



Data learning

Курс “Машинное обучение”
Лабораторная работа



Non-parametric multi-dimensional density estimation

Харитонов Е.А., М16-524
Вариант 2-09

2017

Исходные данные

Данные: сущности, характеризуемые двумя признаками x_1 и x_2

Размер выборки: 500

Представление: данные удобно визуализировать в качестве точек на координатной плоскости с координатами (x_1, x_2)

x_1	x_2
1.5998	0.76857
0.95352	-1.0024
1.7588	0.86545
2.679	1.9814
1.6792	0.59262
1.3089	0.15303
2.4106	1.5793
2.4737	1.1979
2.0354	1.0006
2.8452	1.5663
2.8618	1.4816
2.6181	1.21
1.4692	0.18198
1.8713	-0.03735
...	...

Рисунок 0. Пример исходных данных

Визуализация данных

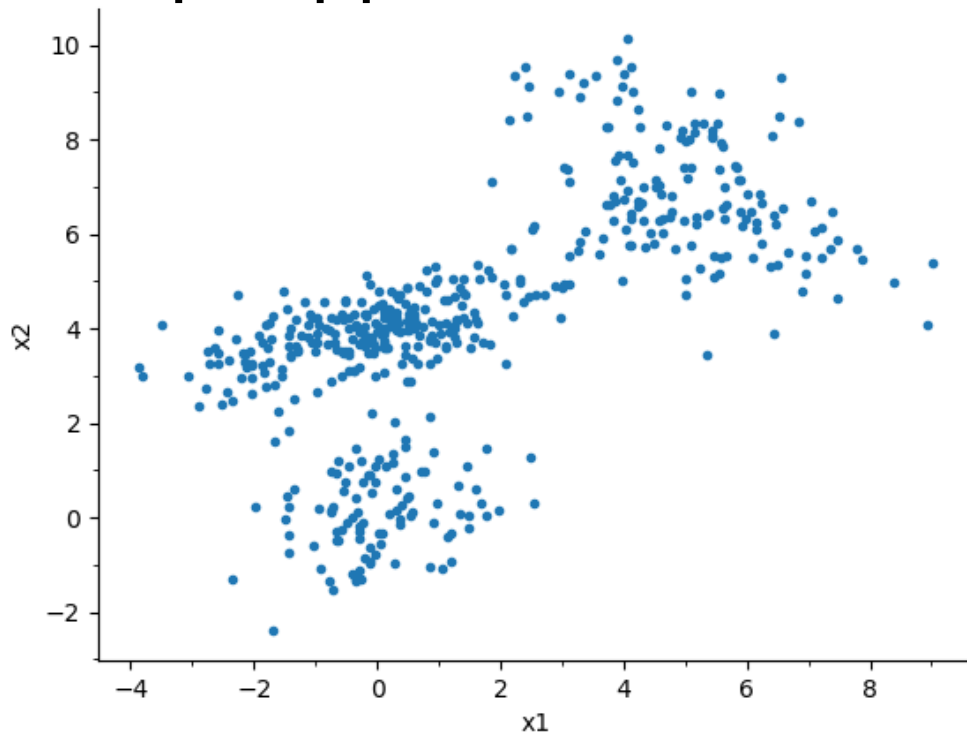
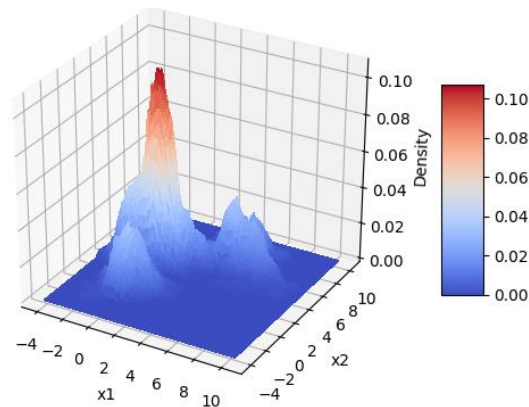
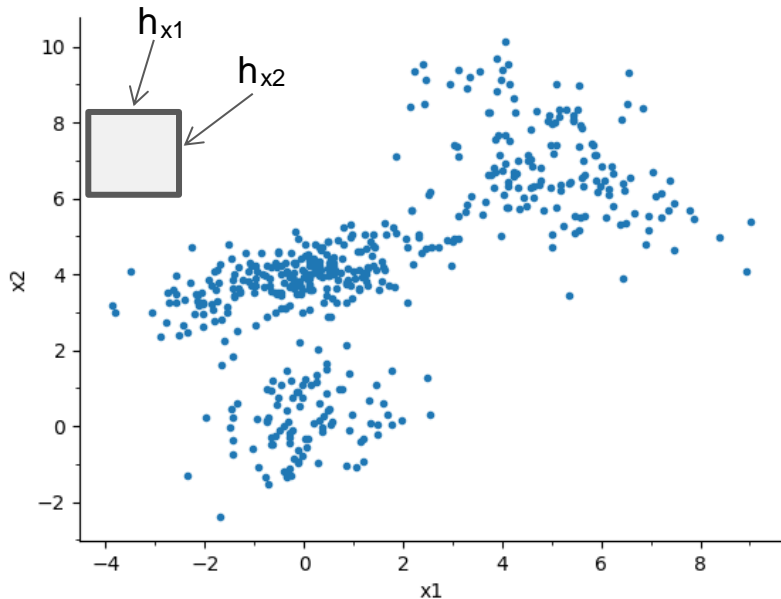


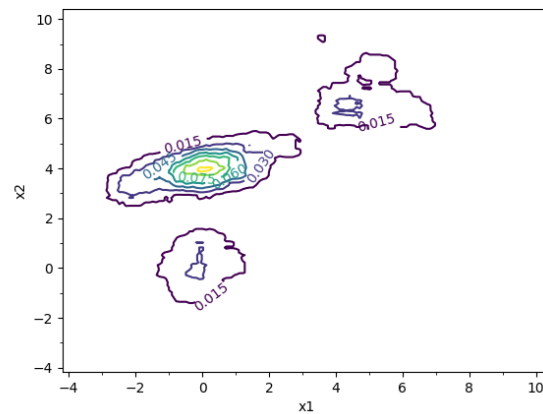
Рисунок 1. Визуализация исходных данных в виде точек на координатной плоскости

Восстановление плотности распределения

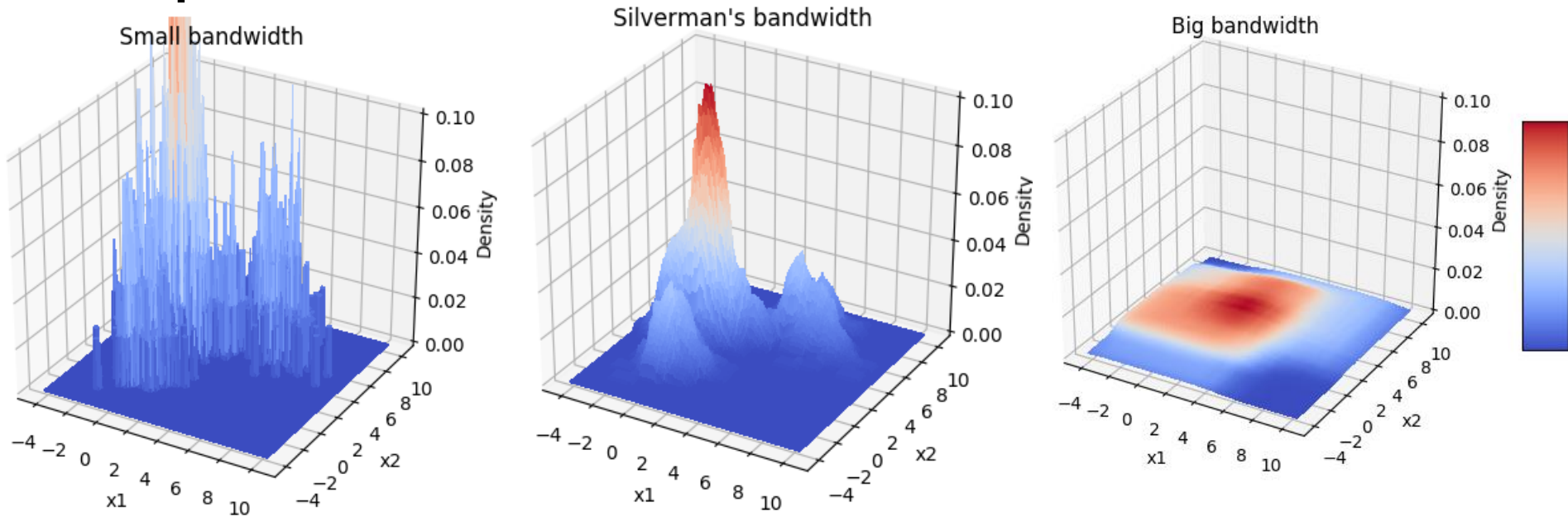
(на примере box kernel)



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



Ширина окна



Рисунки 2-4. Восстановленная плотность распределения с различными размерами окна. Слева - 0.15, 0.15 (слишком узкое), в центре – 0.69, 0.50 (по Сильверману), справа – 5.0, 5.0 (слишком широкое).

Ширина окна (правило Сильвермана)

$$h_{MISE} = 0.9 \min \left(\tilde{\sigma}, \frac{\Delta}{1.34} \right) n^{-1/5}$$

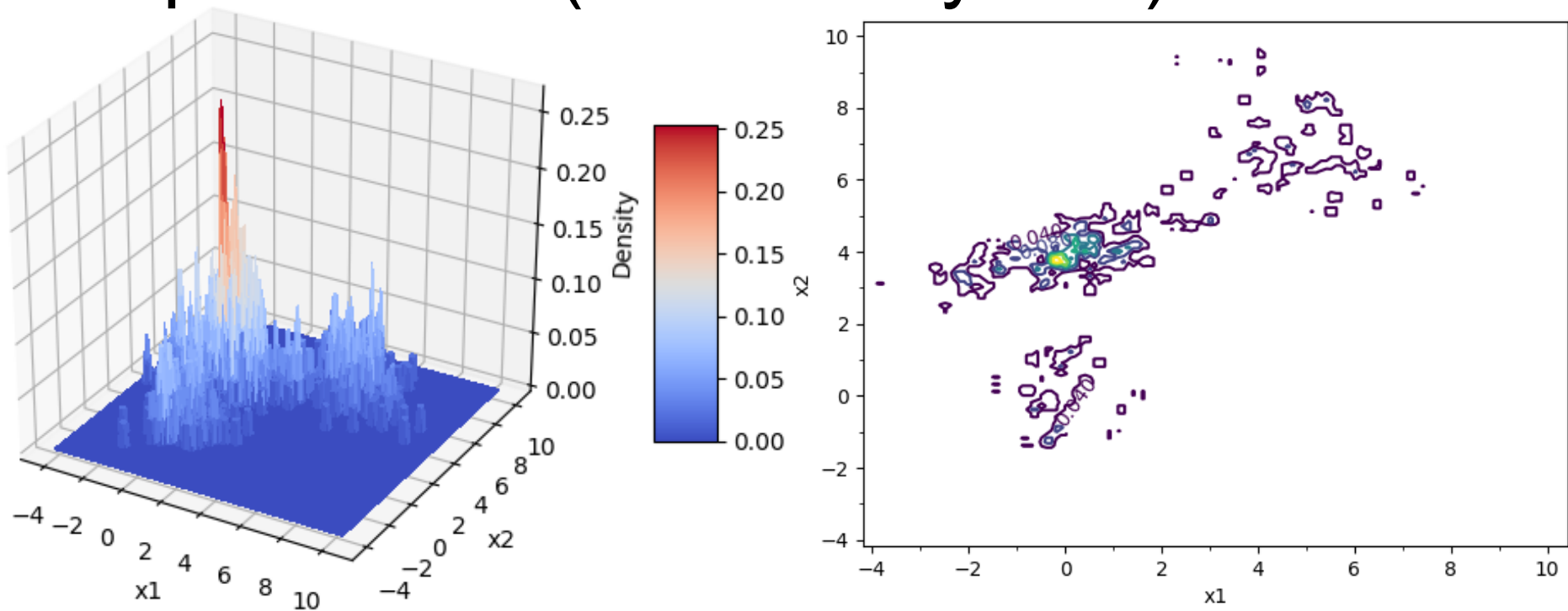
где

$$\Delta = x_{0.75} - x_{0.25} \quad (\text{interquartile range})$$

n – размер выборки

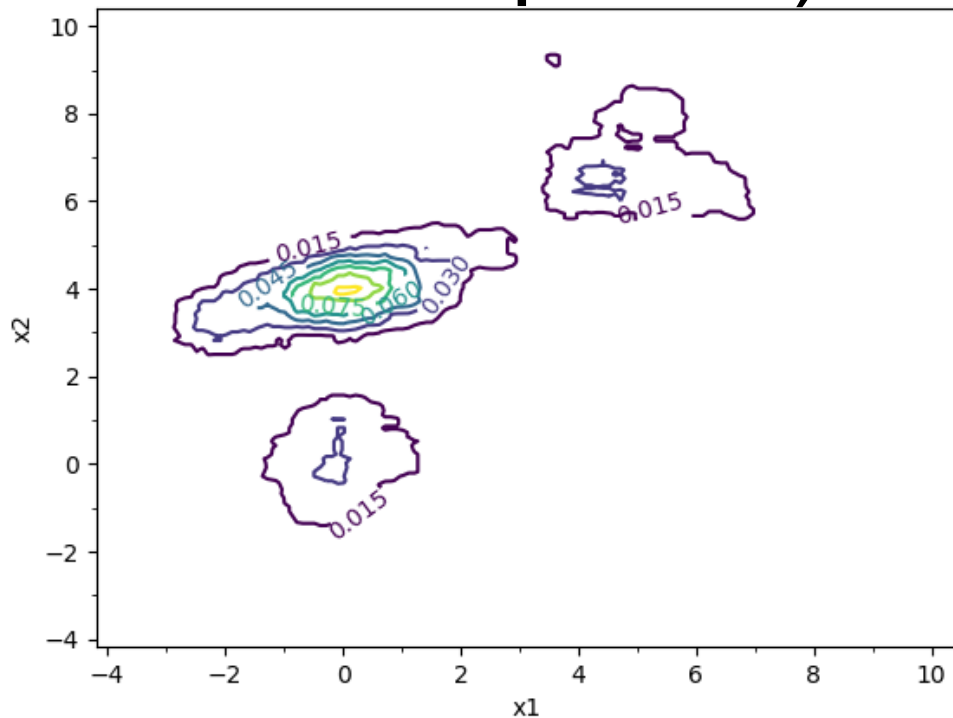
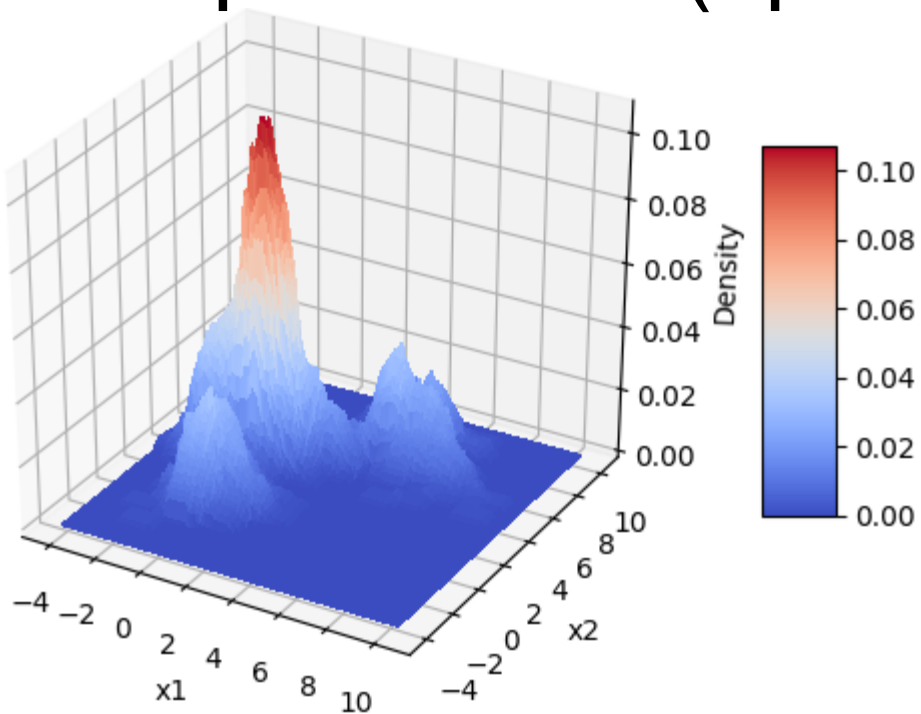
$\tilde{\sigma}$ – среднеквадратическое отклонение

Ширина окна (слишком узкое)



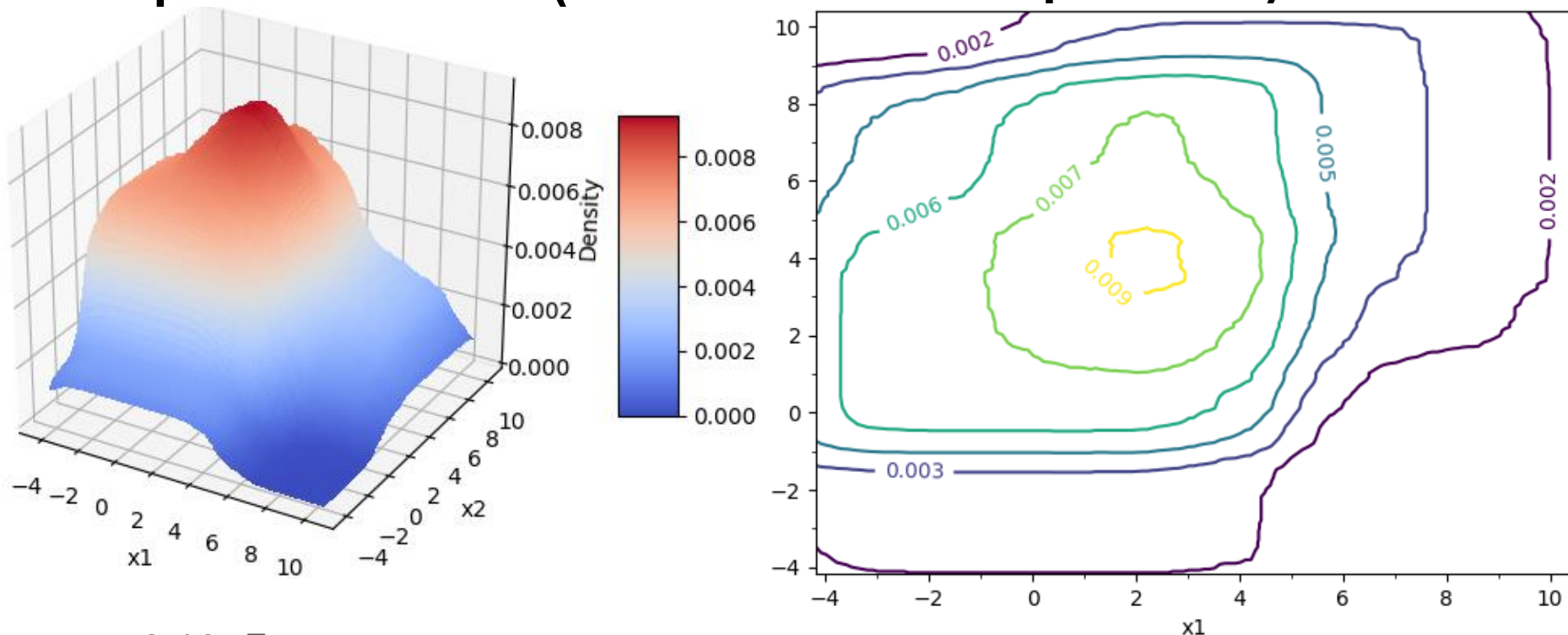
Рисунки 5,6. Диаграммы поверхности и контуров восстановленной плотности распределения со слишком узким размером окна – $(0.15, 0.15)$

Ширина окна (правило Сильвермана)



Рисунки 7,8. Диаграммы поверхности и контуров восстановленной плотности распределения с размером окна по правилу Сильвермана – (0.69, 0.50)

Ширина окна (слишком широкое)



Рисунки 9,10. Диаграммы поверхности и контуров восстановленной плотности распределения со слишком широким размером окна – (5.0, 5.0)

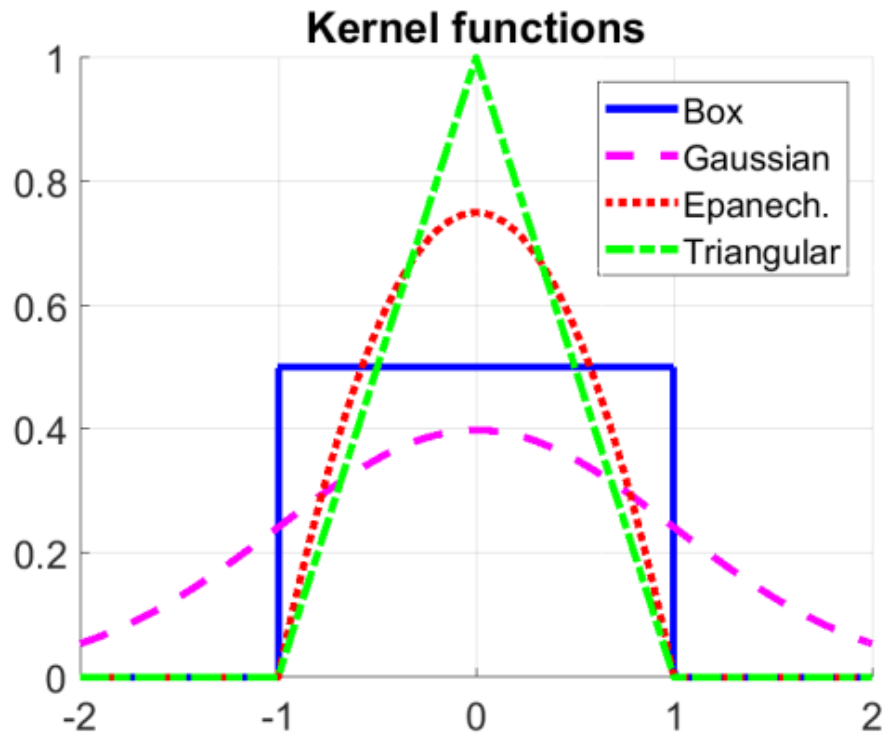
Выбор функции ядра (kernel)

Box (uniform) kernel: $K(u) = \begin{cases} \frac{1}{2}, & |u| < 1 \\ 0, & \text{otherwise} \end{cases}$

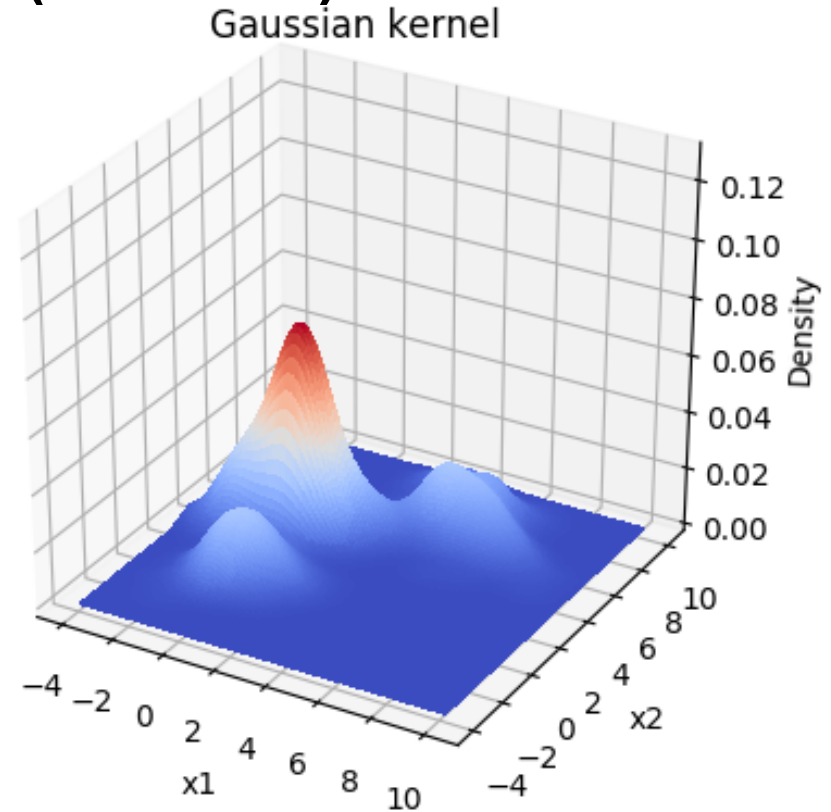
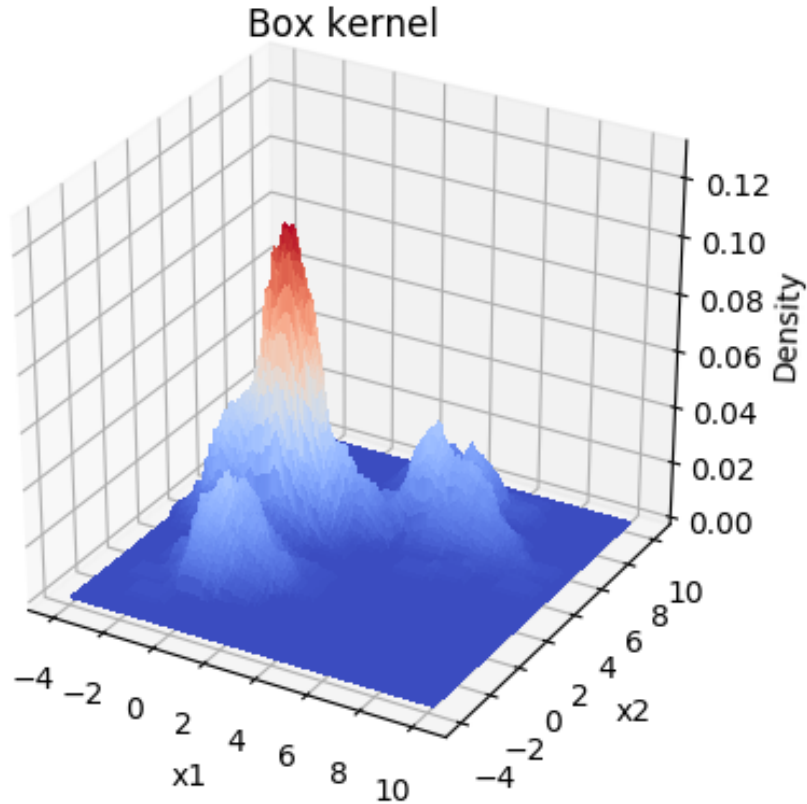
Gaussian kernel: $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$

Epanechnikov kernel: $K(u) = \begin{cases} \frac{3}{4}(1 - u^2), & |u| < 1 \\ 0, & \text{otherwise} \end{cases}$

Triangular kernel: $K(u) = \begin{cases} 1 - |u|, & |u| < 1 \\ 0, & \text{otherwise} \end{cases}$

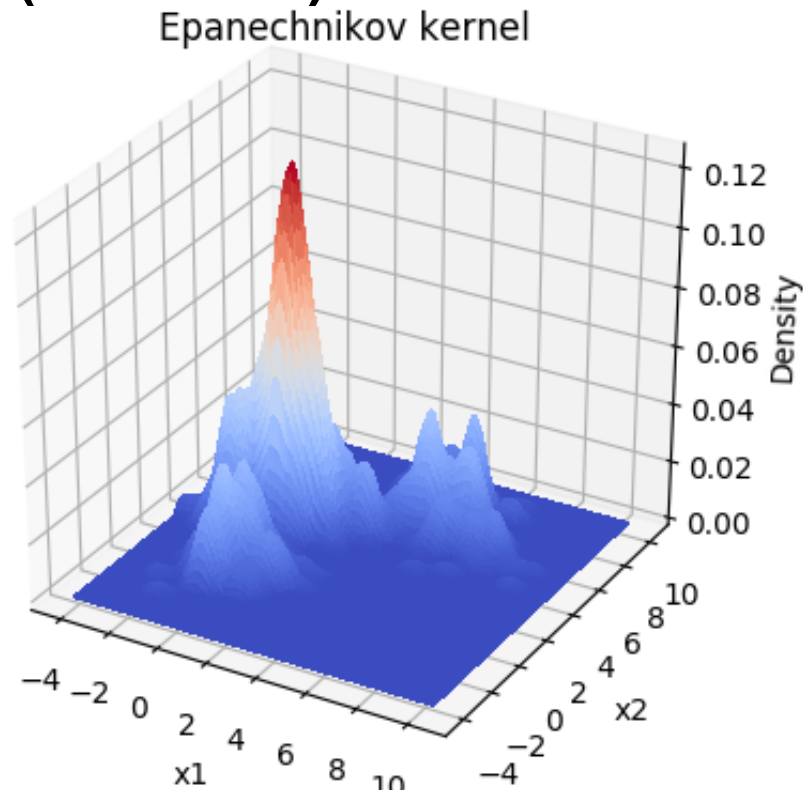
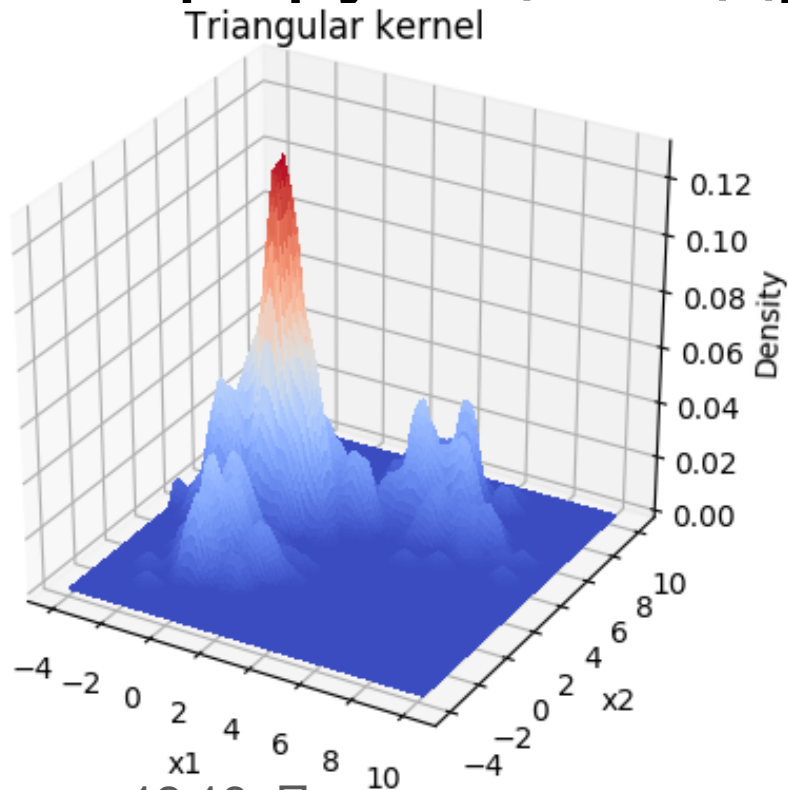


Выбор функции ядра (kernel)



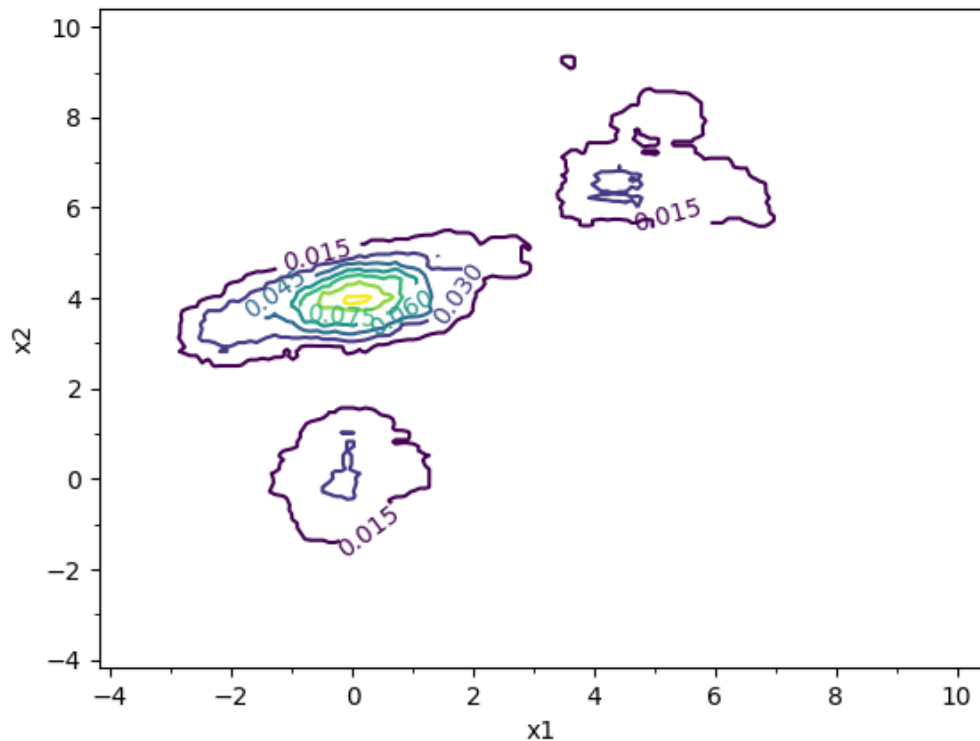
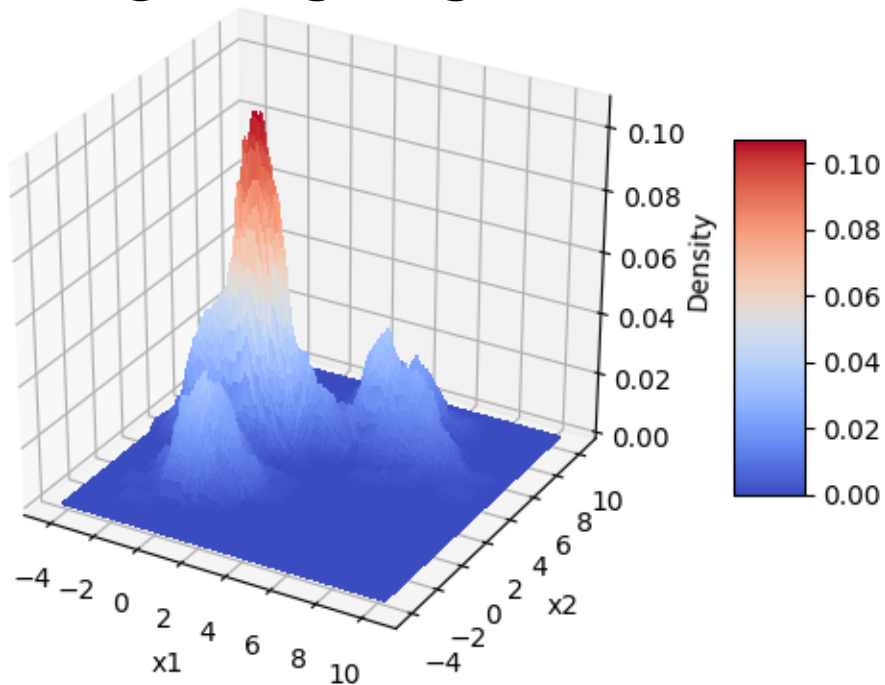
Рисунки 10,11. Плотности распределения восстановленные с помощью ядер Box kernel (слева) и Gaussian kernel (справа). Ширина окна по Сильверману.

Выбор функции ядра (kernel)



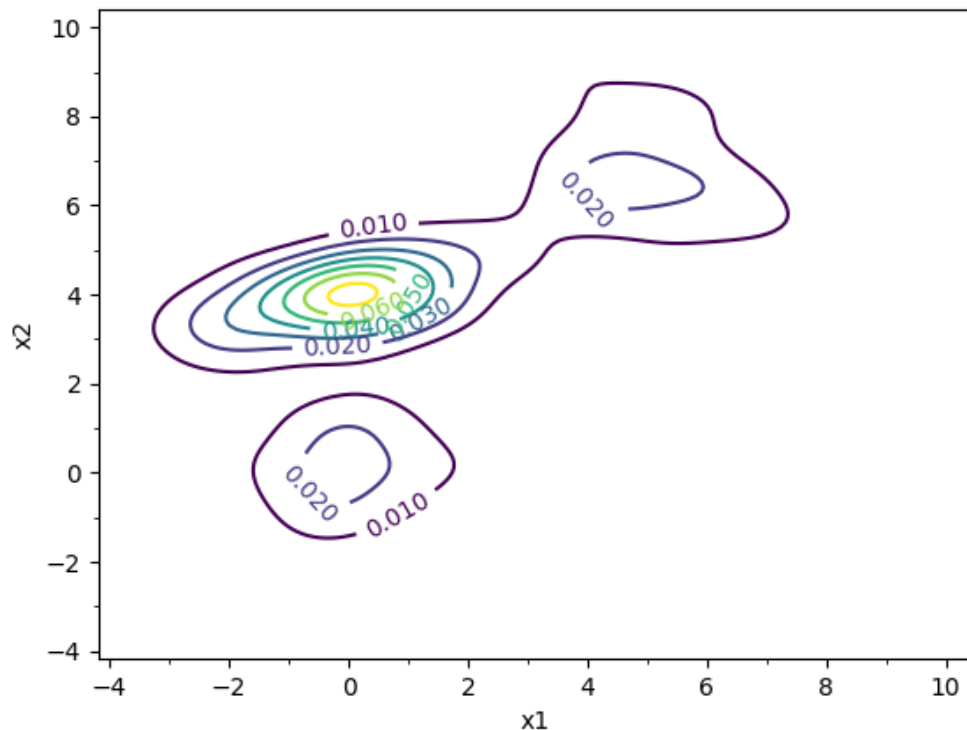
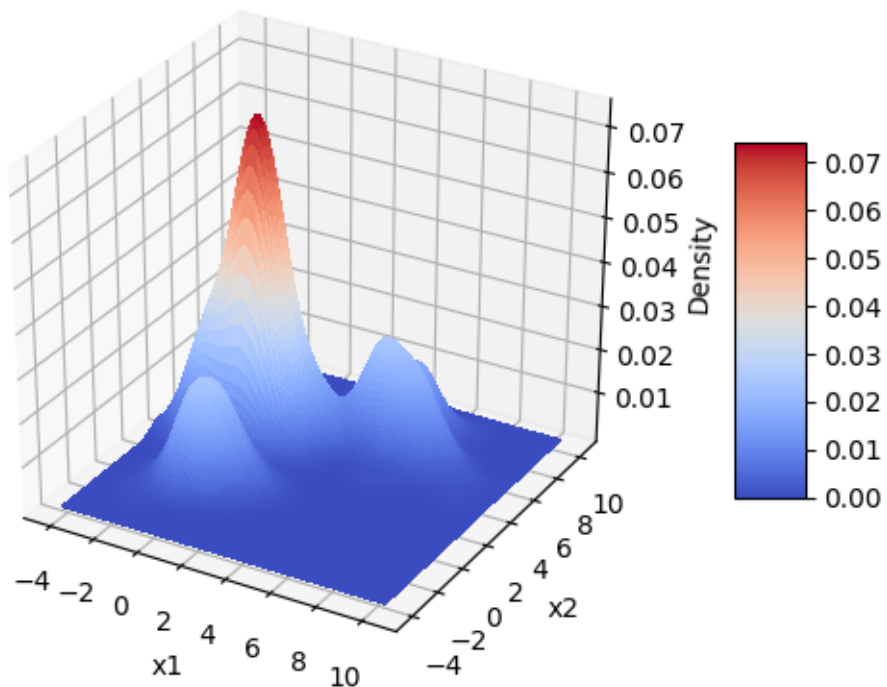
Рисунки 12,13. Плотности распределения восстановленные с помощью ядер Triangular kernel (слева) и Epanechnikov kernel (справа). Ширина окна по Сильверману.

Box kernel



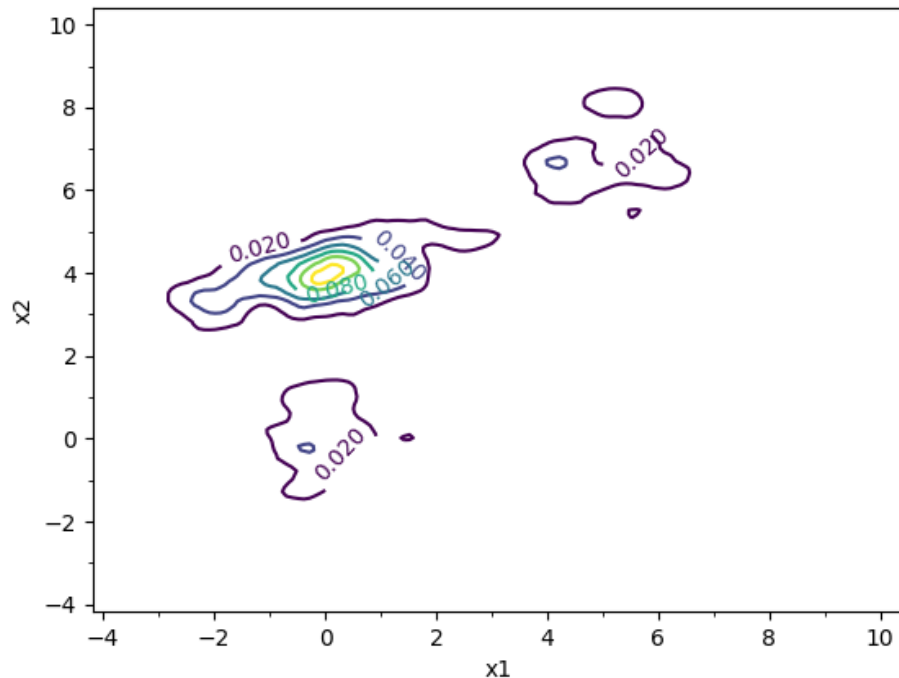
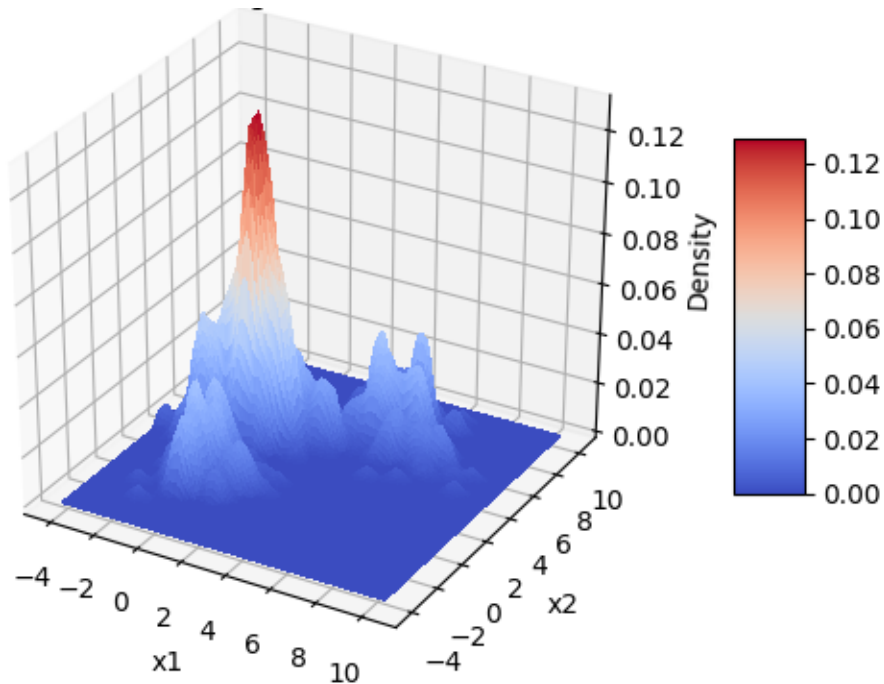
Рисунки 14,15. Диаграммы поверхности и контуров плотности распределения, восстановленные с помощью ядра Box kernel. Ширина окна по Сильверману.

Gaussian kernel



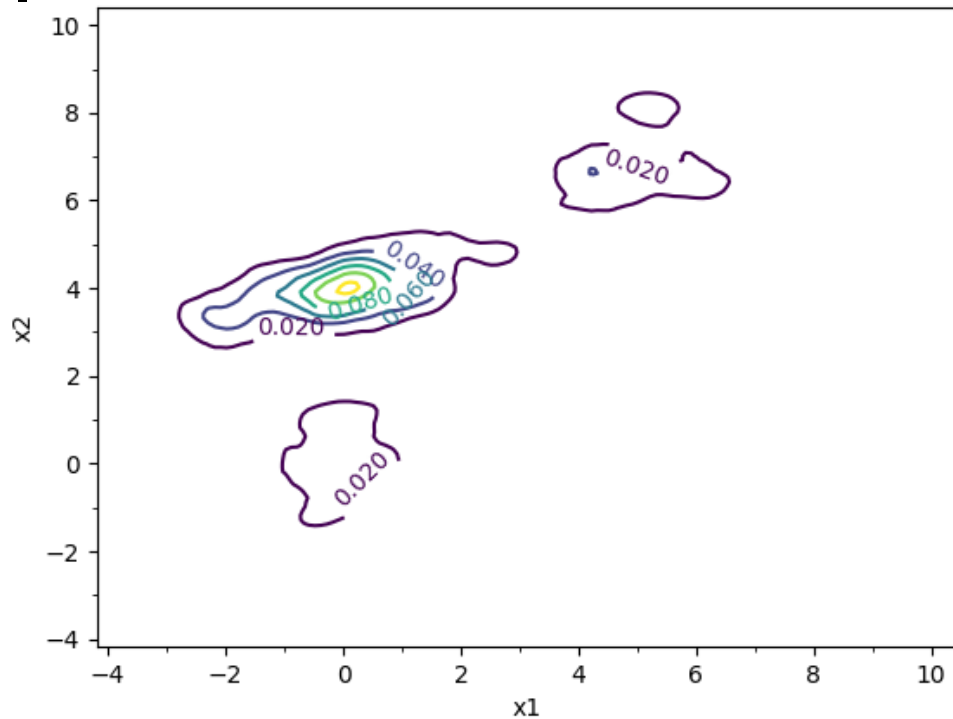
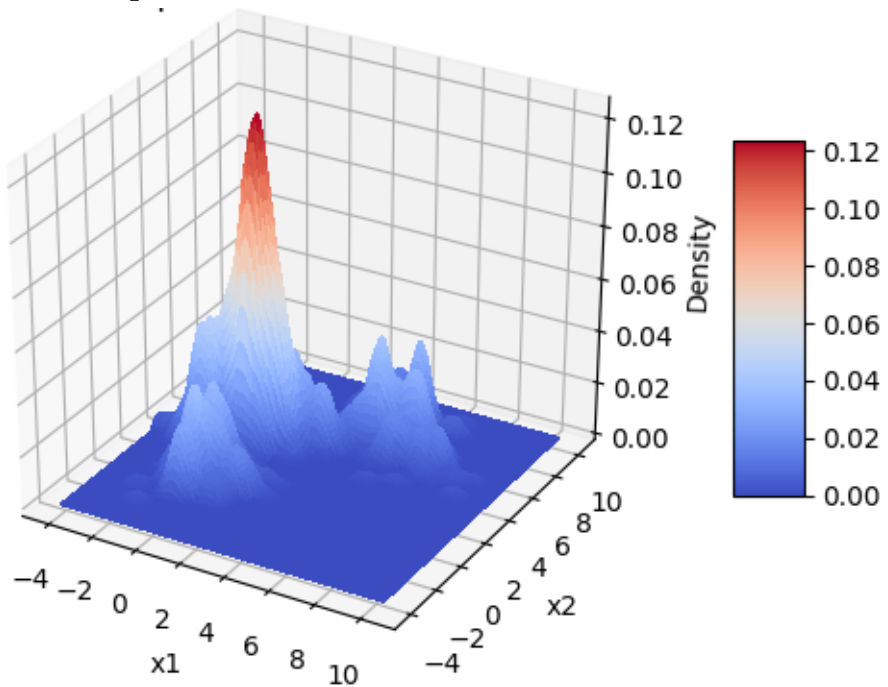
Рисунки 16,17. Диаграммы поверхности и контуров плотности распределения, восстановленные с помощью ядра Gaussian kernel. Ширина окна по Сильверману.

Triangular kernel



Рисунки 18,19. Диаграммы поверхности и контуров плотности распределения, восстановленные с помощью ядра Triangular kernel. Ширина окна по Сильверману.

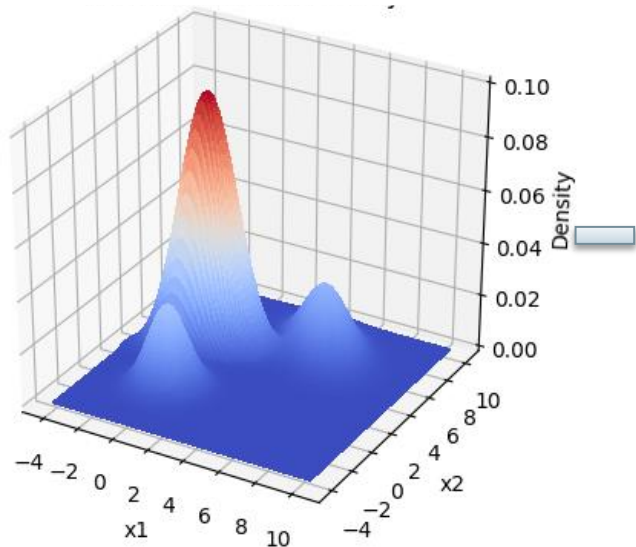
Epanechnikov kernel



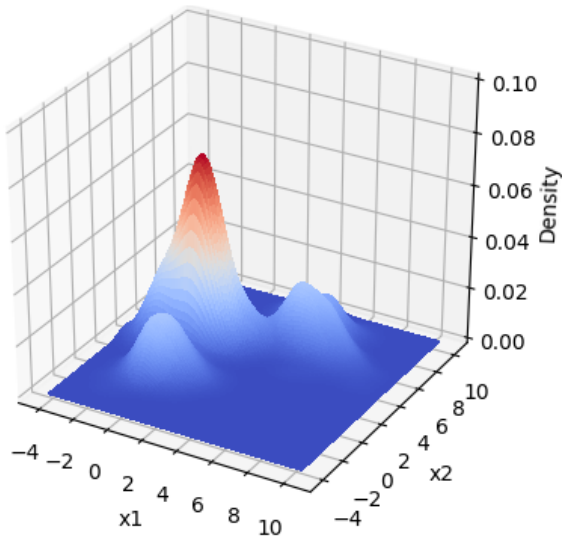
Рисунки 20, 21. Диаграммы поверхности и контуров плотности распределения, восстановленные с помощью ядра Epanechnikov kernel. Ширина окна по Сильверману.

Анализ смещения

Реальная плотность



Восстановленная плотность



Смещение

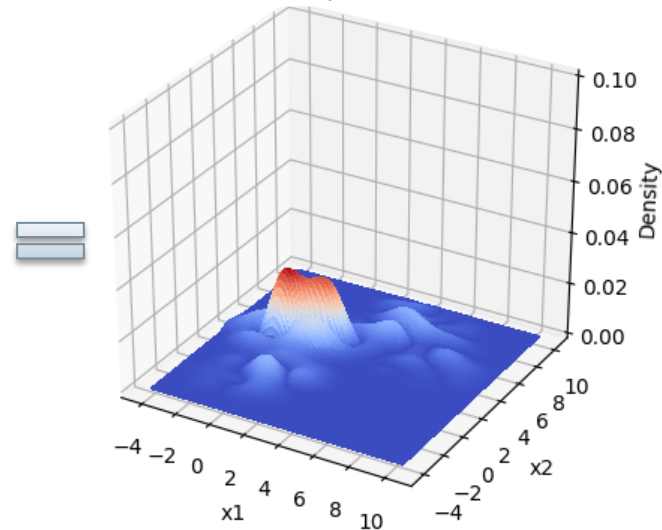
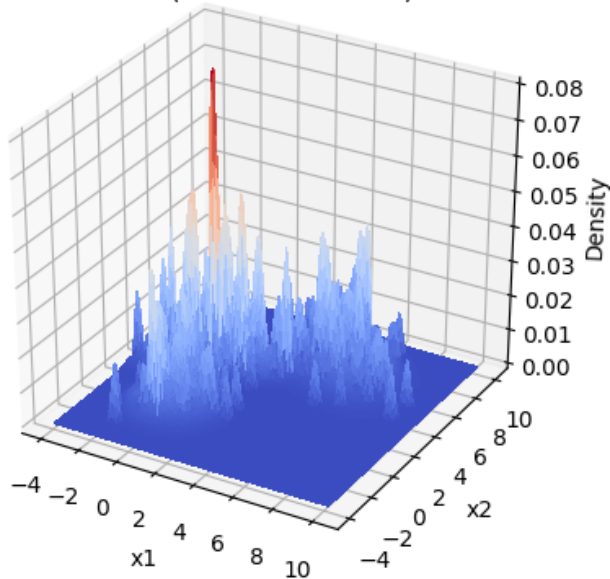


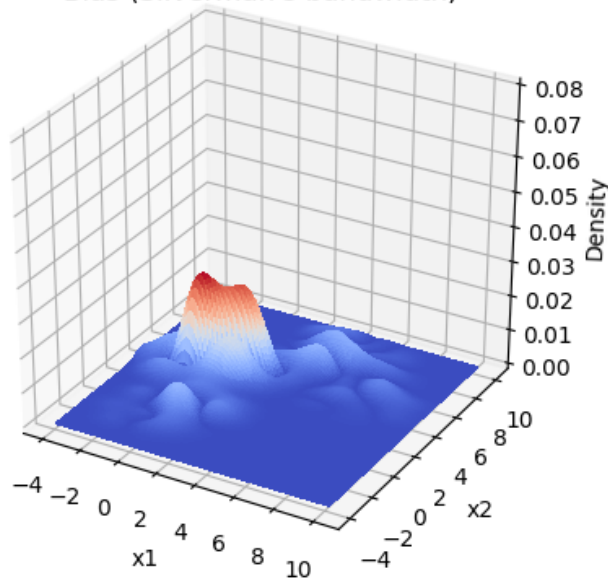
Рисунок 22. Вычисление смещения. Из плотности реального распределения вычитается (в данном случае по модулю) восстановленное распределение.

Смещение при разной ширине окна

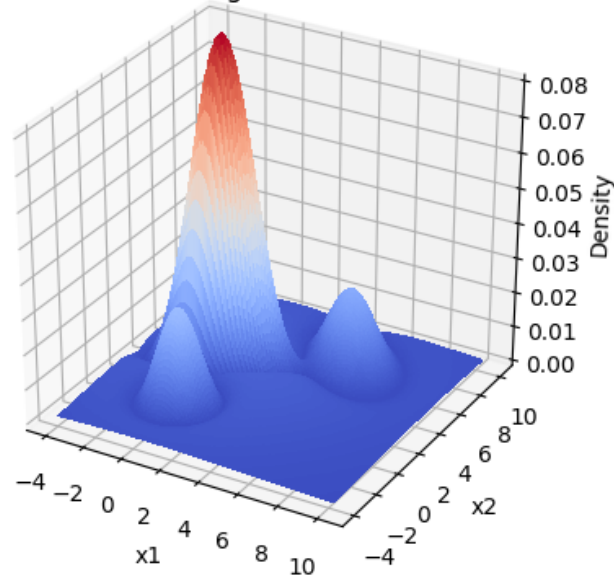
Bias (small bandwidth)



Bias (Silverman's bandwidth)

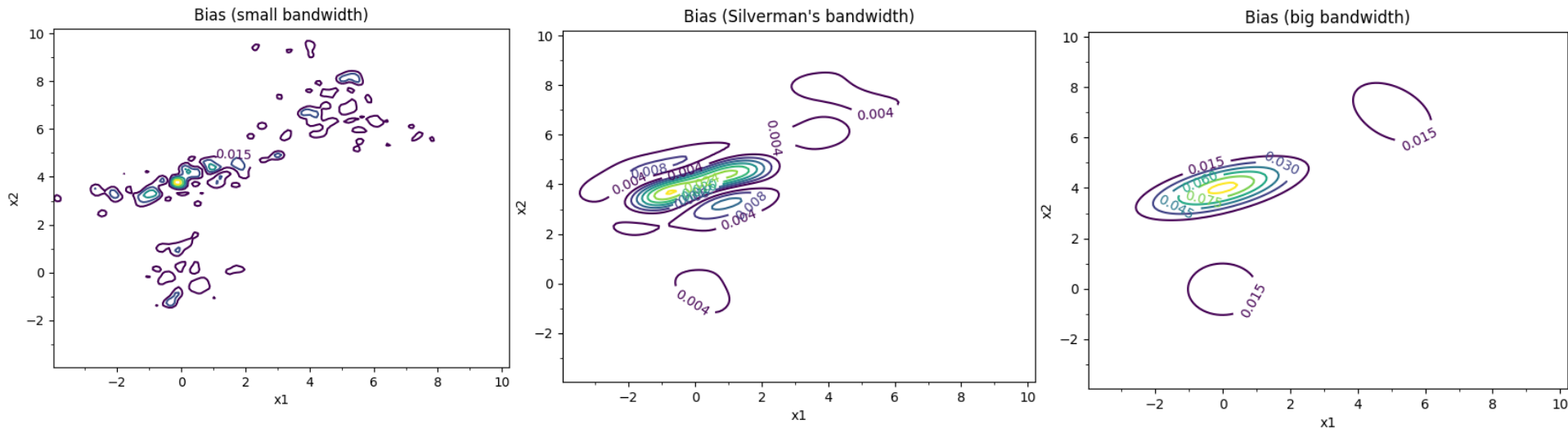


Bias (big bandwidth)



Рисунки 23-25. Смещения восстановленной плотности распределения с различными размерами окна. Слева - 0.15, 0.15 (слишком узкое), в центре – 0.69, 0.50 (по Сильверману), справа – 5.0, 5.0 (слишком широкое). Диаграммы поверхности.

Смещение при разной ширине окна



Рисунки 26, 27. Смещения восстановленной плотности распределения с различными размерами окна. Слева - 0.15, 0.15 (слишком узкое), в центре – 0.69, 0.50 (по Сильверману), справа – 5.0, 5.0 (слишком широкое). Диаграммы контуров.

Анализ дисперсии

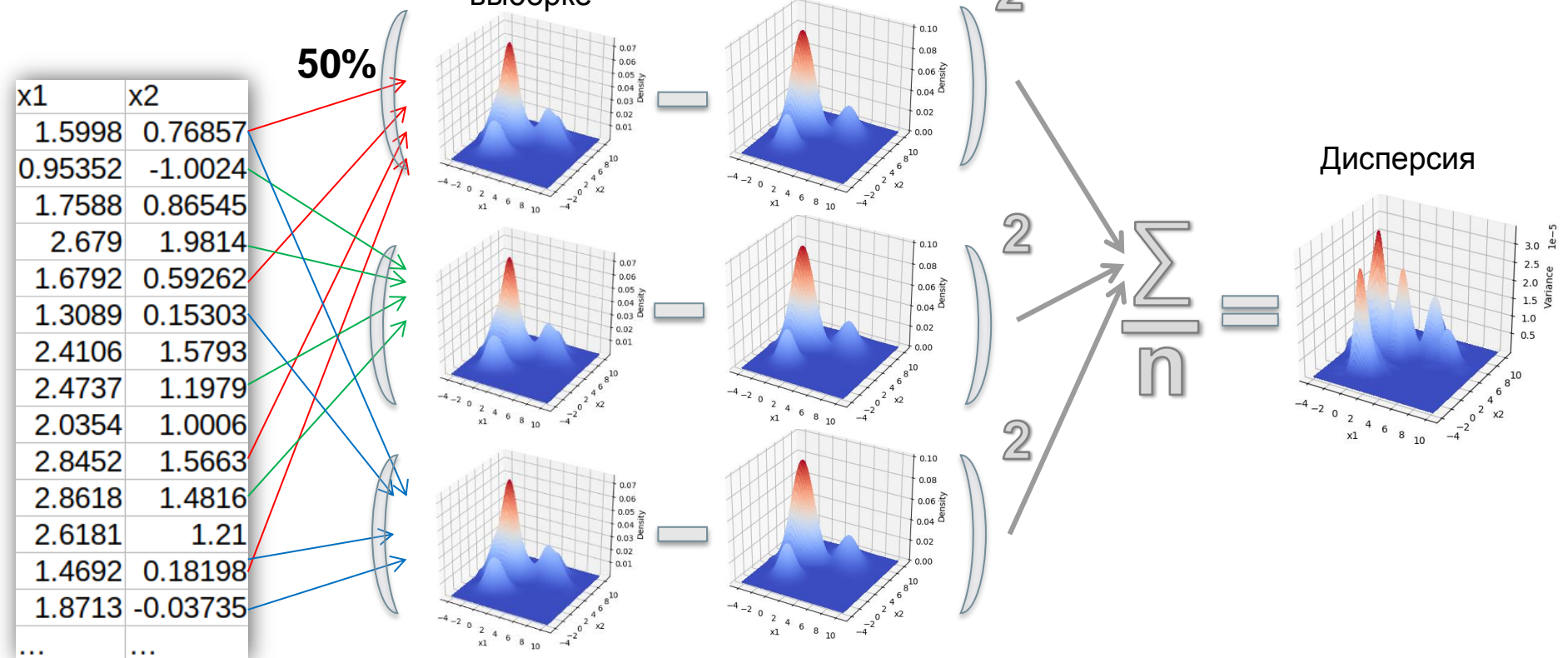


Рисунок 28. Визуализация алгоритма вычисления дисперсии.

Дисперсия (слишком узкое окно)

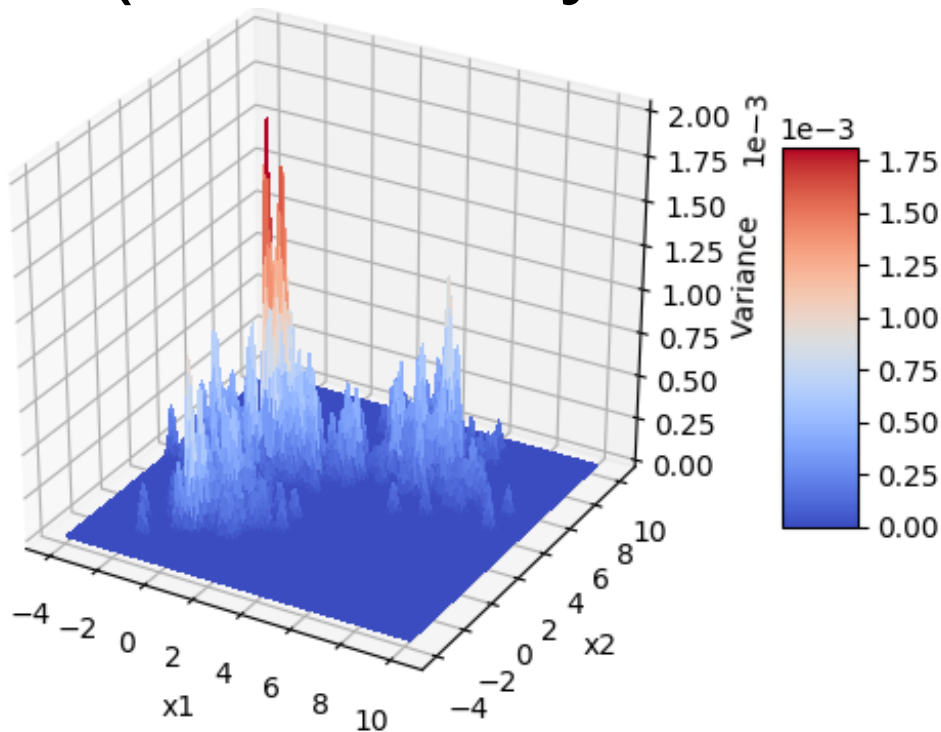


Рисунок 29. Дисперсия восстановленной плотности распределения со слишком узким размером окна – $(0.15, 0.15)$. 10 выборок по 50% данных.

Дисперсия (окно по Сильверману)

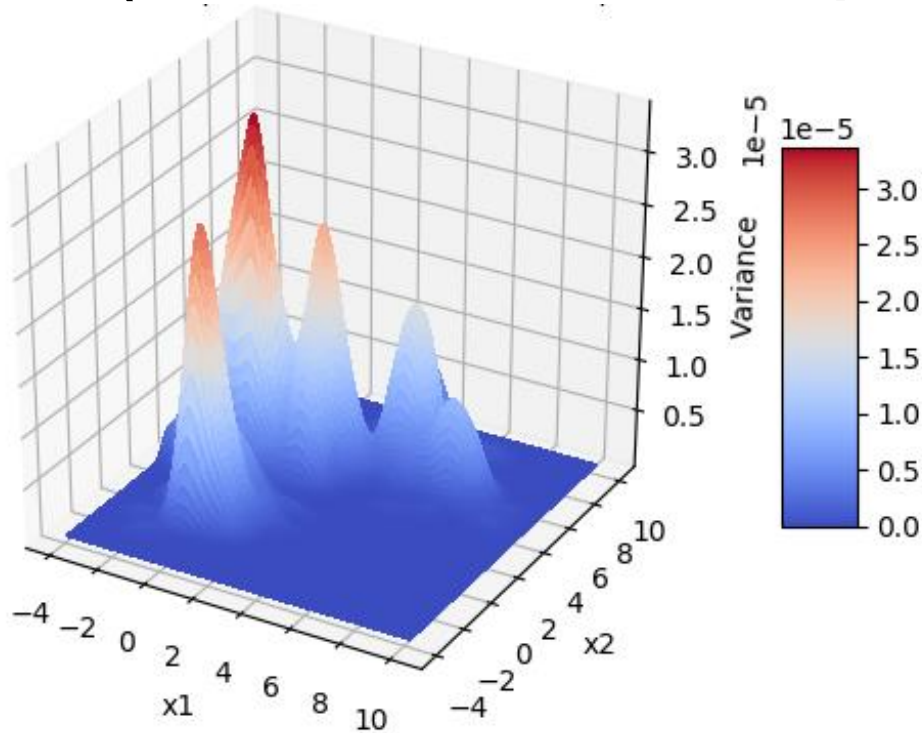


Рисунок 30. Дисперсия восстановленной плотности распределения с размером окна по правилу Сильвермана – $(0.69, 0.50)$. 10 выборок по 50% данных.

Дисперсия (слишком широкое окно)

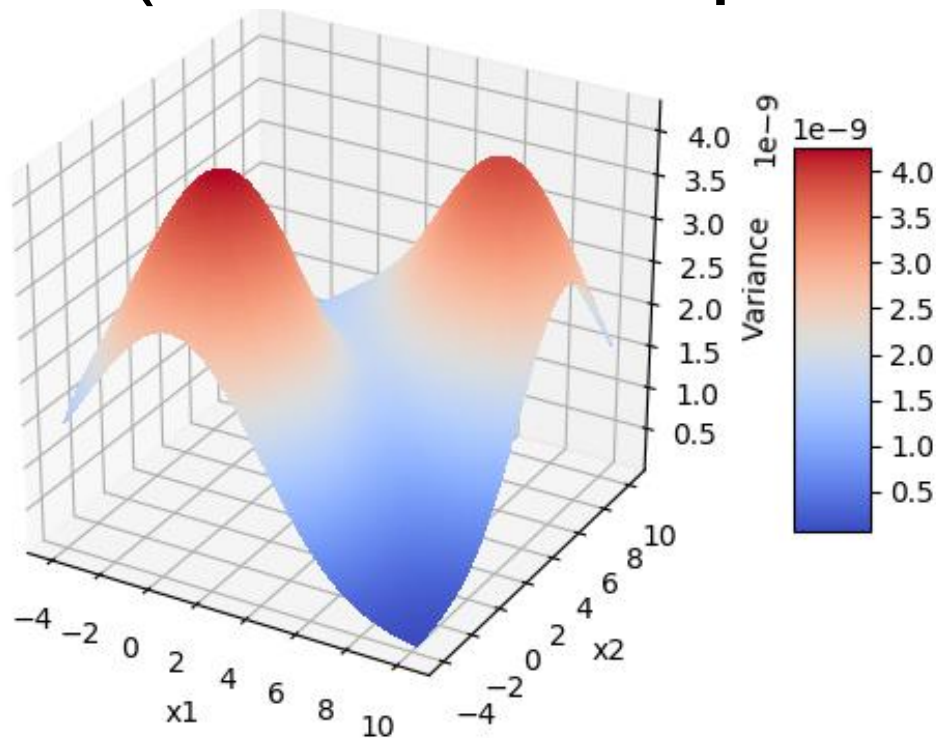


Рисунок 31. Дисперсия восстановленной плотности распределения со слишком широким размером окна – (5.0, 5.0). 10 выборок по 50% данных.

Анализ в фиксированных точках

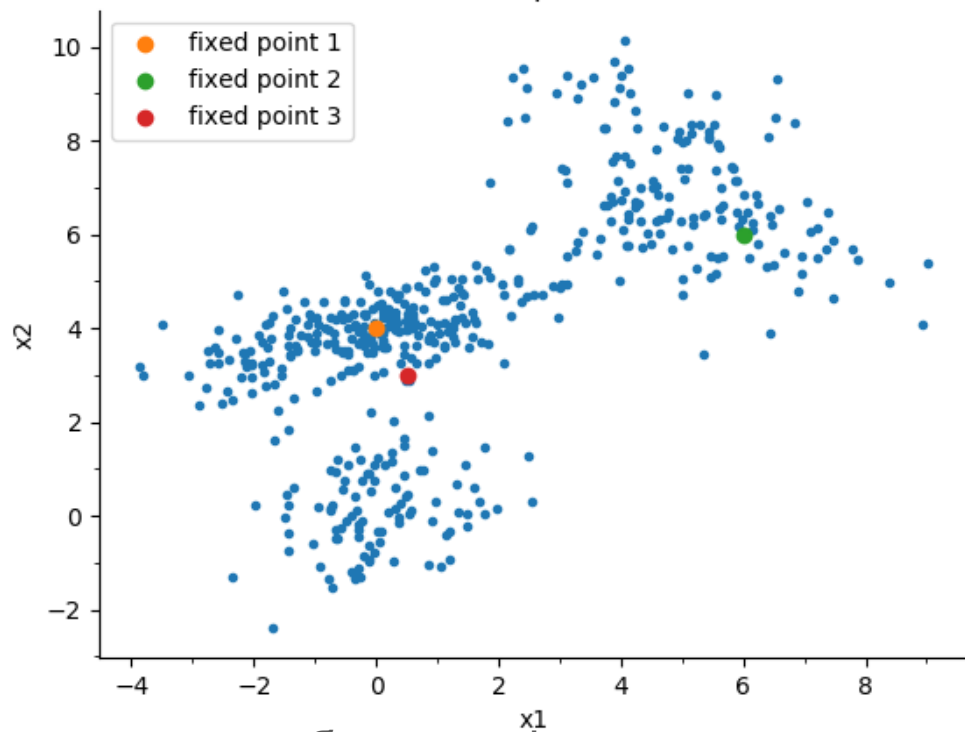


Рисунок 32. Визуализация выбранных фиксированных точек на фоне исходных данных на координатной плоскости

Смещение в фиксированных точках

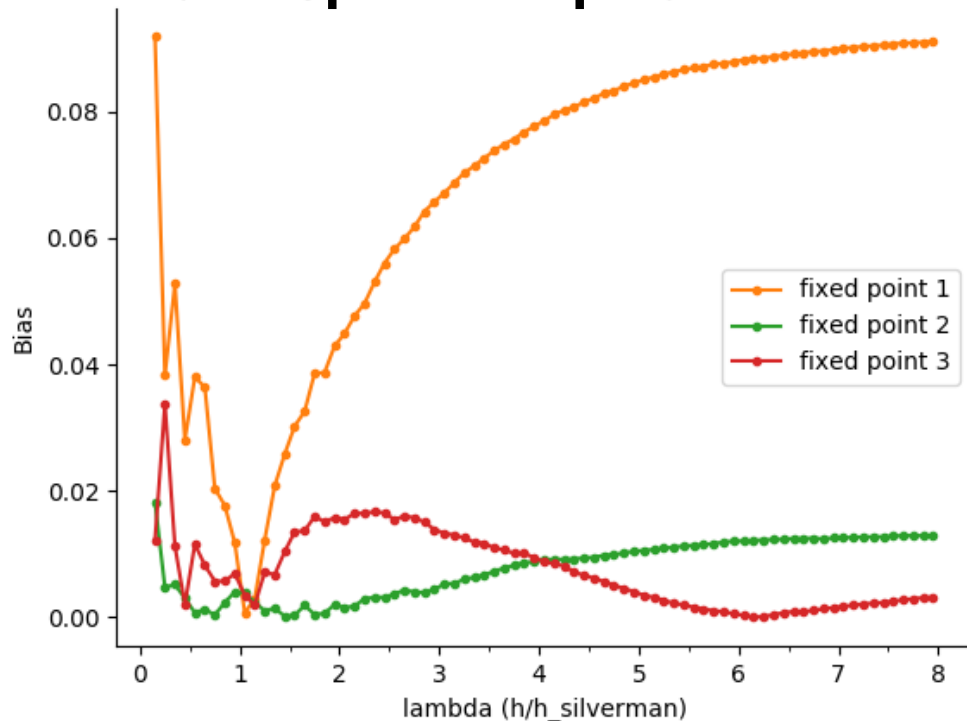


Рисунок 32. Зависимость смещения от коэффициента пропорциональности λ в различных фиксированных точках (Gaussian kernel)

Смещение в фиксированных точках

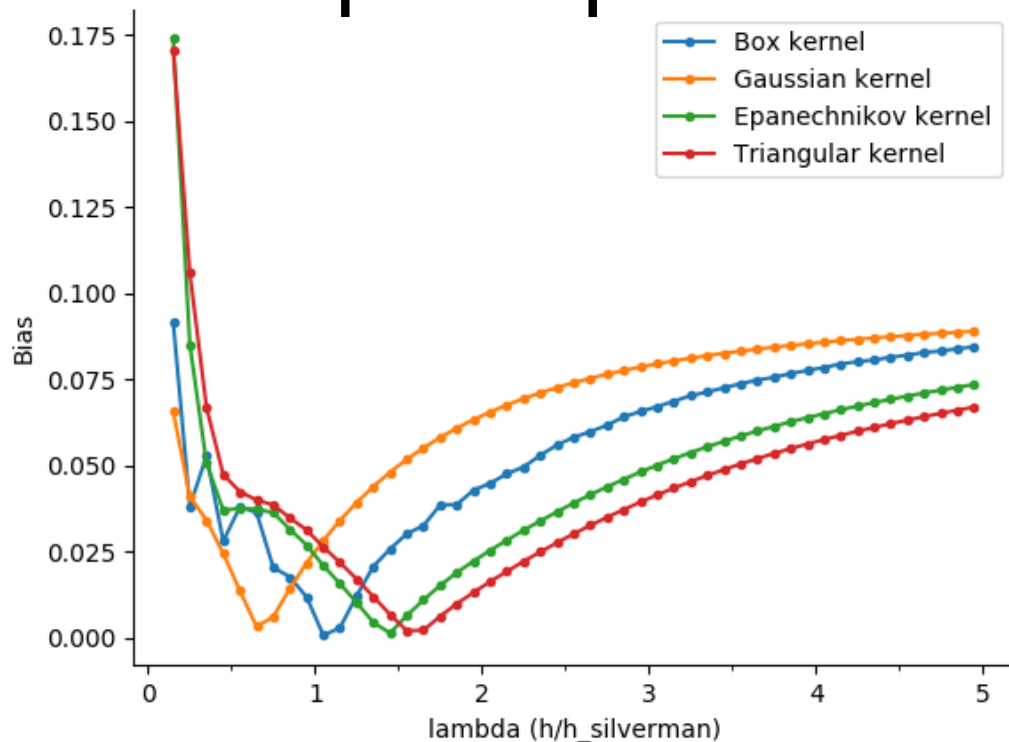
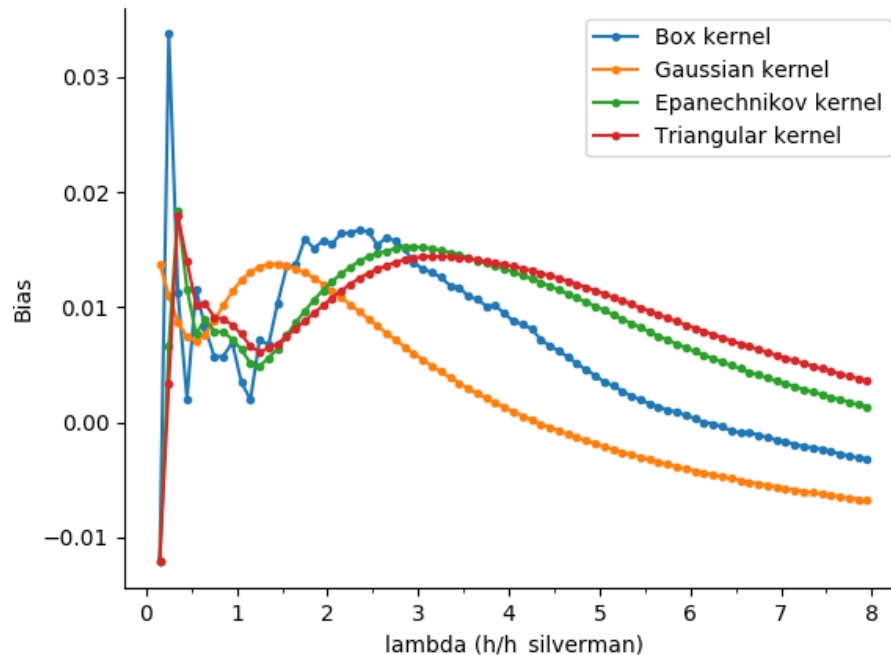
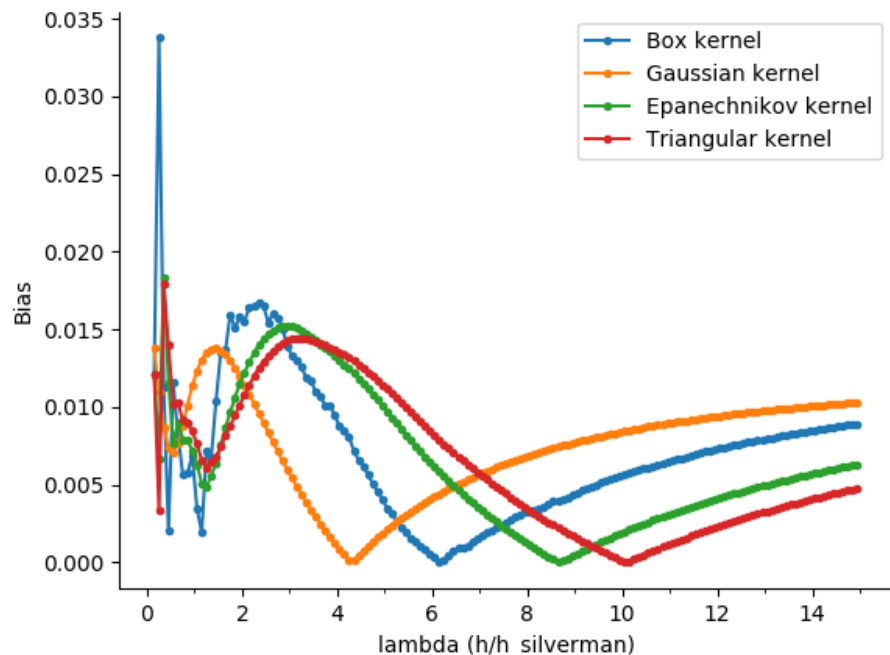


Рисунок 33. Зависимость смещения от коэффициента пропорциональности λ при различных типах ядра в фиксированной точке 1

Смещение в фиксированных точках



Рисунки 34, 35. Зависимость смещения от коэффициента пропорциональности λ при различных типах ядра в фиксированной точке 3. Слева смещение по модулю.

Дисперсия в фиксированных точках

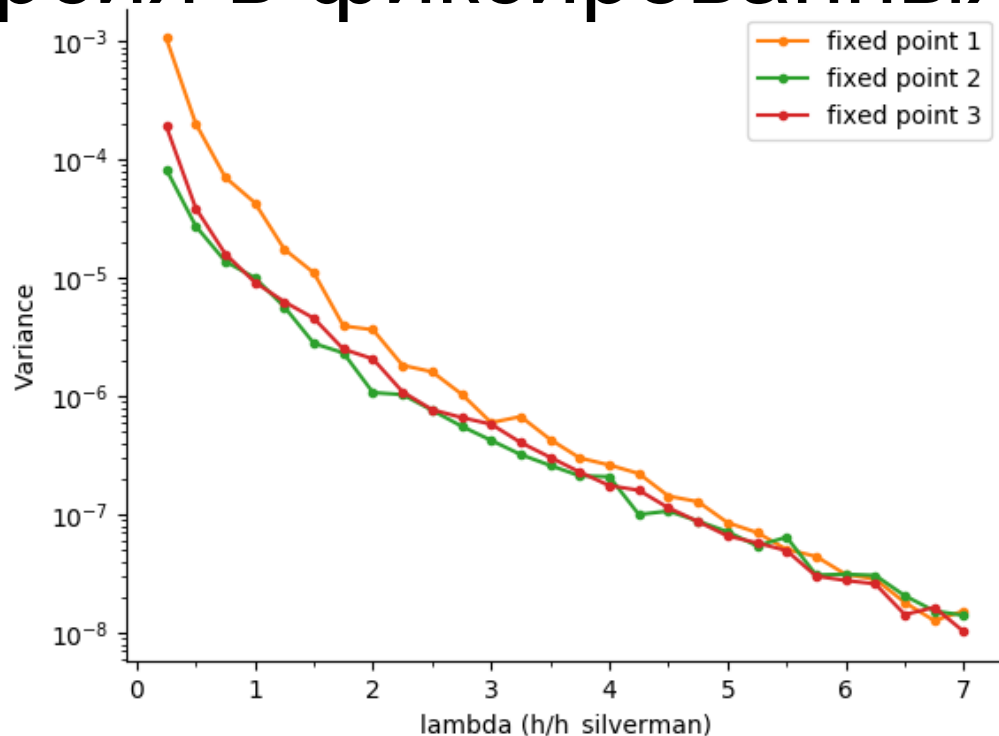


Рисунок 36. Зависимость дисперсии от коэффициента пропорциональности λ в различных фиксированных точках (Gaussian kernel)

Анализ MISE

$$MISE(\tilde{f}) = \int_{-\infty}^{\infty} \varepsilon^2(x) dx = \int_{-\infty}^{\infty} Var(\tilde{f}(x)) dx + \int_{-\infty}^{\infty} Bias^2(x) dx$$

Дисперсия в фиксированных точках

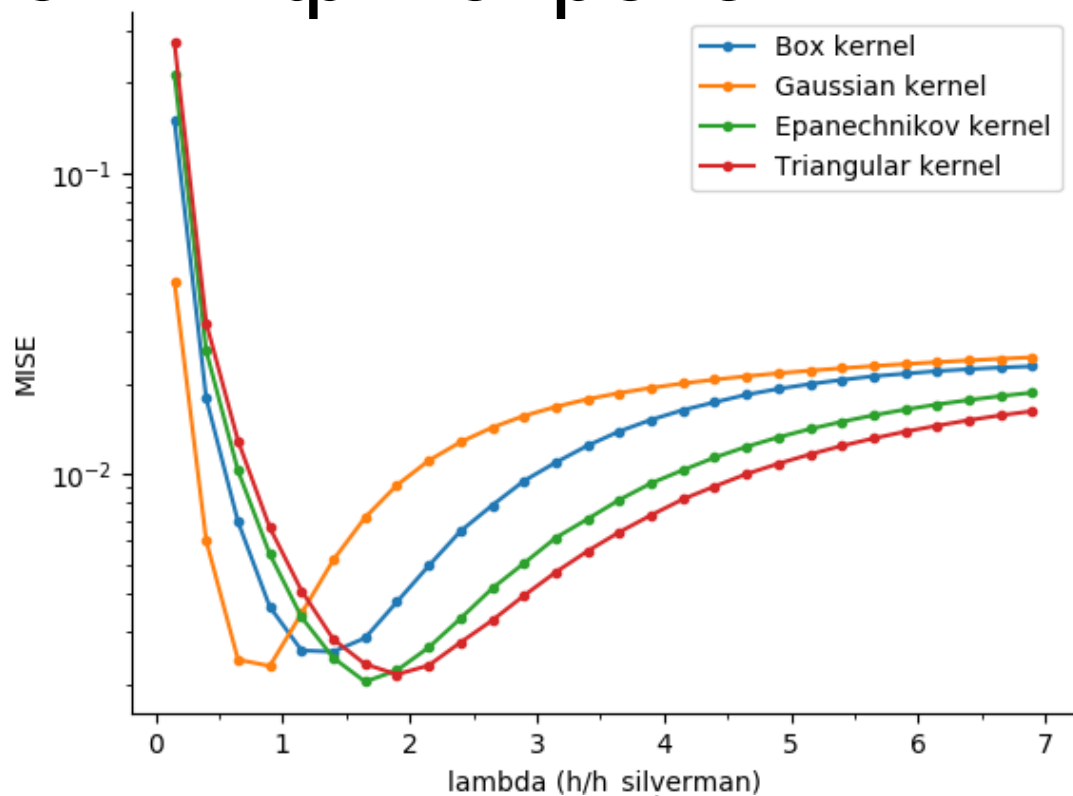


Рисунок 37. Зависимость средней интегральной ошибки восстановления плотности (MISE) от коэффициента пропорциональности λ при различных типах ядра в фиксированной точке 1

Выводы (ширина окна)

Для восстановления плотности распределения необходимо выбрать вид и ширину окна.

При увеличении ширины окна:

- смещение резко убывает, затем (немного не доходя до Сильвермана) начинает плавно расти
- дисперсия резко убывает примерно до Сильвермана, потом плавно убывает
- MISE убывает до Сильвермана, потом растёт

Оптимальную ширину окна в зависимости от вида окна следует искать в интервале $0.7h-1.5h$

Выводы (вид окна)

Наименьшая ошибка с шириной Сильвермана получилась у Box Kernel.

Gauss kernel - самый плавный

Epanichnekov & Triangular kernel - несколько лишних пиков

Выбор вида окна зависит от предполагаемого распределения данных.