Яндекс

Skoltech
Skolkovo Institute of Science and Technology

Непараметрическое оценивание

Непараметрическая оценка плотности. Потери, риск. Ядерная оценка плотности.

ПМИ ФКН ВШЭ, 20 октября 2018 г.

Денис Деркач[1]

[1] ФКН ВШЭ

Оглавление

Непараметрическая оценка плотности

Постановка задачи

Пусть задана выборка: $X_1, \ldots, X_n \sim F$.
F - абсолютно непрерывная функция распределения с неизвестной плотностью p. Необходимо оценить p в точке x, т.е. построить $\hat{p}_n(x) = \hat{p}_n(x; X_1, \ldots, X_n)$.

NB: Ранее в аналогичных задачах мы искали
$p \in \{p(x;\theta), \theta \in \Theta\}, \Theta \subset \mathbb{R}^n$, где есть зависимость от параметров.

MSE, функция риска

Пусть в точке x_0 построена оценка $\hat{p}_n(x_0)$ плотности. Рассматривая квадратичную функцию потерь, приходим к следующему понятию.

Определение

Mean Square Error:

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

MISE

Если же построена оценка $\hat{p}_n(x)\,\forall x \in \mathbb{R}$, то

Определение

Mean Integrated Squared Error:

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

Bias

Определение

Смещение (bias)

$bias(x_0) = \mathbb{E}_p \hat{p}_n(x_0) - p(x_0)$

Разложение ошибки

Лемма

$$MSE(\hat{p}_n, p, x_0) = bias^2(x_0) + \mathrm{Var_p\hat{p}_n(x_0)} =$$
$$= [\mathbb{E}_p\hat{p}_n(x_0) - p(x_0)]^2 + \mathbb{E}_p[\hat{p}_n(x_0) - \mathbb{E}_p\hat{p}_n(x_0)]^2$$

Лемма

$$MISE(\hat{p}_n, p) = \int\limits_{\mathbb{R}} bias^2(x)dx + \int\limits_{\mathbb{R}} \mathrm{Var_p\hat{p}_n(x)dx}$$

Гистограмма

Простейший способ оценить плотность - построить гистограмму

Возьмём интервал $[a, b) \ni X_1, \dots X_n$

Поделим его на M равных частей Δ_i размера $h = \frac{b-a}{M}$:

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

Пусть ν_i - число элементов выборки, попавших в Δ_i;

Определение

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

При $x \in \Delta_i$ и малом h: $\mathbb{E}_p \hat{p}_n(x) = \frac{\mathbb{E}\nu_j}{nh} = \frac{\int_{\Delta_j} p(u)du}{h} \approx \frac{p(x)h}{h} = p(x)$

Непараметрическая оценка плотности

- Постановка задачи
- Потери, риск
- Определение параметра сглаживания
- Доверительная трубка для плотности
- Ядерная оценка плотности
  - Выбор ширины ядра
  - Доверительный интервал для усредненной плотности
  - Многомерный случай

Непараметрическая регрессия

- Доверительная трубка для функции регрессии
- Многомерный случай

Гистограмма: определение параметра сглаживания

Рассмотрим выбор h - параметра сглаживания

Проведём вычисления для $x_0 \in \Delta_j$:

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

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

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

Определение параметра сглаживания

$$\int_a^b bias^2(x_0)dx_0 = \sum_{j=0}^{N-1} \int_{\Delta_j} bias^2(x_0)dx_0 =$$

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

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

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

$$\approx \left(\int_a^b [p'(x)]^2 dx \right) \frac{h^2}{12}.$$

Определение параметра сглаживания

$$\mathbb{V}ar_p\hat{p}_n(x_0) = \mathbb{V}ar_p\frac{\nu_j}{nh} = \frac{1}{(nh)^2}\mathbb{V}ar_p\nu_j =$$

$$= \frac{1}{(nh)^2}n\int\limits_{\Delta_j} p(x)dx(1 - \int\limits_{\Delta_j} p(x)dx) \approx \frac{1}{nh^2}\int\limits_{\Delta_j} p(x)dx$$

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

$$= \frac{1}{nh}\int\limits_a^b p(x)dx = \frac{1}{nh}$$

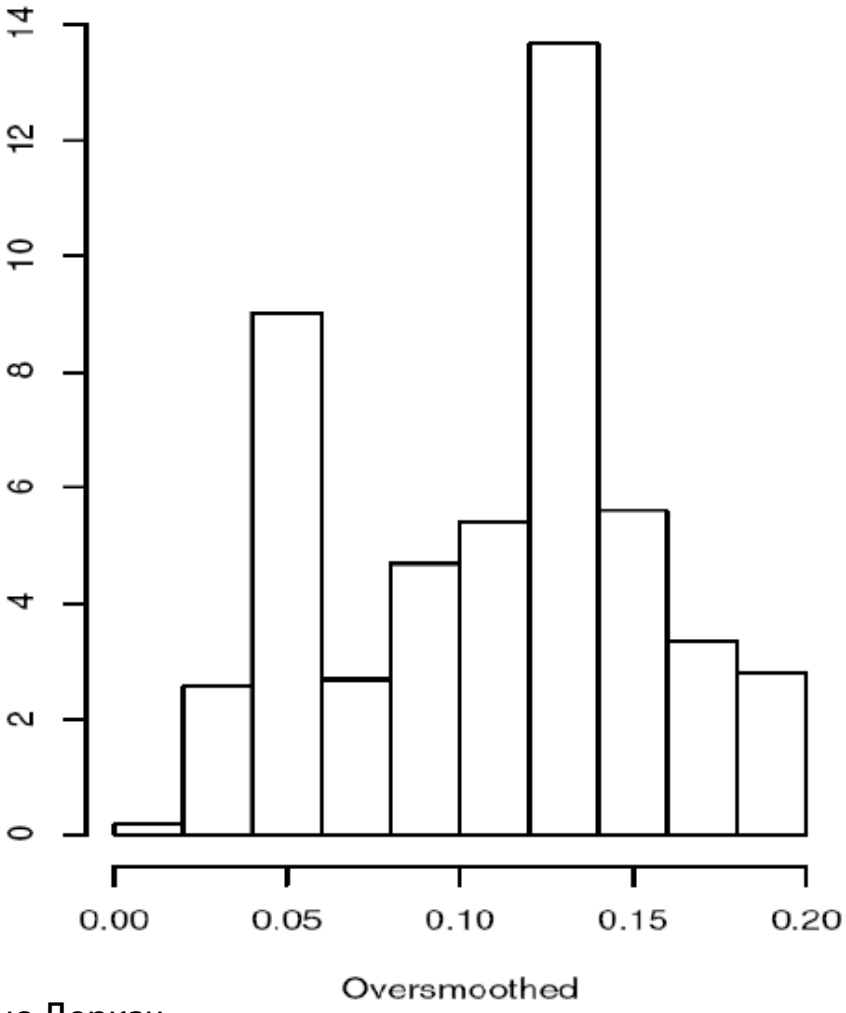
Определение параметра сглаживания

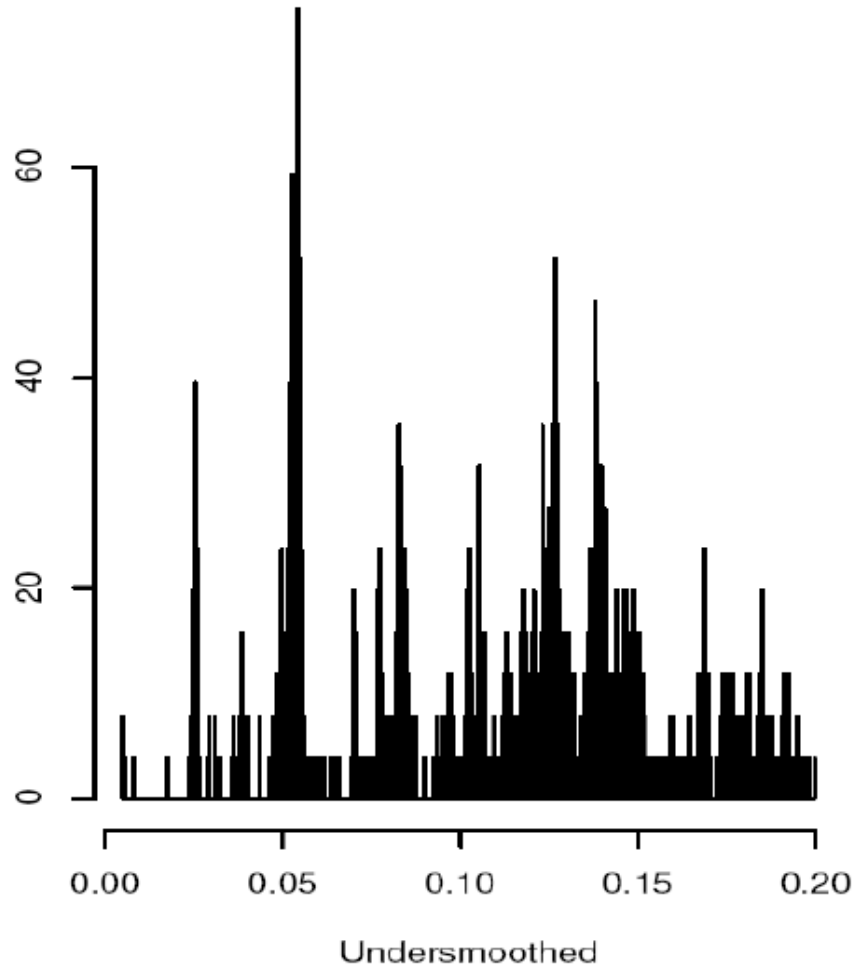
Таким образом,

$$MISE(\hat{p}_n, p) = \left(\int_{\mathbb{R}} [p'(x)]^2 dx \right) \frac{h^2}{12} + \frac{1}{nh}$$

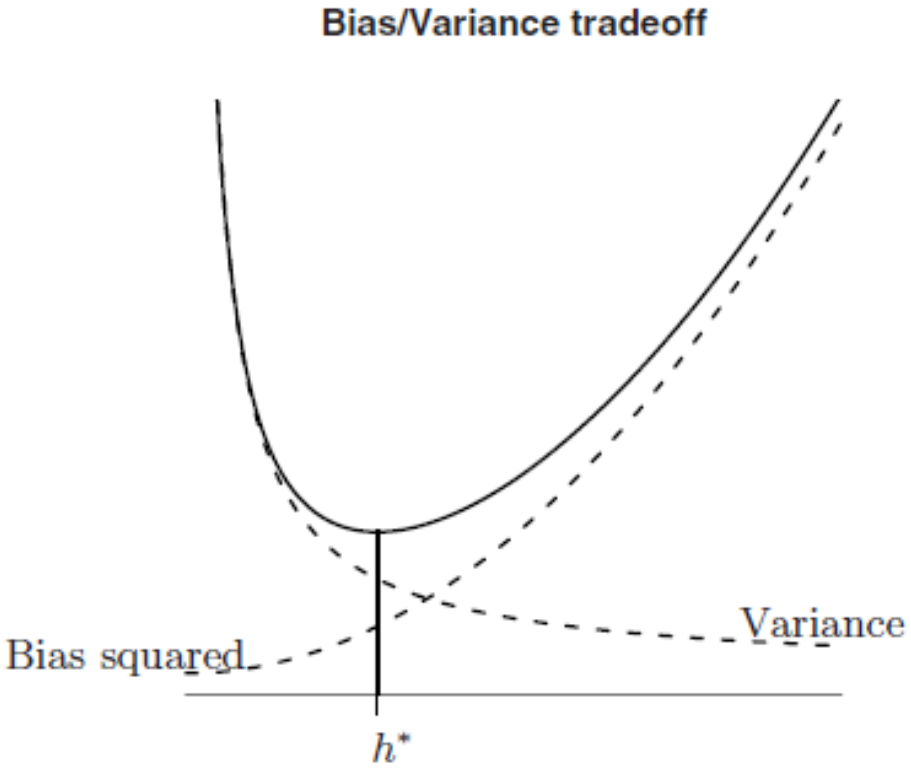
Чем больше h, тем больше смещение и меньше дисперсия, и наоборот. Это называется bias-variance tradeoff.
Ситуации с большим h - oversmoothing, с маленьким - undersmoothing.

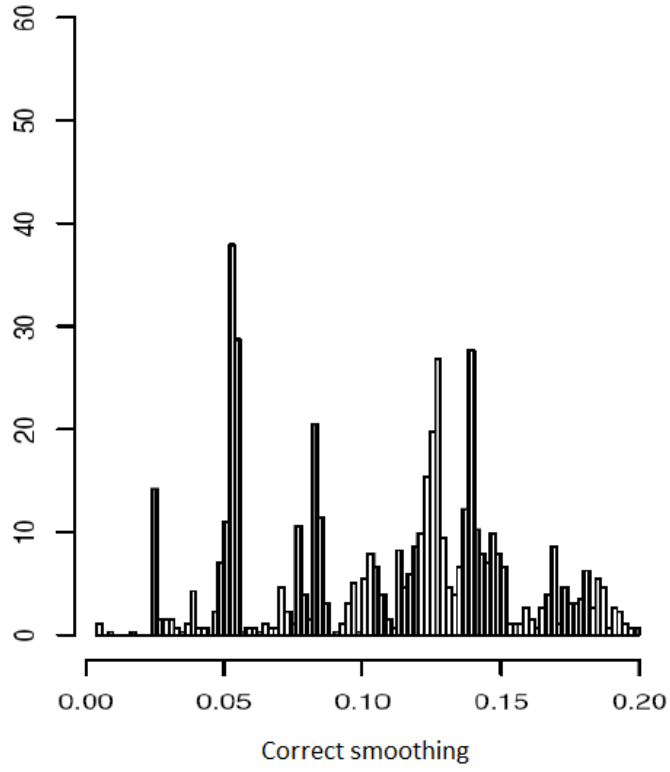
Пример неоптимального сглаживания





Определение параметра сглаживания





Определение параметра сглаживания

Значение h, при котором $MISE$ минимальный

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

При этом

$$MISE(\hat{p}_n, p) \approx \frac{C}{n^{\frac{2}{3}}}, \text{ где } C = \left(\frac{3}{4}\right)^{\frac{2}{3}} \left(\int\limits_{\mathbb{R}} [p'(x)]^2 dx \right)^{\frac{1}{3}}.$$

Таким образом, при использовании гистограммы с оптимальным h, $MISE$ убывает как $n^{-\frac{2}{3}}$:

Определение параметров сглаживания

На практике h^* нельзя вычислить, так как h^* зависит от неизвестной истинной плотности.
Поэтому оценим $MISE$ и минимизируем по h оценку.
Так как:

$$\int\limits_{\mathbb{R}} (\hat{p}_n(x) - p(x))^2 dx = \int\limits_{\mathbb{R}} \hat{p}_n(x)^2 dx - 2 \int\limits_{\mathbb{R}} \hat{p}_n(x) p(x) dx + \int\limits_{\mathbb{R}} p(x)^2 dx,$$

то достаточно оценить и минимизировать только

$$\mathcal{J}(h) = \int\limits_{\mathbb{R}} \hat{p}_n(x)^2 dx - 2 \int\limits_{\mathbb{R}} \hat{p}_n(x) p(x) dx.$$

Определение параметра сглаживания

Определение

Оценка риска с помощью кросс-валидации:

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

где $\hat{p}_{(-i)}$ - оценка гистограммы по выборке без $i-$ого наблюдения.

Теорема

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

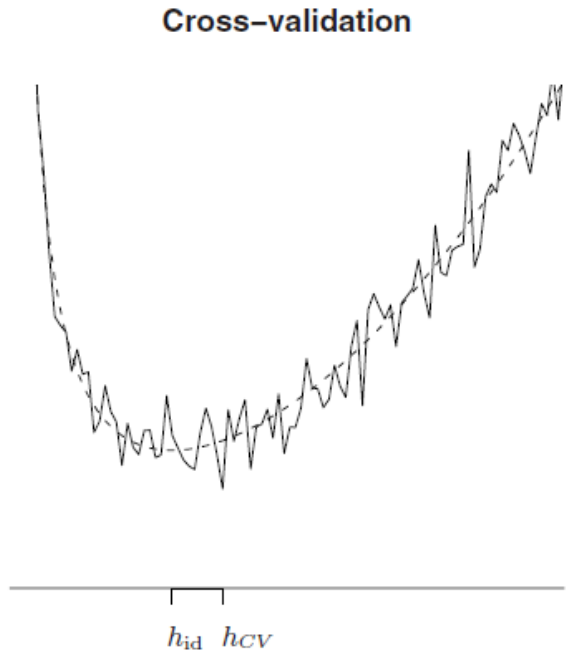
Теорема

Для гистограм оценка функции риска:

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

Определение параметра сглаживания

Типичное поведение $\hat{\mathcal{J}}(h)$ имеет вид:



Таким образом, вместо неизвестного $MISE$ можно минимизировать $\hat{\mathcal{J}}(h)$ и найти оптмимальное h_{cv}, которое будет недалеко от $h_{id} = h^*$.

Непараметрическая оценка плотности

- Постановка задачи
- Потери, риск
- Определение параметра сглаживания
- Доверительная трубка для плотности
- Ядерная оценка плотности
  - Выбор ширины ядра
  - Доверительный интервал для усредненной плотности
  - Многомерный случай

Непараметрическая регрессия

- Доверительная трубка для функции регрессии
- Многомерный случай

Доверительная трубка для плотности

Пусть необходимо построить доверительные интервалы для p.
Для этого будем использовать гистограмму $\hat{p}_n(x)$, определенную ранее.
Определим

$\overline{p_n}(x) = \mathbb{E}\hat{p}_n(x) = \frac{\int_{\Delta_j} p(u)du}{h}$ для $x \in \Delta_j$.

По сути, $\overline{p_n}$ - “гистограммное” усреднение плотности p.

Определение

Пара функций $(p_-(x), p_+(x))$ является $1-\alpha$ доверительной областью (трубкой), если для любого x:

$$\mathbb{P}_p\left(p_-(x) \le \overline{p_n}(x) \le p_+(x)\right) \ge 1-\alpha$$

Доверительная трубка для плотности

Теорема

Пусть $M = M(n)$ - число ячеек в гистограмме $\hat{p}_n$, причем $M(n) \to \infty$ и $\frac{M(n)\log(n)}{n} \to \infty$ при $n \to \infty$.

Определим

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

где $C = \frac{1}{2} z_{\frac{\alpha}{2M}} \sqrt{\frac{M}{n(b-a)}}$

Тогда $(p_-(x), p_+(x))$ является $1-\alpha$ доверительным интервалом.

Доверительная трубка для плотности

Из центральной предельной теоремы

$$\frac{\nu_j}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{x_i \in \Delta_j\} \sim \mathcal{N}\left(\int\limits_{\Delta_j} p(x)dx, \frac{\int\limits_{\Delta_j} p(x)dx(1 - \int\limits_{\Delta_j} p(x)dx)}{n}\right)$$

Согласно дельта-методу $\sqrt{\frac{\nu_j}{n}} \sim \mathcal{N}\left(\sqrt{\int\limits_{\Delta_j} p(x)dx}, \frac{1}{4n}\right)$. Более того, можно показать, что $\sqrt{\frac{\nu_j}{n}}$ приблизительно независимы. Тогда $2\sqrt{n}\left(\sqrt{\frac{\nu_j}{n}} - \sqrt{\int\limits_{\Delta_j} p(x)dx}\right) \approx \xi_j$, где $\xi_0, \ldots, \xi_{M-1} \sim \mathcal{N}(0, 1)$.

Доверительная трубка

$$
\begin{aligned}
A &= \{p_-(x) \leq \overline{p_n}(x) \leq p_+(x) \forall x\} = \\
&= \{\sqrt{p_-(x)} - c \leq \sqrt{\overline{p_n}(x)} \leq \sqrt{p_+(x)} + c \forall x\} = \\
&= \{\max_x |\sqrt{\hat{p}(x)} - \sqrt{\overline{p_n}(x)}| \leq c\}
\end{aligned}
$$

Доверительная трубка для плотности

Тогда $\mathbb{P}(A^c) = \mathbb{P}\{\max\limits_x |\sqrt{\hat{p}_n(x)} - \sqrt{\overline{p_n}(x)}| > c\} =$

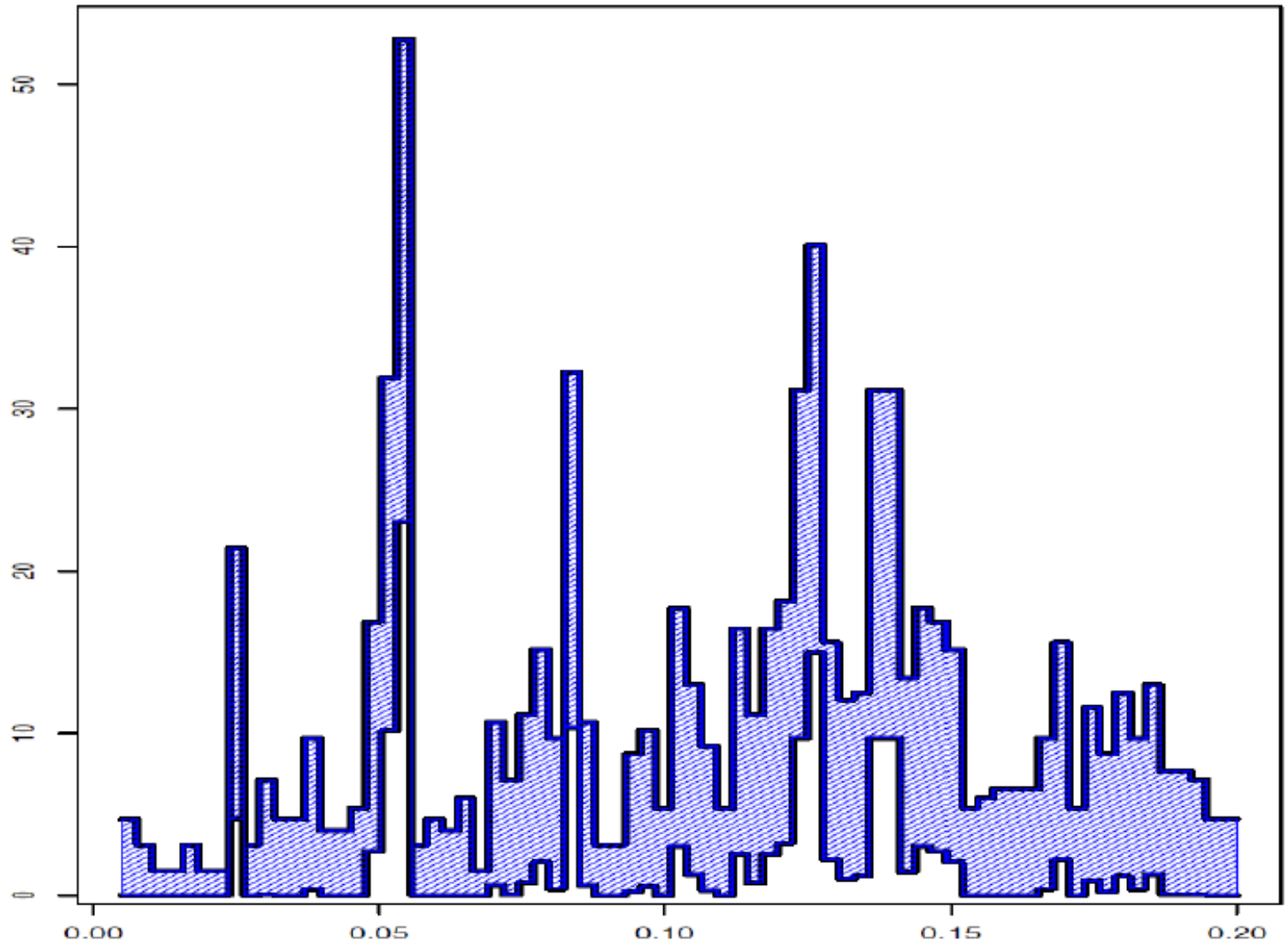
$$\mathbb{P}\left\{\max_{j=\overline{0,M-1}} \left|\sqrt{\frac{\nu_j}{nh}} - \sqrt{\frac{\int\limits_{\Delta_j} p(x)dxh}{h}}\right| > c\right\} \approx$$

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

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

т.е. для предъявленных $p_-(x), p_+(x)$ выполнено определение доверительной трубки.

Доверительная трубка для плотности



Комментарии о доверительных трубках

- Важным условием для предыдущего вывода является наличие большого количества семплов n. В случае малого количества семплов ситуация может отличаться, в зависисости от использованного метода оценки.
- Разные отрасли используют разные определения ширины доверительной трубки (от 68% до 100%).

Непараметрическая оценка плотности

- Постановка задачи
- Потери, риск
- Определение параметра сглаживания
- Доверительная трубка для плотности
- Ядерная оценка плотности
  - Выбор ширины ядра
  - Доверительный интервал для усредненной плотности
  - Многомерный случай

Непараметрическая регрессия

- Доверительная трубка для функции регрессии
- Многомерный случай

Ядерная оценка плотности

Позволяет получить более гладкие по сравнению с гистограммной оценки, быстрее сходящиеся к плотности.

Определение

Ядро - функция K такая, что

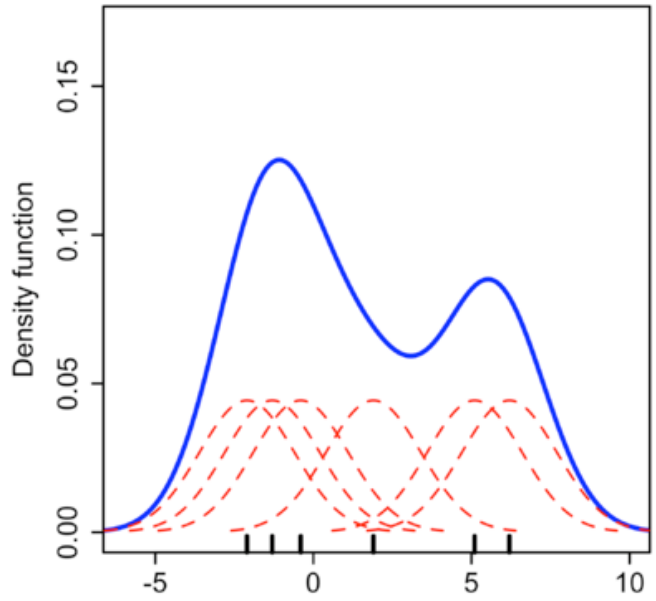
$$K(x) \geq 0, \int\limits_{\mathbb{R}} K(x)dx = 1, \int\limits_{\mathbb{R}} xK(x)dx = 0, \sigma_K^2 \equiv \int\limits_{\mathbb{R}} x^2 K(x)dx$$

Ядерная оценка плотности

Определение

Ядерная оценка плотности имеет вид:

$\hat{p}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$, h — ширина ядра



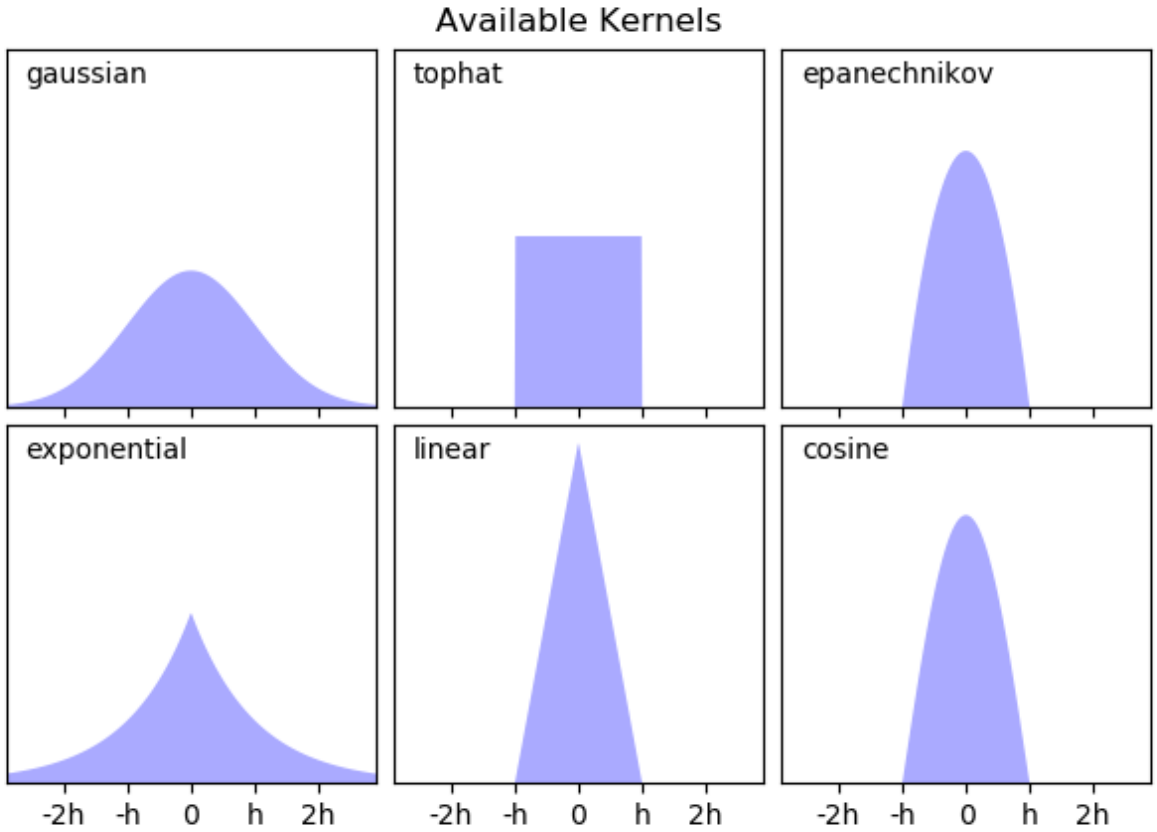
Виды ядер

Examples

- $K(x) = \frac{1}{2}\mathbb{I}\{|x| < 1\}$ — прямоугольное ядро
- $K(x) = (1 - |x|)\mathbb{I}\{|x| < 1\}$ — треугольное ядро
- $K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ — Гауссовское ядро
- $K(x) = \frac{3}{4}(1 - x^2)\mathbb{I}\{|x| < 1\}$ — ядро Епанечникова

Далее мы будем рассматривать только гладкие ядра.

Примеры ядер



Вид ядерной функции K влияет на “качество” оценки не так сильно, как выбор ширины ядра h.

Picture credit:http://scikit-learn.org/stable/modules/density.html

Выбор ширины ядра

Теорема

$$MISE(\hat{p}_n, p) \approx \frac{1}{4}\sigma_K^4 h^4 \int\limits_{\mathbb{R}} (p''(x))^2 dx + \frac{1}{nh} \int\limits_{\mathbb{R}} (K(x))^2 dx$$

Минимум достигается при $h = h^*$:

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

При этом $MISE(\hat{p}_n, p) = O\left(n^{-\frac{4}{5}}\right)$

Выбор ширины ядра

Воспользуемся bias-variance decomposition:

$$bias(x) = \mathbb{E}_p \hat{p}_n(x) - p(x) =$$

$$\int\limits_{\mathbb{R}} (\frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - x_i}{h})) p(x_1) \dots p(x_n) dx_1 \dots dx_n -$$

$$\frac{1}{n} \sum_{i=1}^{n} \int\limits_{\mathbb{R}} K(z) p(x) dz \approx \int\limits_{\mathbb{R}} K(z)[-p'(x)zh + p''(x)\frac{(zh)^2}{2}]dz =$$

$$\frac{1}{2}\sigma_K^2 h^2 p''(x)$$

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

Выбор ширины ядра

$$\int\limits_{\mathbb{R}} \mathbb{V}ar_p \hat{p}_n(x)dx = \int\limits_{\mathbb{R}} \mathbb{V}ar_p[\frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-x_i}{h})dx =$$

$$\frac{1}{(nh)^2}\sum_{i=1}^{n}\int\limits_{\mathbb{R}} \mathbb{V}ar_p K(\frac{x-x_i}{h})dx \leq \frac{1}{(nh)^2}\sum_{i=1}^{n}\int\limits_{\mathbb{R}} \mathbb{E}_p K(\frac{x-x_i}{h})^2 dx =$$

$$\frac{1}{(nh)^2}\sum_{i=1}^{n}\int\limits_{\mathbb{R}}\int\limits_{\mathbb{R}} K(\frac{x-x_i}{h})^2 p(x_i)dx_i dx =$$

$$\frac{1}{(nh)^2}\sum_{i=1}^{n}\int\limits_{\mathbb{R}} p(x_i)\int\limits_{\mathbb{R}} K(\frac{x-x_i}{h})^2 dx dx_i =$$

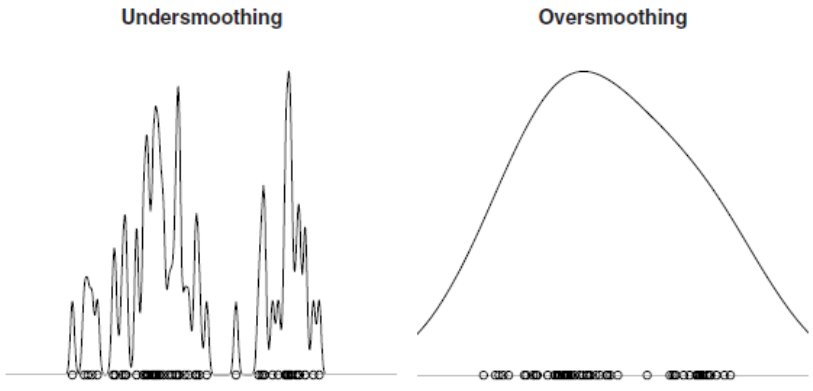
$$\frac{1}{(nh)^2}\sum_{i=1}^{n}\int\limits_{\mathbb{R}} p(x_i)dx_i h\int\limits_{\mathbb{R}} K^2(z)dz = \frac{1}{nh}\int\limits_{\mathbb{R}} K^2(z)dz$$

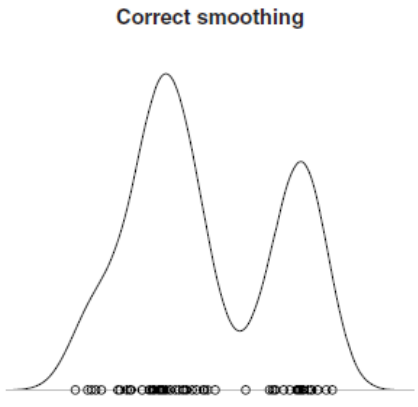
Минимум $MISE(\hat{p}_n, p)$ достигается в некотором h^*.

Выбор ширины ядра

Подставляя h^* в $\hat{p}_n$, получаем, что $MISE = O(n^{-\frac{4}{5}})$, т.е. сходимость ядерной оценки лучше, чем у гистограммы.
Можно показать, что при достаточно общих условиях нельзя получить скорость лучше, чем $n^{-\frac{4}{5}}$.

Выбор ширины ядра





Как и в случае с гистограммной, при больших h имеет место oversmoothing, а при маленьких - undersmoothing из-за bias-variance tradeoff.

Доверительный интервал

Определим $\overline{p_n}(x) = \mathbb{E}\hat{p}_n(x) = \int\limits_{\mathbb{R}} \frac{1}{h} K(\frac{x-u}{h}) p(u) du$. Допустим, что $supp(p) \subset (a, b)$.

Тогда определим $(1-\alpha)$ доверительную трубку.

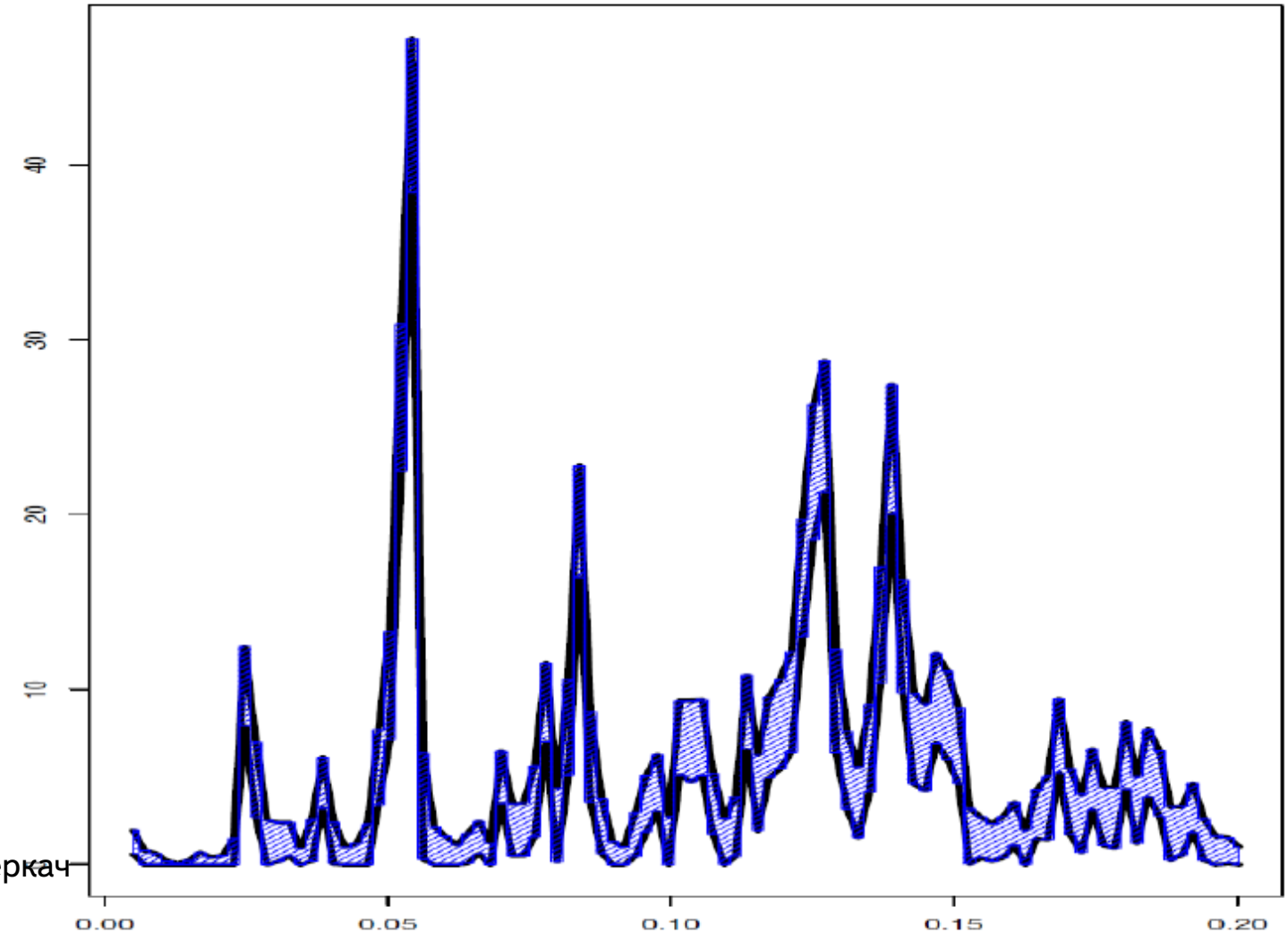
$$p_-(x) = \hat{p}_n(x) - \frac{z_\alpha}{\sqrt{n}} s(x),$$

$$p_+(x) = \hat{p}_n(x) + \frac{z_\alpha}{\sqrt{n}} s(x).$$

Где $s^2(x) = \frac{1}{n-1} \sum\limits_{i=1}^{n} [Y_i(x) - \overline{Y_n}(x)]^2$, $Y_i(x) = \frac{1}{h} K(\frac{x-X_i}{h})$,

$z_\alpha = \Phi^{-1}\left(\frac{1+(1-\alpha)^{\frac{w}{b-a}}}{2}\right)$, $\Phi-$ функция стандартного нормального распределения. w - эффективная ширина ядра.

Доверительный интервал для усредненной плотности



Ядерная оценка плотности: многомерный случай

Пусть теперь данные многомерные, то есть $i-$ое наблюдение - вектор размерности d:

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

Пусть $h = [h_1, \dots, h_d]^T$ - вектор ширины ядра вдоль каждого измерения.
Тогда:

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

где $x = [x_1, \dots, x_d]^T$ — произвольная точка в $\mathbb{R}^d$

Ядерная оценка плотности: многомерный случай

Для такой оценки риск

$$MISE(\hat{p}_n, p) \approx$$

$$\frac{1}{4}\sigma_K^4 \left[\sum_{j=1}^{d} h_j^4 \int\limits_{\mathbb{R}^d} p_{jj}^2(x)dx + \sum_{j \neq k} h_j^2 h_k^2 \int\limits_{\mathbb{R}^d} p_{jj}(x) p_{kk}^2(x)dx \right] +$$

$$\frac{\left(\int\limits_{\mathbb{R}^d} K^2(x)dx \right)^d}{nh_1 \cdot \ldots \cdot h_d},$$

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

Оптимальная ширина ядра $h_i^* \approx cn^{-\frac{1}{4+d}}$

При этом риск имеет порядок: $MISE(\hat{p}_n, p) = O(n^{-\frac{4}{4+d}})$.

Проклятие размерности

Оптимальный порядок риска $O(n^{-\frac{4}{4+d}})$, т.е. наблюдаем "проклятье размерности" - при росте d скорость сходимости к истинной плотности падает.

Рассмотрим таблицу объёмов выборки, необходимых для того, чтобы средний квадрат ошибки в нуле был меньше 0.1 в зависимости от размерности наблюдений в случае многомерной нормальной плотности и оптимальной ширины ядра:

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

где d — размерность данных, n — необходимый объём выборки.

Непараметрическая регрессия

Непараметрическая оценка плотности

- Постановка задачи
- Потери, риск
- Определение параметра сглаживания
- Доверительная трубка для плотности
- Ядерная оценка плотности
  - Выбор ширины ядра
  - Доверительный интервал для усредненной плотности
  - Многомерный случай

Непараметрическая регрессия

- Доверительная трубка для функции регрессии
- Многомерный случай

Непараметрическая регрессия

Пусть имеется n наблюдений: $(X_1, Y_1) \dots (X_n, Y_n)$,
сгенерированных из совместной плотности $p(x, y)$.
Наблюдения связаны соотношением:
$Y_i = r(X_i) + \varepsilon_i, \varepsilon_i - i.i.d, \mathbb{E}\varepsilon_i = 0, \mathbb{V}ar\varepsilon_i = \sigma^2$
Необходимо оценить функцию регрессии:

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

Непараметрическая регрессия

Определение

Пусть $\hat{p}_n(x)$ и $\hat{p}_n(x,y)-$ ядерные оценки плотностей по выборкам $\{X_1,\ldots,X_n\}$ и $\{(X_1,Y_1)\ldots,(X_n,Y_n)\}$ соответственно с ядром K. Тогда, если $\hat{p}_n(x) \neq 0$, то

$$\hat{r}_n(x) = \frac{\int\limits_{\mathbb{R}} y\hat{p}_n(x,y)dy}{\hat{p}_n(x)}$$

.

Непараметрическая регрессия

Для оценки $r(x)$ используется оценка Надарая-Ватсона:

Определение

$\hat{r}_n^{NW} = \sum_{i=1}^{n} w_i(x) Y_i$, где $w_i = \frac{K(\frac{x - X_i}{h})}{\sum_{j=1}^{n} K(\frac{x - X_j}{h})}$, K заданная ядерная функция

Таким образом,это взвешенная сумма Y_i, где точки близкие к x имеют больший вес.

NB: оценку Надарайя-Ватсона можно применять и в случае, когда X_i — фиксированные и детерминированные числа (например, $X_i = \frac{i}{n}$).

Непараметрическая регрессия

Перейдём к риску и выбору ширины ядра.

Теорема

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

Оптимальная ширина ядра: $h^* = cn^{-\frac{1}{5}}$

Порядок риска при этой ширине: $MISE(\hat{r}_n^{NW}, r) = O(n^{-\frac{4}{5}})$

Непараметрическая регрессия

Опять же, h^* нельзя выписать на практике, так как она зависит от неизвестных $r(x), p(x)$.

Поэтому минимизируют по h оценку риска

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

где $\hat{r}^{NW}_{(-i)}$ - оценка Надарайя-Ватсона, построенная по выборке, из которой удалено наблюдение (X_i, Y_i)

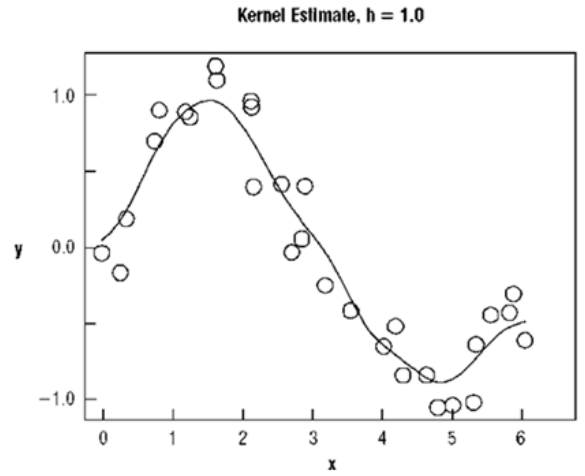
Теорема

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

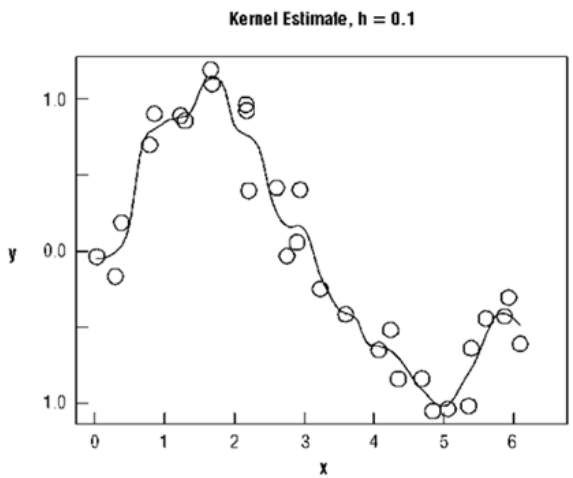
Непараметрическая регрессия

Как и в случае с гистограммной и ядерной оценкой плотности, наблюдается bias-variance tradeoff: при больших h имеет место oversmoothing - оценка слишком сглажена, а при маленьких h имеет место undersmoothing - оценка излишне подстроилась под данные.

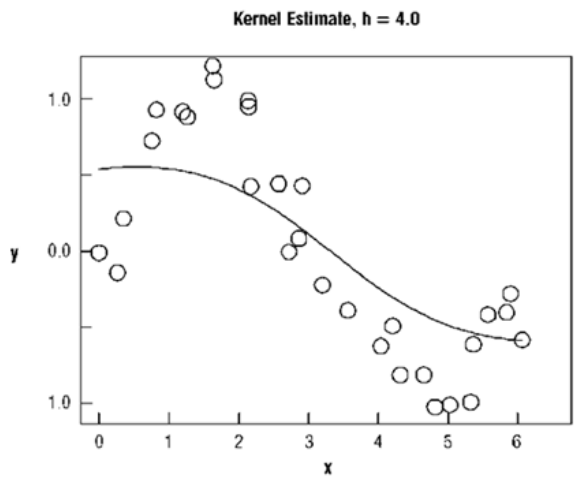
Непараметрическая регрессия



Correct smoothing



Undersmoothing



Oversmoothing

Непараметрическая оценка плотности

- Постановка задачи
- Потери, риск
- Определение параметра сглаживания
- Доверительная трубка для плотности
- Ядерная оценка плотности
  - Выбор ширины ядра
  - Доверительный интервал для усредненной плотности
  - Многомерный случай

Непараметрическая регрессия

- Доверительная трубка для функции регрессии
- Многомерный случай

Доверительная трубка для функции регрессии

Построим доверительную область. Сначала оценим σ^2. Пусть X_i упорядочены по возрастанию. Предполагая, что $r(x)$ - гладкая функция, получаем $r(X_{i+1}) - r(X_i) \approx 0$.
Тогда:

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

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

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

Будем строить доверительную область для сглаженной версии $\overline{r}_n(x) = \mathbb{E}(\hat{r}_n^{NW}(x))$ настоящей функции регрессии r.

Доверительная трубка для функции регрессии

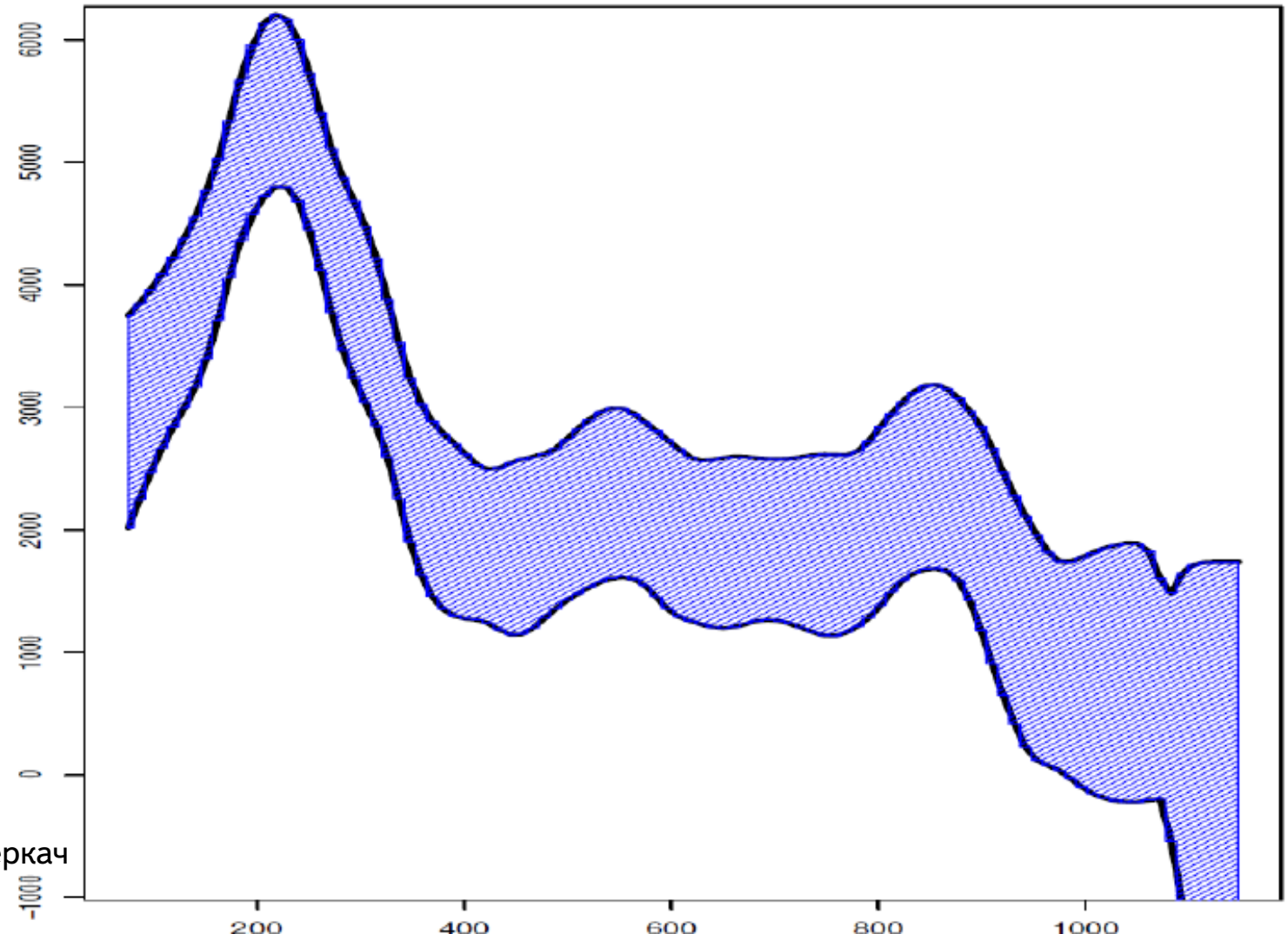
Приближенный $(1-\alpha)$ доверительный интервал для $\overline{r}_n(X)$ имеет вид:

$$r_-(x) = \hat{r}_n^{NW}(x) - C$$
$$r_+(x) = \hat{r}_n^{NW}(x) + C,$$

Где $\hat{\sigma}, w_i$ - определены выше, $C = z_\alpha \hat{\sigma} \sqrt{\sum_{i=1}^{n} w_i^2(x)}$,

$z_\alpha = \Phi^{-1}\left(\frac{1+(1-\alpha)^{\frac{w}{b-a}}}{2}\right)$, Φ — функция стандартного нормального распределения, w - эффективная ширина ядра, $X_1, \ldots, X_n \in (a; b)$

Доверительная трубка для функции регрессии



Непараметрическая оценка плотности

- Постановка задачи
- Потери, риск
- Определение параметра сглаживания
- Доверительная трубка для плотности
- Ядерная оценка плотности
  - Выбор ширины ядра
  - Доверительный интервал для усредненной плотности
  - Многомерный случай

Непараметрическая регрессия

- Доверительная трубка для функции регрессии
- Многомерный случай

Непараметрическая регрессия: многомерный случай

Если $X = [X_1, \ldots, X_n]^T$, то из-за проклятия размерности бесполезно обобщать оценку Надарайя-Ватсона аналогично способу, как мы делаем в ядерной оценке плотности.
Вместо этого можно рассмотреть аддитивную модель

- $Y = \sum_{j=1}^{d} r_j(X^j) + \alpha + \varepsilon$

  или

- $Y = \sum_{j=1}^{d} r_j(X^j) + \sum_{j<k} r_{jk}(X^j X^k) + \alpha + \varepsilon$

Непараметрическая регрессия: многомерный случай

Подготовка первой аддитивной модели:

Algorithm (Backfitting)

Инициализация : $\hat{\alpha} = \overline{Y_n}; \hat{r}_1, \dots \hat{r}_d$

Пока не стабилизируется $\hat{r}_1, \dots, \hat{r}_d$ повторять;

- Для всех $j = 1, \dots, d$:
  1. Вычислить $\widetilde{\varepsilon}_i = Y_i - \hat{\alpha} - \sum\limits_{k \neq j} \hat{r}_k(X_i^k), i = 1, \dots, n$
  2. Получить $\hat{r}_j(X^j)$ - функцию регрессии $\widetilde{\varepsilon}_i$ на j-ую компоненту X^j (то есть в качестве наблюдений имеем $\{(X_1^j, \widetilde{\varepsilon}_1), \dots, (X_n^j, \widetilde{\varepsilon}_n)\}$
  3. Положить $\hat{r}_j := \hat{r}_j - \frac{1}{n} \sum\limits_{i=1}^{n} \hat{r}_j(X_i^j)$

Доверительная трубка для регрессии

Замечание: Построенная доверительная трубка как и в случае с гистограммной и ядерной оценками плотности не является в точности доверительным интервалом для функции регрессии, но подходит для сглаженной версии. Например, доверительный интервал для плотности в случае ядерной оценки является на самом деле доверительным интервалом для функции, равной сглаженной с помощью этого же ядра истинной плотности, получить доверительный интервал для самой плотности сложно по следующим причинам.

Пусть $\hat{p}_n(x)$ - оценка плотности $p(x)$.

Обозначим $\mathbb{E}\hat{p}_n(x) = \overline{p_n}(x), \mathbb{V}ar\hat{p}_n(x) = S_n(x)$, тогда

$$\frac{\hat{p}_n(x) - p_n(x)}{S_n(x)} = \frac{\hat{p}_n(x) - \overline{p_n}(x)}{S_n(x)} + \frac{\overline{p_n}(x) - p_n(x)}{S_n(x)}$$

Доверительная трубка для регрессии

Продолжение замечания:

Обычно, по центрельной предельной теореме, первое слагаемое сходится к стандартному нормальному распределению, используя которое и строится доверительный интервал. Второе слагаемое равняется отношению смещения к стандартному отклонению. При параметрическом оценивании смещение обычно значительно меньше, чем стандартное отклонение, то есть второе слагаемое стремится к нулю при увеличении объёма выборки. В непараметрическом оценивании оптимальное сглаживание приводит к тому, что смещение и стандартное отклонение “балансируются”. В таком случае второе слагаемое может не стремиться к нулю даже при больших объёмах выборки, поэтому доверительный интервал не будет центрирован на истинную плотность.