Непараметрическое оценивание сигналов

Можно получить состоятельную оценку функции распределения без каких-либо предположений о том, какому классу она принадлежит

В случае оценки плотности распределения $f(x)$ **или функции регрессии**
1 $r(x) = \mathbb{E}(Y|X = x)$ **это не так – необходимы различные предположения о гладкости оцениваемых функций**

Гистограмма – основной параметр – количество ячеек – параметр сглаживания. Если ячеек мало – гистограмма сильно “сглажена”, то есть является сильно смещенной оценкой плотности, в противном случае она эта оценка имеет большую дисперсию

Выбор значения сглаживающего параметра – основная задача непараметрического оценивания

а) Выбор оптимального соотношения между смещением и дисперсией

Пусть g - неизвестная функция (плотность или функция регрессии), $\widehat{g}_n$ - ее оценка, полученная на основе выборки

$$\widehat{g}(x)$$

2

функция, построенная на основе данных

точка, в которой необходимо оценить значение функции $\widehat{g}(\cdot)$

$\widehat{g}_n(x)$ - случайная функция, оцененная для детерминированного значения абциссы

Интегральный квадрат ошибки (ISE – Integrated Square Error)

$$L(g, \widehat{g}_n) = \int (g(u) - \widehat{g}_n(u))^2 \, du$$

Риск или средний интегральный квадрат ошибки (MISE – Mean Integrated Square Error)

$$R(f, \widehat{f}) = \mathbb{E}\Big(L(g, \widehat{g}) \Big)$$

3

<u>Лемма</u>. Функция риска моет быть записана в виде

$$R(g, \widehat{g}_n) = \int b^2(x) \, dx + \int v(x) \, dx$$ **,**

где

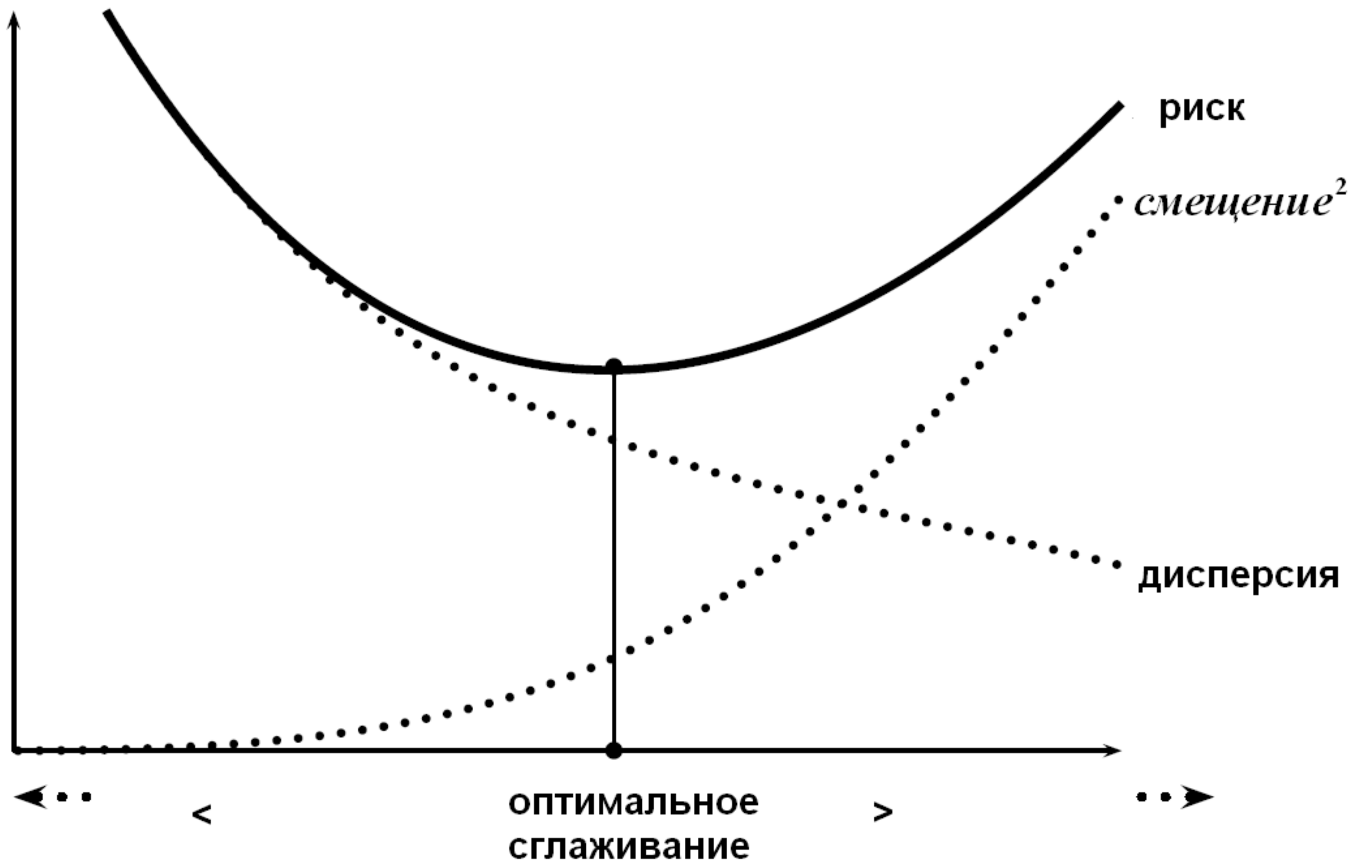
$$b(x) = \mathbb{E}(\widehat{g}_n(x)) - g(x)$$

равно смещению $\widehat{g}_n(x)$ **при фиксированном** x **и**

$$v(x) = \mathbb{V}(\widehat{g}_n(x)) = \mathbb{E}\Bigg(\big(\widehat{g}_n(x) - \mathbb{E}(\widehat{g}_n(x)\big)^2\big) \Bigg)$$

равно дисперсии $\widehat{g}_n(x)$ **при фиксированном** x

Риск = смещение² + дисперсия



b) Гистограммы

Пусть $X_1, \ldots, X_n$ **- i.i.d. выборка из отрезка** $[0, 1]$**, плотность распределения элементов которой равна** f**. Для заданного** m **определим ячейки**

$$B_1 = \left[0, \frac{1}{m}\right), B_2 = \left[\frac{1}{m}, \frac{2}{m}\right), \ \ldots, \ B_m = \left[\frac{m-1}{m}, 1\right]$$

Положим ширину равной $h = 1/m$**. Пусть** ν_j **- количество элементов выборки, попавших в отрезок** B_j**,** $\widehat{p}_j = \nu_j / n$**,** $p_j = \int_{B_j} f(u)du$

Гистограмма (оценка плотности) равна

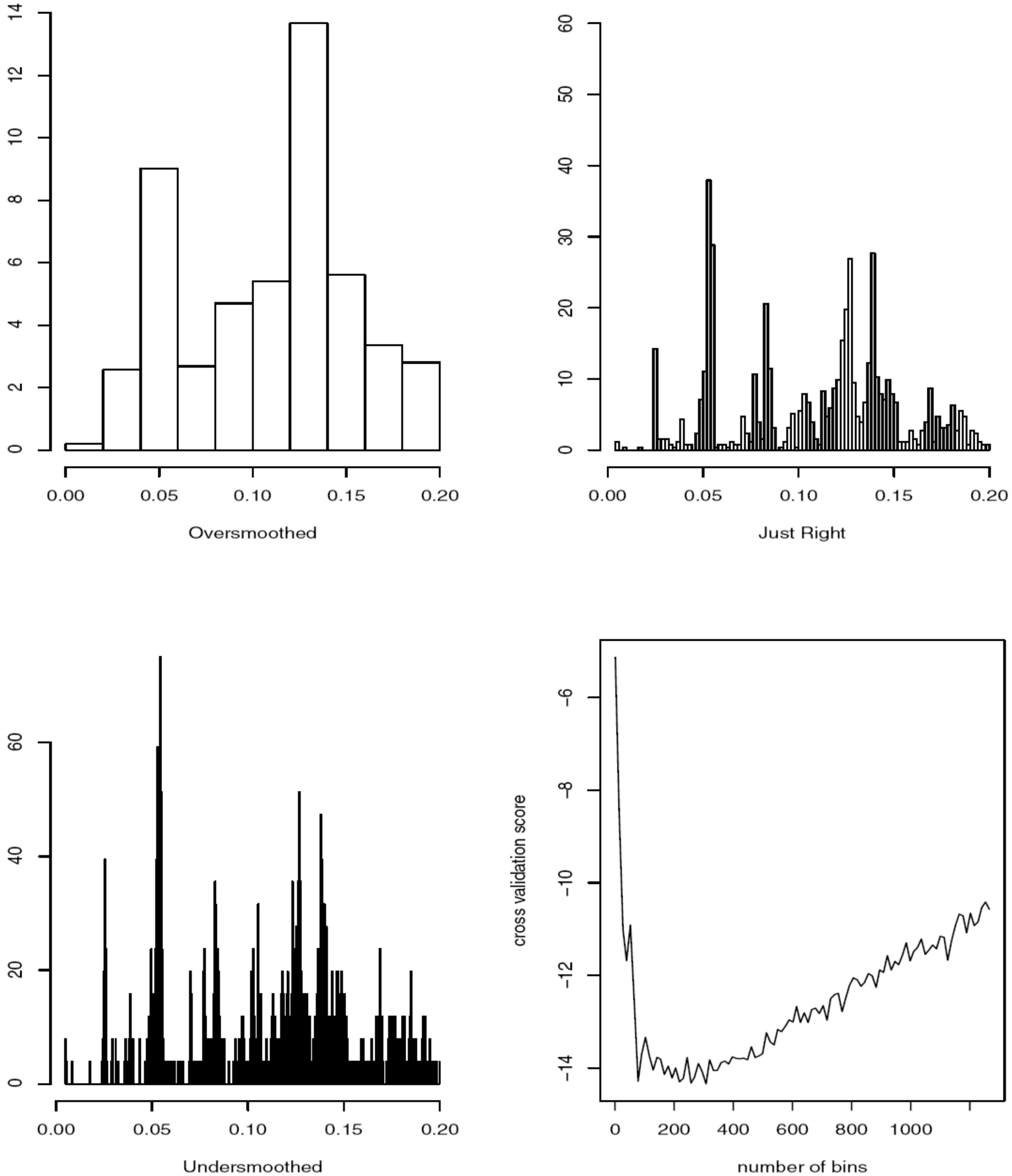
$$\widehat{f}_n(x) = \begin{cases} \widehat{p}_1/h & x \in B_1 \\ \widehat{p}_2/h & x \in B_2 \\ \vdots & \vdots \\ \widehat{p}_m/h & x \in B_m \end{cases}$$

$$\widehat{f}_n(x) = \sum_{j=1}^{n} \frac{\widehat{p}_j}{h} I(x \in B_j)$$

Очевидно, что при $x \in B_j$ **и малом значении** h **будет выполнено, что**

$$\mathbb{E}(\widehat{f}_n(x)) = \frac{\mathbb{E}(\widehat{p}_j)}{h} = \frac{p_j}{h} = \frac{\int_{B_j} f(u)du}{h} \approx \frac{f(x)h}{f(x)} = f(x)$$

6 **<u>Пример</u>. n = 1.266 точек, каждая точка равна расстоянию от земли до некоторой галактики. Наиболее правильная оценка – правая верхняя гистограмма. Из этой гистограммы следует наличие нескольких кластеров. То, как размер и количество кластеров зависят от времени позволяет астрономам понять эволюцию вселенной.**


0 2 4 6 8 10 12 14
0.00 0.05 0.10 0.15 0.20
Oversmoothed
0 10 20 30 40 50 60
0.00 0.05 0.10 0.15 0.20
Just Right
0 20 40 60
0.00 0.05 0.10 0.15 0.20
Undersmoothed
cross validation score
-6 -8 -10 -12 -14
0 200 400 600 800 1000
number of bins

<u>Теорема</u>. Пусть x **и** m **фиксированы, причем** B_j **содержит точку** x**. Тогда**

$$\mathbb{E}(\widehat{f}_n(x)) = \frac{p_j}{h}$$ **и** $$\mathbb{V}(\widehat{f}_n(x)) = \frac{p_j(1-p_j)}{nh^2}$$

Рассмотрим выбор оптимального значения ширины подробнее. Пусть $x \in B_j$**, для любого другого** $u \in B_j$

8

$$f(u) \approx f(x) + (u-x)f'(x)$$

$$p_j = \int_{B_j} f(u)du \approx \int_{B_j} \left(f(x) + (u-x)f'(x) \right) du$$

$$= f(x)h + hf'(x) \left(h \left(j - \frac{1}{2} \right) - x \right)$$

Отсюда получаем, что смещение $b(x)$ **будет равно**

$$\begin{aligned} b(x) &= \mathbb{E}(\widehat{f}_n(x)) - f(x) = \frac{p_j}{h} - f(x) \\ &\approx \frac{f(x)h + hf'(x)\left(h\left(j-\frac{1}{2}\right)-x\right)}{h} - f(x) \\ &= f'(x)\left(h\left(j-\frac{1}{2}\right)-x\right) \end{aligned}$$

Если $\widetilde{x}_j$ **- центр ячейки, то**

$$\begin{aligned} \int_{B_j} b^2(x)\,dx &\approx \int_{B_j} (f'(x))^2 \left(h\left(j-\frac{1}{2}\right)-x\right)^2 dx \\ &\approx (f'(\widetilde{x}_j))^2 \int_{B_j} \left(h\left(j-\frac{1}{2}\right)-x\right)^2 dx \\ &= (f'(\widetilde{x}_j))^2 \frac{h^3}{12} \end{aligned}$$

Значит,

$$\int_0^1 b^2(x)dx \;=\; \sum_{j=1}^{m}\int_{B_j} b^2(x)dx \approx \sum_{j=1}^{m}(f'(\widetilde{x}_j))^2\frac{h^3}{12}$$

$$= \;\frac{h^2}{12}\sum_{j=1}^{m} h\,(f'(\widetilde{x}_j))^2 \approx \frac{h^2}{12}\int_0^1 (f'(x))^2dx$$

10 **Видно, что смещение является возрастающей функцией** h**. Теперь рассмотрим дисперсию. Для малых** h **выполнено, что** $1-p_j\approx 1$**, поэтому**

$$\begin{aligned} v(x) &\approx \frac{p_j}{nh^2} \\ &= \frac{f(x)h + hf'(x)\left(h\left(j-\frac{1}{2}\right)-x\right)}{nh^2} \\ &\approx \frac{f(x)}{nh} \end{aligned}$$

Значит, $\int_0^1 v(x)dx \approx \frac{1}{nh}$

<u>Теорема</u>. Допустим, что $\int (f'(u))^2 du < \infty$**. Тогда**

$$R(\widehat{f}_n, f) \approx \frac{h^2}{12}\int (f'(u))^2 du + \frac{1}{nh}$$

При этом значение h^***, для которого риск принимает минимальное значение, равно**

$$h^* = \frac{1}{n^{1/3}}\left(\frac{6}{\int (f'(u))^2 du}\right)^{1/3}$$

Для такого h^* **выполнено, что**

$$R(\widehat{f}_n, f) \approx \frac{C}{n^{2/3}}$$

С

$$C = (3/4)^{2/3} \left(\int (f'(u))^2 du \right)^{1/3}$$

12

Таким образом, MISE **спадает до нуля со скоростью** $n^{-2/3}$**. Отметим, что большинство параметрических оценок имеет скорость сходимости, равную** n^{-1}**.**

Оптимальное значение h^* **не годится для практики, поскольку оно зависит от неизвестного значения плотности => надо оценить риск и минимизировать его по ширине** h

$$
\begin{aligned}
L(h) &= \int (\widehat{f}_n(x) - f(x))^2\, dx \\
&= \int \widehat{f}_n^{\,2}(x)\, dx - 2\int \widehat{f}_n(x) f(x) dx + \int f^2(x)\, dx
\end{aligned}
$$

$$
J(h) = \int \widehat{f}_n^{\,2}(x)\, dx - 2\int \widehat{f}_n(x) f(x) dx
$$

13 **В дальнейшем риском будем называть величину** $\mathbb{E}(J(h))$

<u>Определение</u>. Оценка риска с помощью перекрестной проверки (cross-validation) равна

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

,

где $\widehat{f}_{(-i)}$ **- оценка гистограммы по выборке без i-го наблюдения**

<u>Теорема</u>.

$$\mathbb{E}(\widehat{J}(x)) \approx \mathbb{E}(J(x))$$

<u>Теорема</u>.

$$\hat{J}(h) = \frac{2}{(n-1)h} - \frac{n+1}{h(n-1)} \sum_{j=1}^{m} \hat{p}_j^2$$

Определим доверительные интервалы для f. Допустим, что $\widehat{f}_n$ - гистограмма с m ячейками и шириной ячеек, равной $h = 1/m$. Мы не можем сделать какие-то достоверные утверждения о локальном поведении настоящей плотности f (будет объяснено позже), взамен мы построим доверительный интервал для усреднения плотности f по ячейкам ширины $h = 1/m$.

Определим

$$\overline{f}_n(x) = \mathbb{E}(\widehat{f}_n(x)) = \frac{p_j}{h}$$ **для** $x \in B_j$**,**

где $p_j = \int_{B_j} f(u)du$**. По сути** $\overline{f}(x)$ **- “гистограммное” усреднение функции** f**.**

15

<u>Определение</u>. Пара функций $(\ell_n(x), u_n(x))$ **является** $1-\alpha$ **доверительным областью (трубкой), если**

$$\mathbb{P}\left(\ell(x) \leq \overline{f}_n(x) \leq u(x) \ \text{ for all } x\right) \geq 1-\alpha$$

<u>Теорема</u>. **Пусть** $m = m(n)$ **- количество ячеек в гистограмме** $\widehat{f}_n$**.** **Предположим, что** $m(n) \to \infty$ **и** $m(n) \log n / n \to 0$ **при** $n \to \infty$**.** **Определим**

$$\ell_n(x) = \left(\max\left\{\sqrt{\widehat{f}_n(x)} - c, 0\right\}\right)^2$$

16

$$u_n(x) = \left(\sqrt{\widehat{f}_n(x)} + c\right)^2$$

при $$c = \frac{z_{\alpha/(2m)}}{2}\sqrt{\frac{m}{n}}$$

Тогда $(\ell_n(x), u_n(x))$ **является** $1 - \alpha$ **доверительным интервалом**

<u>**Доказательство**</u>**. Из центральной предельной теоремы следует, что** $\widehat{p}_j \approx N(p_j, p_j(1-p_j)/n)$**. Согласно дельта-методу**

$$\sqrt{\widehat{p}_j} \approx N(\sqrt{p_j}, 1/(4n))$$

Более того, можно показать, что $\sqrt{\widehat{p}_j}$ **приблизительно независимы. Тогда**

$$2\sqrt{n}\left(\sqrt{\widehat{p}_j} - \sqrt{p_j}\right) \approx Z_j$$

,

где $Z_1, \ldots, Z_m \sim N(0,1)$**. Пусть**

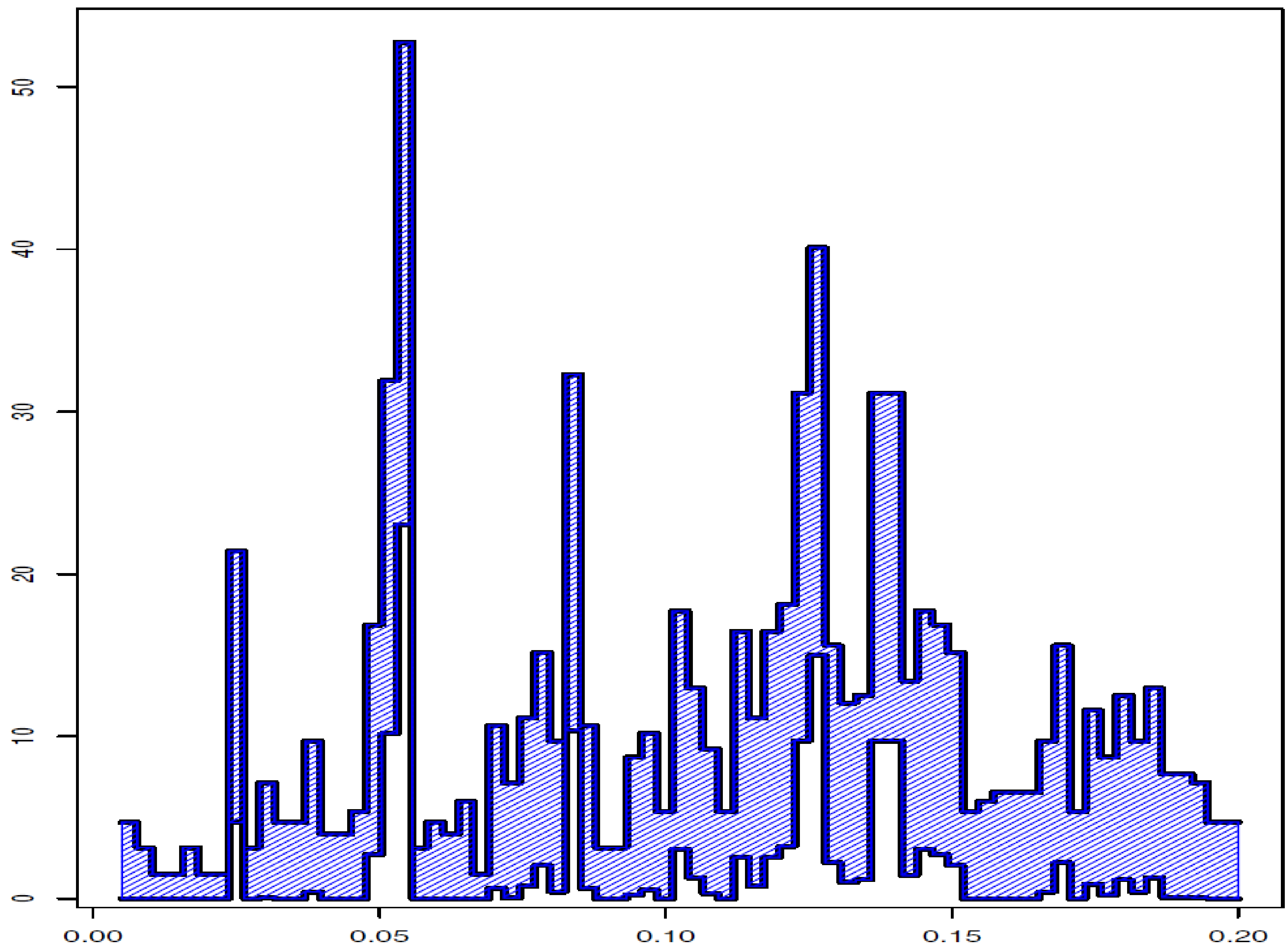
$$A = \left\{ \ell_n(x) \le \overline{f}_n(x) \le u_n(x) \ \text{ for all } x \right\} = \left\{ \max_x \left| \sqrt{\widehat{f}_n(x)} - \sqrt{\overline{f}(x)} \right| \le c \right\}$$

Тогда

$$\mathbb{P}(A^c) \;=\; \mathbb{P}\left(\max_x \left| \sqrt{\widehat{f}_n(x)} - \sqrt{\overline{f}(x)} \right| > c \right) = \mathbb{P}\left(\max_j \left| \sqrt{\frac{\widehat{p}_j}{h}} - \sqrt{\frac{p_j}{h}} \right| > c \right)$$

$$
\begin{aligned}
&= \mathbb{P}\left(\max_j 2\sqrt{n}\left|\sqrt{\widehat{p}_j} - \sqrt{p_j}\right| > z_{\alpha/(2m)}\right) \\
&\approx \mathbb{P}\left(\max_j |Z_j| > z_{\alpha/(2m)}\right) \leq \sum_{j=1}^{m} \mathbb{P}\left(|Z_j| > z_{\alpha/(2m)}\right) \\
&= \sum_{j=1}^{m} \frac{\alpha}{m} = \alpha
\end{aligned}
$$

19



с) Ядерная оценка плотности – позволяет получить более гладкие по сравнению с гистограммой оценки, которые быстрее сходятся к неизвестной величине

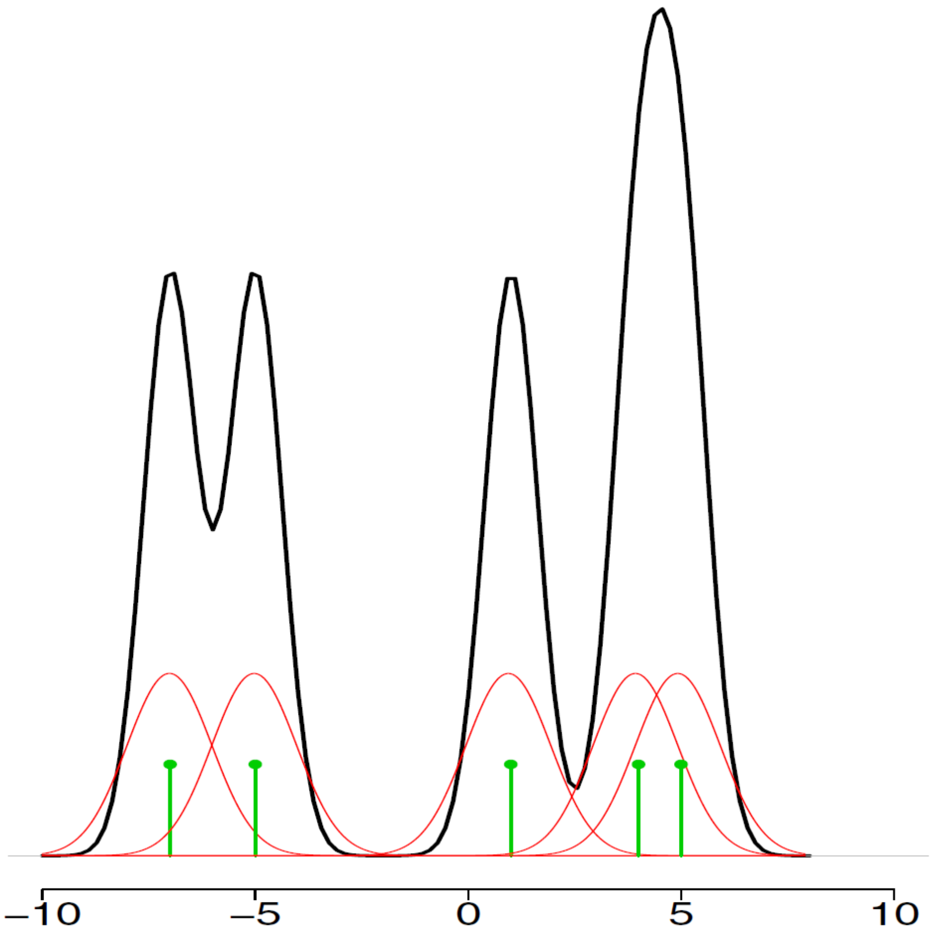
Пусть $X_1, \ldots, X_n$ **обозначает наблюдаемые данные, плотность которых равна** f**. В данном разделе под ядром понимается любая такая гладкая функция** K**, что** $K(x) \geq 0$**,** $\int K(x)\,dx = 1$**,** $\int xK(x)dx = 0$ **и** $\sigma_K^2 \equiv \int x^2 K(x)dx > 0$**.**

Примеры:

Ядро Епанечникова

$$K(x) = \begin{cases} 3/4 \cdot \left(1 - x^2/5\right) / \sqrt{5}, |x| < \sqrt{5} \\ 0 \quad \textit{иначе} \end{cases}$$

Гауссовское (нормальное) ядро $K(x) = (2\pi)^{-1/2} e^{-x^2/2}$



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

h **- ширина ядра**

При $h \to 0$ оценка $\widehat{f}_n$ состоит из резких пиков в каждой из точек выборки. При $h \to \infty$ оценка $\widehat{f}_n$ стремится к равномерному распределению

Вообще говоря, вид ядерной функции K не очень сильно влияет на “качество” оценки, а выбор значения ширины ядра – влияет сильно.

<u>Теорема</u>.

$$R(f, \widehat{f}_n) \approx \frac{1}{4}\sigma_K^4 h^4 \int (f''(x))^2 + \frac{\int K^2(x)dx}{nh}$$

$$\sigma_K^2 = \int x^2 K(x)dx$$

$$h^* = \frac{c_1^{-2/5} c_2^{1/5} c_3^{-1/5}}{n^{1/5}},$$

где $c_1 = \int x^2 K(x)dx$, $c_2 = \int K(x)^2 dx$, $c_3 = \int (f''(x))^2 dx$

$$R(f, \widehat{f}_n) \approx \frac{c_4}{n^{4/5}}$$

для некоторой константы $c_4 > 0$

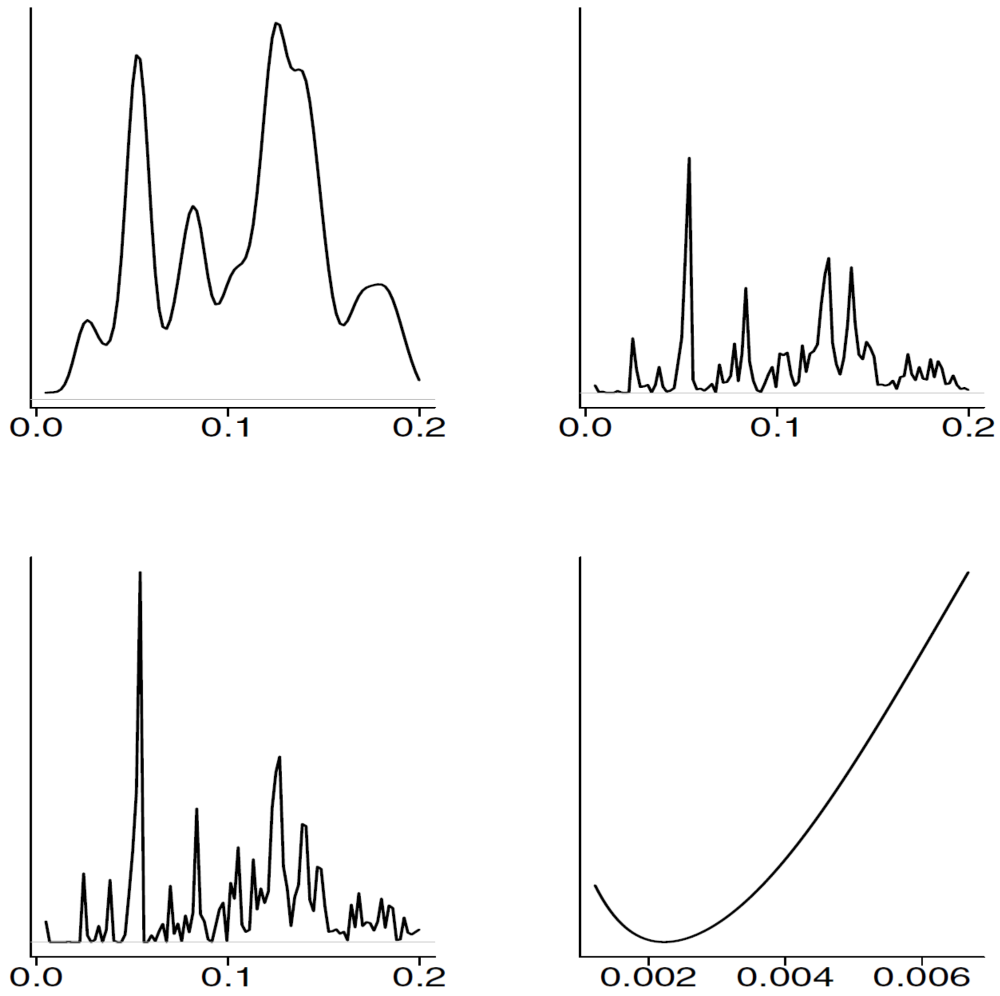
<u>Доказательство</u>. **Пусть** $K_h(x, X) = h^{-1} K\left((x - X)/h\right)$, $\widehat{f}_n(x) = n^{-1} \sum_i K_h(x, X_i)$, $\mathbb{E}[\widehat{f}_n(x)] = \mathbb{E}[K_h(x, X)]$ **и** $\mathbb{V}[\widehat{f}_n(x)] = n^{-1} \mathbb{V}[K_h(x, X)]$. **Тогда**

$$
\begin{aligned}
\mathbb{E}[K_h(x, X)] &= \int \frac{1}{h} K\left(\frac{x-t}{h}\right) f(t)\, dt \\
&= \int K(u) f(x - hu)\, du \\
&= \int K(u) \left[f(x) - h f'(x) + \frac{1}{2} f''(x) + \cdots \right] du \\
&= f(x) + \frac{1}{2} h^2 f''(x) \int u^2 K(u)\, du \cdots
\end{aligned}
$$

поскольку $\int K(x)\, dx = 1$ **и** $\int x\, K(x)\, dx = 0$**. Таким образом, смещение равно**

$$
\mathbb{E}[K_h(x, X)] - f(x) \approx \frac{1}{2} \sigma_k^2 h^2 f''(x)
$$

Аналогично, $$\mathbb{V}[\widehat{f}_n(x)] \approx \frac{f(x)\int K^2(x)\,dx}{n\,h_n}$$



Таким образом, ядерная оценка сходится со скоростью $n^{-4/5}$, а не со скоростью $n^{-2/3}$, как гистограмма. Можно показать, что при достаточно слабых условиях нельзя получить скорость лучше, чем $n^{-4/5}$

Так как оптимальное значение ширины ядра вычислить нельзя, то риск оценивают по формуле

$$\widehat{J}(h) = \int \widehat{f}^2(x)dz - \frac{2}{n}\sum_{i=1}^{n}\widehat{f}_{-i}(X_i)$$

,

$\widehat{f}_{-i}$ **- оценка плотности по выборке, из которой было исключено i-ое наблюдение**

<u>Теорема</u>. Для любого $h > 0$

$$\mathbb{E}\left[\widehat{J}(h)\right] = \mathbb{E}\left[J(h)\right]$$

$$\widehat{J}(h) \approx \frac{1}{hn^2}\sum_i\sum_j K^*\left(\frac{X_i - X_j}{h}\right) + \frac{2}{nh}K(0)$$

27

$$K^*(x) = K^{(2)}(x) - 2K(x)$$

$$K^{(2)}(z) = \int K(z-y)K(y)dy$$

В частности, если K **является** $N(0,1)$ **гауссовским ядром, то** $K^{(2)}(z)$ **-** $N(0,2)$ **гауссовское ядро.**

Далее, выбирается h_n **, минимизирующее** $\widehat{J}(h)$**. Обоснование –**

<u>Теорема</u> (Стоуна). Допустим, что f - ограничена. Пусть $\widehat{f}_h$ обозначает ядерную оценку с шириной ядра, равной h и пусть h_n - ширина ядра, выбранная с помощью кросс-проверки. Тогда

$$\frac{\int \left(f(x) - \widehat{f}_{h_n}(x)\right)^2 dx}{\inf_h \int \left(f(x) - \widehat{f}_h(x)\right)^2 dx} \xrightarrow{\mathrm{P}} 1$$

<u>Замечание</u>. Даже если оказалось, что оценка плотности достаточно "изрезанная", это не значит, что получился плохой результата. Кросс-проверка "лучше знает"!

Построим доверительный интервал для усредненной плотности

$$\overline{f}_n = \mathbb{E}(\widehat{f}_n(x)) = \int \frac{1}{h} K\left(\frac{x-u}{h}\right) f(u)\, du$$

Допустим, что плотность имеет носитель на отрезке (a,b)**. Тогда концы интервала равны**

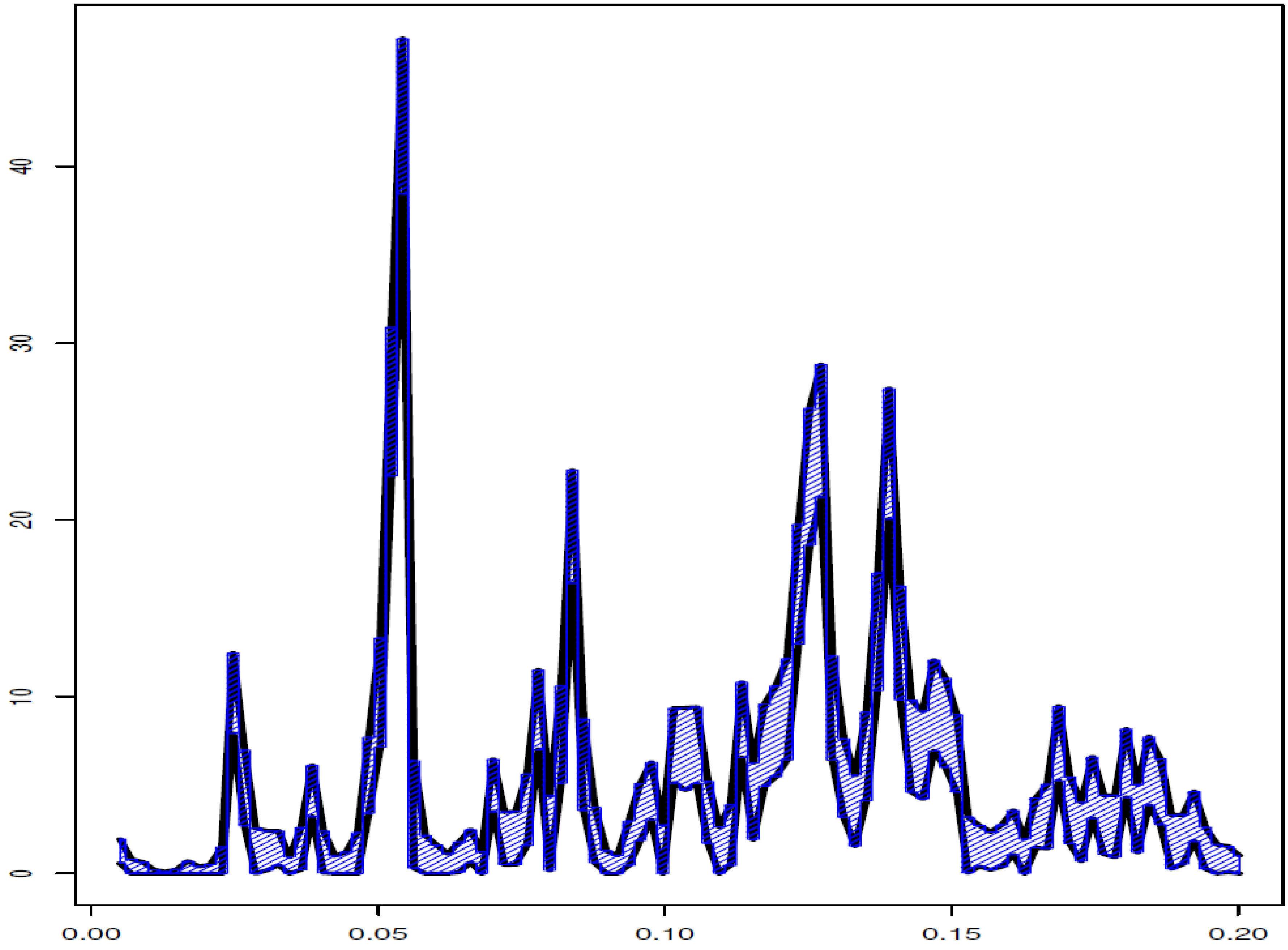
$$\ell_n(x) = \widehat{f}_n(x) - q\ \mathsf{se}(x)$$

$$u_n(x) = \widehat{f}_n(x) + q\ \mathsf{se}(x)$$

$$
\begin{aligned}
\mathsf{se}(x) &= \frac{s(x)}{\sqrt{n}}, \\
s^2(x) &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(x) - \overline{Y}_n(x))^2, \\
Y_i(x) &= \frac{1}{h}K\left(\frac{x - X_i}{h}\right), \\
q &= \Phi^{-1}\left(\frac{1 + (1-\alpha)^{1/m}}{2}\right), \\
m &= \frac{b-a}{\omega}
\end{aligned}
$$

ω - ширина ядра (или эффективная ширина ядра)

Для гауссовского ядра она может быть равна $\omega = 3h$


0
10
20
30
40
0.00
0.05
0.10
0.15
0.20

Допустим, что данные многомерные, то есть $X_i = (X_{i1}, \ldots, X_{id})$**.**

Пусть $h = (h_1, \ldots, h_d)$ **- вектор ширин ядра, тогда**

$$\widehat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i)$$

$$K_h(x - X_i) = \frac{1}{nh_1 \cdots h_d} \left\{ \prod_{j=1}^{d} K\left(\frac{x_i - X_{ij}}{h_j} \right) \right\}$$

Для простоты можно положить $h_j = s_j h$ **, где** s_j **- стандартное отклонение по** j**-ой координате.**

$$R(f, \widehat{f}_n) \approx \frac{1}{4}\sigma_K^4 \left[\sum_{j=1}^{d} h_j^4 \int f_{jj}^2(x)dx + \sum_{j \neq k} h_j^2 h_k^2 \int f_{jj} f_{kk} dx\right] + \frac{\left(\int K^2(x)dx\right)^d}{n h_1 \cdots h_d}$$

33 f_{jj} **- вторая частная производная** f**. Оптимальная ширина ядра равна** $h_i \approx c_1 n^{-1/(4+d)}$**, при этом риск будет иметь порядок** $n^{-4/(4+d)}$ **=> проклятие размерности**

Рассмотрим таблицу объемов выборки, которая нужна, чтобы относительный средний квадрат ошибки в нуле был меньше, чем 0.1**, когда плотность многомерная нормальная и используется оптимальное значение ширины ядра**

1	4
2	19
3	67
4	223
5	768
6	2790
7	10,700
8	43,700
9	187,000
10	842,000

Отсюда следует, что “объем выборки равен 842 000 в пространстве размерности 10” < = > “объем выборки равен 4 в пространстве размерности 1”

d) Непараметрическая регрессия

Рассмотрим пары значений $(x_1, Y_1), \ldots, (x_n, Y_n)$**, связанных соотношением**

$$Y_i = r(x_i) + \epsilon_i$$

где $\mathbb{E}(\epsilon_i) = 0$**.**

Можно считать, что x_i **принимают фиксированные значения, поскольку надо оценить функцию** $r(x) = \mathbb{E}(Y|X = x)$

<u>Определение</u>. Оценка Надарайя-Ватсона имеет вид

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

$$w_i(x) = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)}$$

K **- заданная ядерная функция**

$$r(x) = \mathbb{E}(Y|X=x) = \int y f(y|x) dy = \frac{\int y f(x,y) dy}{\int f(x,y) dy}$$

<u>Теорема</u>. Допустим, что $\mathbb{V}(\epsilon_i) = \sigma^2$**. Риск для оценки Надарая-Ватсона равен**

$$R(\widehat{r}_n, r) \approx \frac{h^4}{4}\left(\int x^2 K^2(x)dx\right)^4 \int \left(r''(x) + 2r'(x)\frac{f'(x)}{f(x)}\right)^2 dx + \int \frac{\sigma^2 \int K^2(x)dx}{nhf(x)}dx$$

Оптимальная ширина ядра убывает со скоростью $n^{-1/5}$**, при этом риск спадает со скоростью** $n^{-4/5}$

На практике ширину ядра выбирают так, чтобы оценка риска

$$\widehat{J}(h) = \sum_{i=1}^{n}(Y_i - \widehat{r}_{-i}(x_i))^2$$

была минимальной

Теорема.

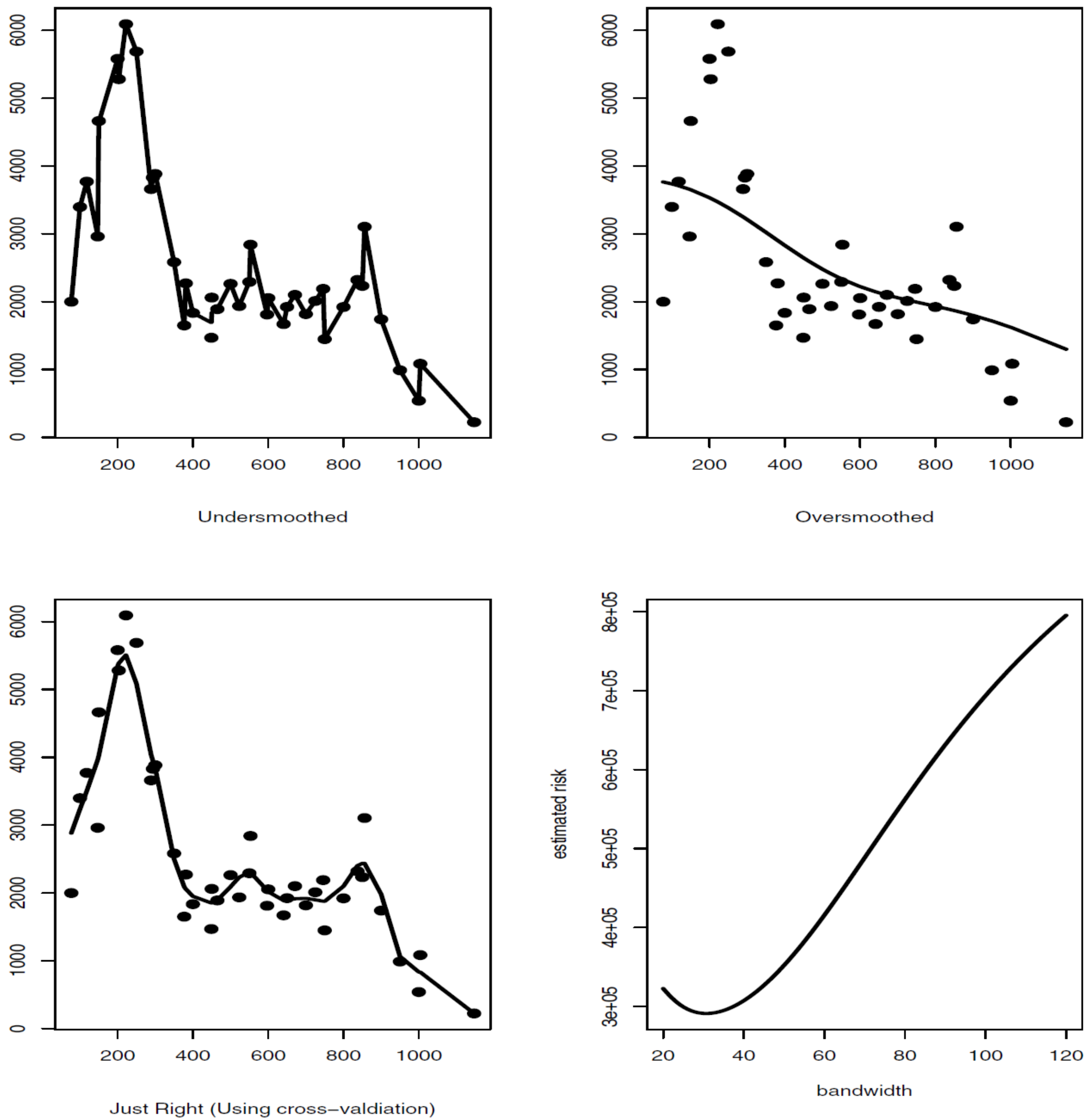
$$\widehat{J}(h)=\sum_{i=1}^{n}(Y_i-\widehat{r}(x_i))^2\frac{1}{\left(1-\frac{K(0)}{\sum_{j=1}^{n}K\left(\frac{x_i-x_j}{h}\right)}\right)^2}$$

Пример. Cosmic microwave background (CMB; космический микроволновой

фон). Данные состоят из n **пар** $(x_1,Y_1),\ \ldots,\ (x_n,Y_n)$**, где** x_i **- мультипольный момент,** Y_i **- оценка спектра мощности флуктуации температур. Наблюдения – излучение, оставшееся после большого взрыва. Если** $r(x)$ **- настоящее значение спектра мощности, то**

$$Y_i=r(x_i)+\epsilon_i$$

Положение и размер пиков функции $r(x)$ **позволяет выявить важные свойства поведения вселенной**


Undersmoothed
Oversmoothed
Just Right (Using cross-valdiation)
estimated risk
bandwidth

Определим доверительный интервал. Сначала оценим σ^2. Допустим, что x_i упорядочены по возрастанию. Предполагая, что $r(x)$ - гладкая функция, получаем, что $r(x_{i+1}) - r(x_i) \approx 0$

$$Y_{i+1} - Y_i = \left[r(x_{i+1}) + \epsilon_{i+1} \right] - \left[r(x_i) + \epsilon_i \right] \approx \epsilon_{i+1} - \epsilon_i$$

$$\mathbb{V}(Y_{i+1} - Y_i) \approx \mathbb{V}(\epsilon_{i+1} - \epsilon_i) = \mathbb{V}(\epsilon_{i+1}) + \mathbb{V}(\epsilon_i) = 2\sigma^2$$

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

Будем строить доверительное множество для сглаженной версии $\overline{r}_n(x) = \mathbb{E}(\widehat{r}_n(x))$ настоящей функции регрессии r.

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

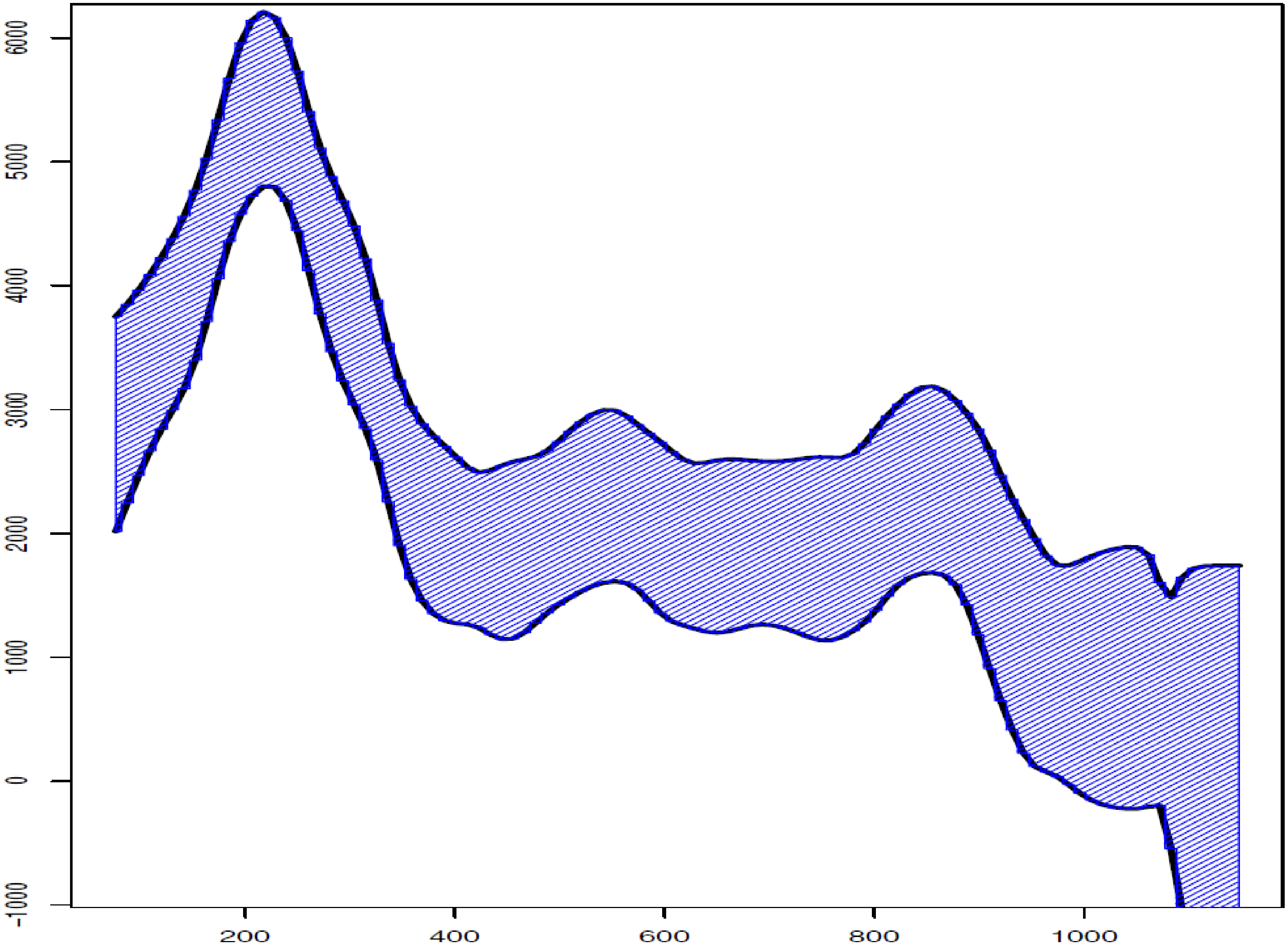
$$\ell_n(x) = \widehat{r}_n(x) - q\ \widehat{\mathsf{se}}(x),\quad u_n(x) = \widehat{r}_n(x) + q\ \widehat{\mathsf{se}}(x)$$

$$\begin{aligned}
\widehat{\mathsf{se}}(x) &= \widehat{\sigma}\sqrt{\sum_{i=1}^{n} w_i^2(x)}, \\
q &= \Phi^{-1}\left(\frac{1+(1-\alpha)^{1/m}}{2}\right), \\
m &= \frac{b-a}{\omega},
\end{aligned}$$

41

$\widehat{\sigma}$ определена выше, а ω - ширина ядра

42



Рассмотрим обобщение на многомерный случай $X = (X_1, \ldots, X_p)$ **очевидно, однако, бесполезно – гнет размерности. Можно уменьшить количество параметров за счет дополнительных ограничений, налагаемых на модель, например, предположив, что это аддитивная модель**

$$Y = \sum_{j=1}^{p} r_j(X_j) + \epsilon$$

$$Y = \sum_{j=1}^{p} r_j(X_j) + \sum_{j<k} r_{jk}(X_j X_k) + \epsilon$$

Подгонка аддитивной модели

1. “Инициализировать” $r_1(x_1),\ \ldots,\ r_p(x_p)$

2. Для $j = 1, \ldots, p$:

(a) Пусть $\epsilon_i = Y_i - \sum_{s \neq j} r_s(x_i)$

(b) r_j **- оценка функции, полученная с помощью регрессии** ϵ_i **на** j**-я переменную.**

3. STOP, если оценка функции перестала изменяться.

Эта модель принадлежит классу полупараметрических моделей (semiparametric model)

Замечание. Отметим, что доверительные интервалы, которые были подсчитаны, не является в точности доверительными интервалами для функции регрессии или плотности, но являются таковыми для сглаженных версий этих величин. Например, доверительный интервал (в случае ядерной оценки плотности с шириной h) для плотности является на самом деле доверительным интервалом для функции, равной сглаженной с помощью этого же ядра истинной плотности. Получить доверительный интервал для реальной плотности очень сложно, и вот почему.

Пусть $\widehat{f}_n(x)$ обозначает оценку функции $f(x)$. Обозначим среднее и стандартное отклонение функции $\widehat{f}_n(x)$ через $\overline{f}_n(x)$ и $s_n(x)$. Тогда

$$\frac{\widehat{f}_n(x) - f(x)}{s_n(x)} = \frac{\widehat{f}_n(x) - \overline{f}_n(x)}{s_n(x)} + \frac{\overline{f}_n(x) - f(x)}{s_n(x)}$$

Обычно 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 оценки, то есть второе слагаемое стремится к нулю при увеличении объема 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 объемах выборки, поэтому доверительный интервал не будет центрирован по отношению к оцениваемой величине.