delta_j} (a + (j + \frac{1}{2})h - x_0)^2 dx_0$$

$$= \sum_{j=0}^{N-1} [p'(a + (j + \frac{1}{2})h)]^2 \left(-\frac{(a + (j + \frac{1}{2})h - x_0)^3}{3} \right) \Bigg|_{\Delta_j} \approx \left(\int_a^b [p'(x)]^2 dx \right) \frac{h^2}{12}.$$

Histogram: Choosing Smoothness Parameter

Note that $\nu_j \sim Bin(\int\limits_{\Delta_j} p(x)dx, n)$. Then

$$\mathbb{V}_p \widehat{p}_n(x_0) = \mathbb{V}_p \frac{\nu_j}{nh} = \frac{1}{(nh)^2} \mathbb{V}_p \nu_j =$$
$$= \frac{1}{(nh)^2} n \int\limits_{\Delta_j} p(x)dx(1 - \int\limits_{\Delta_j} p(x)dx) \approx \frac{1}{nh^2} \int\limits_{\Delta_j} p(x)dx.$$

$$\int\limits_a^b \mathbb{V}_p \widehat{p}_n(x_0)dx_0 \approx \sum_{j=0}^{N-1} \left(\frac{1}{nh^2} \int\limits_{\Delta_j} p(x)dx \right) h = \frac{1}{nh} \int\limits_a^b p(x)dx = \frac{1}{nh}.$$

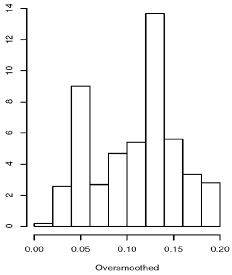
Histogram: Choosing Smoothness Parameter

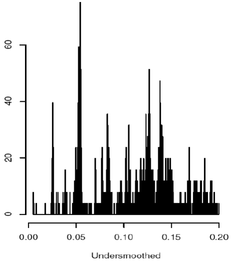
From that we get:

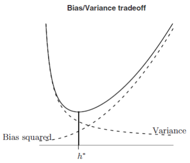
$$MISE(\widehat{p}_n, p) \approx \left(\int [p'(x)]^2 dx \right) \frac{h^2}{12} + \frac{1}{nh}.$$

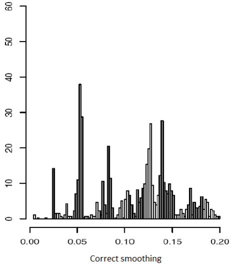
- Increasing h increases the bias and lowers the variance and vise versa.
- This is called bias-variance tradeoff.
- When h is too large – oversmoothing, too small – undersmoothing.

Histogram: Choosing Smoothness Parameter









Histogram: Choosing Smoothness Parameter

The value of h, for which $MISE$ is minimal:

$$h^* = \frac{1}{n^{\frac{1}{3}}} \left(\frac{6}{\int \left[p'(x)\right]^2 dx} \right)^{\frac{1}{3}}.$$

Also,

$$MISE(\widehat{p}_n, p) \approx \frac{C}{n^{\frac{2}{3}}}, \quad \text{where } C = \left(\frac{3}{4}\right)^{\frac{2}{3}} \left(\int \left[p'(x)\right]^2 dx \right)^{\frac{1}{3}}.$$

Hence, if we use a histogram with the optimal value of h, $MISE$ will be decreasing at the rate of $n^{-\frac{2}{3}}$.

Histogram: Choosing Smoothness Parameter

- In practice we can not find h^* since it depends on the unknown true density.
- Instead, we can estimate $MISE$ and minimize this estimate over h.

Since

$$\int \left(\widehat{p}_n(x) - p(x)\right)^2 dx = \int \widehat{p}_n(x)^2 dx - 2 \int \widehat{p}_n(x) p(x) dx + \int p(x)^2 dx,$$

then it suffices to estimate and minimize just

$$\mathcal{J}(h) = \int \widehat{p}_n(x)^2 dx - 2 \int \widehat{p}_n(x) p(x) dx.$$

Histogram: Choosing Smoothness Parameter

Definition

Risk estimate using cross-validation:

$$\widehat{\mathcal{J}}(h) = \int \left[\widehat{p}_n(x)\right]^2 dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{p}_{(-i)}(X_i),$$

where $\widehat{p}_{(-i)}$ – histogram estimate using all but i-th observation.

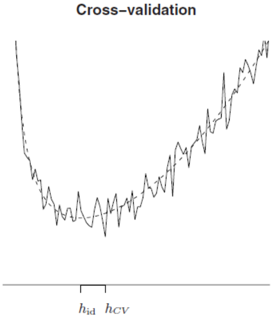
Theorem

For any $h > 0$

$$\mathbb{E}\widehat{\mathcal{J}}(h) = \mathbb{E}\mathcal{J}(h).$$

Histogram: Choosing Smoothness Parameter

Typical behavior of $\widehat{\mathcal{J}}(h)$ looks like this:



This way, instead of unknown $MISE$ we can minimize $\widehat{\mathcal{J}}(h)$ and find optimal h_{cv}, that will be not far from $h_{id} = h^*$.

Kernel Density Estimation

This method gives faster converging and more smooth estimates than the histogram.

Definition

Kernel is a function K that satisfies the following properties:

$$K(x) \geq 0, \ \int K(x)dx = 1, \ \int xK(x)dx = 0, \ \sigma_K^2 \equiv \int x^2 K(x)dx.$$

Examples

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

In what follows, we will only consider smooth kernels.

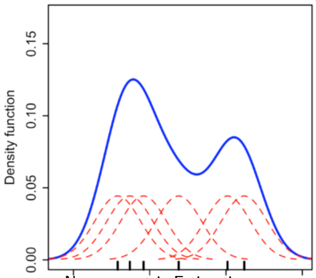
Kernel Density Estimation

Definition

Kernel density estimate has the following form:

$$\widehat{p}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

where h is called <u>*bandwidth*</u>.



Choosing Bandwidth

Particular form of the kernel function K is far less important for the 'quality' of estimation than the bandwidth h.

Theorem

$$MISE(\widehat{p}_n, p) \approx \frac{1}{4}\sigma_K^4 h^4 \int \left(p''(x)\right)^2 dx + \frac{1}{nh}\int \left(K(x)\right)^2 dx.$$

Minimum is attained for $h = h^*$:

$$h^* = \left(\frac{1}{n}\frac{\int \left(K(x)\right)^2 dx}{\left(\int x^2 K(x) dx\right)^2 \left(\int p''(x)^2 dx\right)}\right)^{\frac{1}{5}}.$$

In that case $MISE(\widehat{p}_n, p) = O\left(n^{-\frac{4}{5}}\right)$.

Choosing Bandwidth

Proof:

Apply bias-variance decomposition:

- $bias(x) = \mathbb{E}_p \widehat{p}_n(x) - p(x) = \int \Big(\frac{1}{nh} \sum\limits_{i=1}^{n} K(\frac{x - x_i}{h}) \Big) p(x_1) \dots p(x_n) dx_1 \dots dx_n - \frac{1}{n} \sum\limits_{i=1}^{n} \int K(z) p(x) dz \approx \int K(z) \Big[-p'(x) z h + p''(x) \frac{(zh)^2}{2} \Big] dz = \frac{1}{2} \sigma_K^2 h^2 p''(x).$

- $\int bias^2(x) dx = \frac{1}{4} \sigma_K^4 h^4 \int [p''(x)]^2 dx.$

Choosing Bandwidth

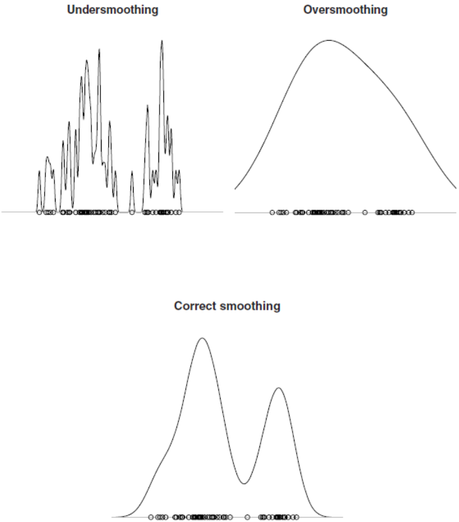
Proof (continued):

- $\int \mathbb{V}_p \widehat{p}_n(x)dx = \int \mathbb{V}_p \Big[\frac{1}{nh}\sum\limits_{i=1}^{n} K(\frac{x-x_i}{h})dx\Big] = \frac{1}{(nh)^2}\sum\limits_{i=1}^{n}\int \mathbb{V}_p K(\frac{x-x_i}{h})dx \leq$
  $\frac{1}{(nh)^2}\sum\limits_{i=1}^{n}\int \mathbb{E}_p K(\frac{x-x_i}{h})^2 dx = \frac{1}{(nh)^2}\sum\limits_{i=1}^{n}\int\int K(\frac{x-x_i}{h})^2 p(x_i)dx_i dx =$
  $\frac{1}{(nh)^2}\sum\limits_{i=1}^{n}\int p(x_i)\int K(\frac{x-x_i}{h})^2 dx dx_i = \frac{1}{(nh)^2}\sum\limits_{i=1}^{n}\int p(x_i)dx_i h\int K^2(z)dz =$
  $\frac{1}{nh}\int K^2(z)dz.$

Choosing Bandwidth

- Minimum of $MISE(\widehat{p}_n, p)$ is attained at some value h^*.
- Pluging h^* into $\widehat{p}_n$, we get that $MISE = O(n^{-\frac{4}{5}})$, i.e. convergence is better for KDE than for the histogram.
- It can be shown that under very general conditions it is not possible to get convergence better than $n^{\frac{4}{5}}$.
- As with the histogram, large h can lead to oversmoothing and small – to undersmoothing due to bias-variance tradeoff.

Choosing Bandwidth



Multiple Dimensions

Now let the observations be multidimensional, i.e. i-th observation is a d-dimesional vector:

$$X_i = \left[X_i^1, \dots X_i^d\right]^T.$$

Let $h = [h_1, \dots, h_d]^T$ be a vector of bandwidth values for each dimension. Then:

$$\widehat{p}_n(x) = \frac{1}{nh_1 \cdot \ldots \cdot h_d} \sum_{i=1}^{n} \left[\prod_{j=1}^{d} K\left(\frac{x_j - X_i^j}{h_j} \right) \right],$$

where $x = [x_1, \dots, x_d]^T$ is an arbitrary point in $\mathbb{R}^d$.

Multiple Dimensions

- For this estimate risk $MISE(\widehat{p}_n, p) \approx$
$\frac{1}{4}\sigma_K^4 \left[\sum_{j=1}^{d} h_j^4 \int p_{jj}^2(x)dx + \sum_{j \neq k} h_j^2 h_k^2 \int p_{jj}(x) p_{kk}(x)dx\right] + \frac{\left(\int K^2(x)dx\right)^d}{nh_1 \cdot \ldots \cdot h_d}$,
where $p_{jj}(x) = \frac{\partial^2 p(x)}{\partial x_j^2}$.

- Optimal bandwidth $h_i^* \approx cn^{-\frac{1}{4+d}}$.

- In that case risk has asymptotic $MISE(\widehat{p}_n, p) = O(n^{-\frac{4}{4+d}})$.

Curse of Dimensionality

Optimal risk scales as $O\left(n^{-\frac{4}{4+d}}\right)$, i.e. we observe so-called "curse of dimensionality": as d grows, rate of convergence to the true density decreases rapidly.

The following table contains samples sizes required to achive mean squared error at zero less than 0.1 with a KDE vs the dimensionality of data (assuming optimal bandwidth):

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

where d is the dimensionality and n is the sample size.

Nonparametric Density Estimation

- Problem Statement
- Histograms for Density Estimation
- Kernel Density Estimation

Nonparametric Regression

- Nadaraya-Watson Estimator
- Confidence Band for Regression Function

Nonparametric Regression

Consider a sample of n observations: $(X_1, Y_1), \ldots, (X_n, Y_n)$, generated from the joint density $p(x, y)$.

Observations follow the following relationship:

$$Y_i = r(X_i) + \varepsilon_i, \quad \varepsilon_i - i.i.d, \quad \mathbb{E}\varepsilon_i = 0, \quad \mathbb{V}\varepsilon_i = \sigma^2.$$

The task is to estimate the regression function:

$$r(x) = \mathbb{E}(Y \mid X = x) = \int y\, p(y \mid x) dy = \frac{\int y\, p(x, y) dy}{\int p(x, y) dy} = \frac{\int y\, p(x, y) dy}{p(x)}.$$

Skoltech
Skolkovo Institute of Science and Technology

Nonparametric Regression

Definition

Let $\widehat{p}_n(x)$ and $\widehat{p}_n(x, y)$ be the KDEs of the density build on samples $\{X_1, \dots, X_n\}$ and $\{(X_1, Y_1) \dots, (X_n, Y_n)\}$ respectively with kernel K. If $\widehat{p}_n(x) \neq 0$, then

$$\widehat{r}_n^{NW}(x) = \frac{\int y \, \widehat{p}_n(x, y) dy}{\widehat{p}_n(x)}.$$

Note that such estimator can be applied even when X_i are fixed and deterministic, e.g. $X_i = \frac{i}{n}$.

Nonparametric Regression

To estimate $r(x)$ Nadaraya-Watson estimator is used:

Definition

Nadaraya-Watson estimator:

$$\widehat{r}_n^{NW}(x) = \sum_{i=1}^{n} w_i(x) Y_i,$$

where $w_i(x) = \dfrac{K\left(\frac{x-X_i}{h}\right)}{\sum\limits_{j=1}^{n} K\left(\frac{x-X_j}{h}\right)}$, and K is a kernel function.

Hence, it is a weighted sum of Y_is, where points close to x have higher weight.

Nonparametric Regression

Now lets consider the risk and choose bandwidth.

Theorem

$$MISE(\widehat{r}_n^{NW}, r) \approx \frac{h^4}{4}\left(\int x^2 K^2(x)dx\right)^4 \int \left(r''(x) + 2r'(x)\frac{p'(x)}{p(x)}\right)^2 dx + \frac{1}{h}\int \frac{\sigma^2 \int K^2(x)dx}{np(x)}dx.$$

- Optimal bandwidth $h^* = cn^{-\frac{1}{5}}$.
- In that case risk scales as $MISE(\widehat{r}_n^{NW}, r) = O(n^{-\frac{4}{5}})$.

Nonparametric Regression

As before, h^* can not be obtained, since it depends on unknown $r(x)$ and $p(x)$. So again we optimize an estimate of the risk over h:

$$\widehat{\mathcal{J}}(h) = \sum_{i=1}^{n} \left(Y_i - \widehat{r}^{NW}_{(-i)}(X_i)\right)^2,$$

where $\widehat{r}^{NW}_{(-i)}$ is Nadaraya-Watson estimate using the sample with observation (X_i, Y_i) removed.

Theorem

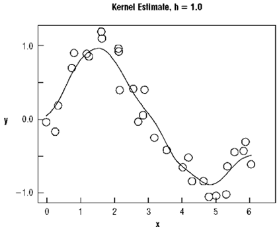
$$\widehat{\mathcal{J}}(h) = \sum_{i=1}^{n} \left(Y_i - \widehat{r}^{NW}(X_i)\right)^2 \frac{1}{\left(1 - \frac{K(0)}{\sum\limits_{j=1}^{n} K\left(\frac{X_i - X_j}{h}\right)}\right)^2}.$$

Nonparametric Regression

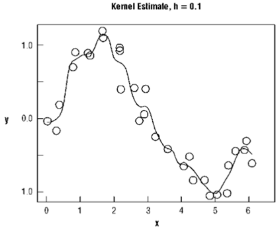
Similar to histogram and KDE, we can observe the bias-variance tradeoff:

- large h produces oversmoothing – too many fine details are removed,
- for small h we have undersmoothing – the estimate is adapted to the noise.

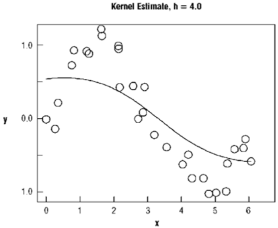
Nonparametric Regression



Correct smoothing



Undersmoothing



Oversmoothing

Skoltech
Skolkovo Institute of Science and Technology

Confidence Band for Regression Function

- First estimate σ^2.
- Let X_i be in increasing order.
- Assume that $r(x)$ is a smooth function, we get $r(X_{i+1}) - r(X_i) \approx 0$.

Then:

$$Y_{i+1} - Y_i = [r(X_{i+1}) + \varepsilon_{i+1}] - [r(X_i) + \varepsilon_i] \approx \varepsilon_{i+1} - \varepsilon_i.$$

$$\mathbb{V}(Y_{i+1} - Y_i) \approx \mathbb{V}(\varepsilon_{i+1} - \varepsilon_i) = \mathbb{V}\varepsilon_{i+1} + \mathbb{V}\varepsilon_i = 2\sigma^2.$$

$$\Rightarrow \widehat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2.$$

Lets construct confidence band for the smoothed version $\bar{r}_n(x) = \mathbb{E}\widehat{r}_n^{NW}(x)$ of the true regression function $r(x)$.

Confidence Band for Regression Function

An approximate $(1-\alpha)$ confidence interval for $\bar{r}_n(X)$ is given by:

$$r_-(x) = \widehat{r}_n^{NW}(x) - z_\alpha \widehat{\sigma} \sqrt{\sum_{i=1}^{n} w_i^2(x)},$$

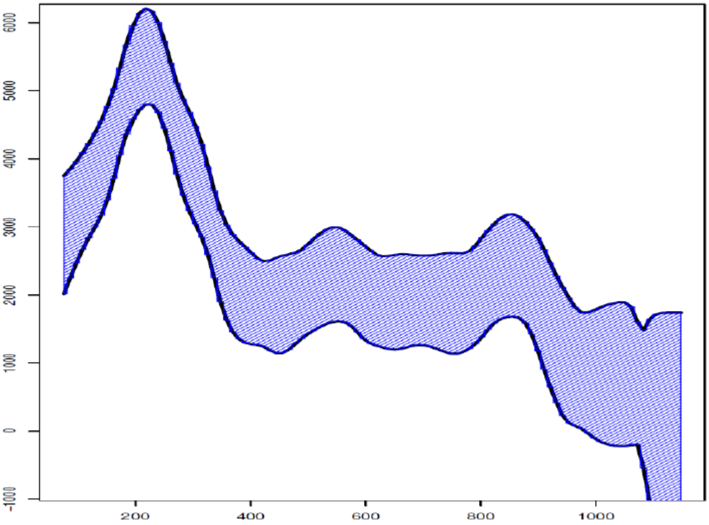
$$r_+(x) = \widehat{r}_n^{NW}(x) + z_\alpha \widehat{\sigma} \sqrt{\sum_{i=1}^{n} w_i^2(x)},$$

where $\widehat{\sigma}, w_i$ were defined earlier.

$z_\alpha = \Phi^{-1}\left(\frac{1+(1-\alpha)^{\frac{h}{b-a}}}{2}\right)$, where Φ is CDF of the standard normal distribution,

h is bandwidth, $X_1, \ldots, X_n \in (a; b)$.

Confidence Band for Regression Function



Skoltech
Skolkovo Institute of Science and Technology

Confidence Band for Regression Function

- The constructed confidence band, as was the case with the histogram and KDE, is not a confidence band for the regression, but works for the smoothed version.
- E.g., confidence interval for the density in the case of KDE is actually a confidence interval for a function given by smoothing the true density with the same kernel.
- Obtaining confidence interval for the density itself is difficult for the following reason.
- Let $\widehat{f}_n(x)$ be an estimate of the function $f(x)$.
- Denote $\bar{f}_n(x) = \mathbb{E}\widehat{f}_n(x),\ s_n(x) = \sqrt{\mathbb{V}\widehat{f}_n(x)}$, then

$$\frac{\widehat{f}_n(x) - f_n(x)}{s_n(x)} = \frac{\widehat{f}_n(x) - \bar{f}_n(x)}{s_n(x)} + \frac{\bar{f}_n(x) - f_n(x)}{s_n(x)}.$$

Confidence Band for Regression Function

$$\frac{\widehat{f}_n(x) - f_n(x)}{s_n(x)} = \frac{\widehat{f}_n(x) - \bar{f}_n(x)}{s_n(x)} + \frac{\bar{f}_n(x) - f_n(x)}{s_n(x)}.$$

- Usually, according to CLT, the first summand converges to the standard normal distribution, using which we can construct the confidence interval.
- Second summand equals the ration of bias to standard deviation.
- In the case of parametric estimation, bias is usually smaller than standard deviation, i.e. the second summand approaches zero with increasing sample size.
- In nonparametrics, smoothing leads to "balancing" bias and standard deviation.
- In that case the second summand may not be close to zero even for large sample sizes, so the confidence interval will not be centered around the true density.

Thank you for your attention!