Непараметрическое моделирование

ЦМФ

Содержание

- гистограммы
- ядерные оценки в одномерном случае
- ядерные оценки в многомерном случае

Гистограмма

Наиболее простая и широко используемая непараметрическая оценка плотности распределения

Пусть мы имеем выборку $(y_1, \dots, y_n)$. Разделим всю область определения случайной величины на несколько интервалов, тогда оценка плотности запишется в виде

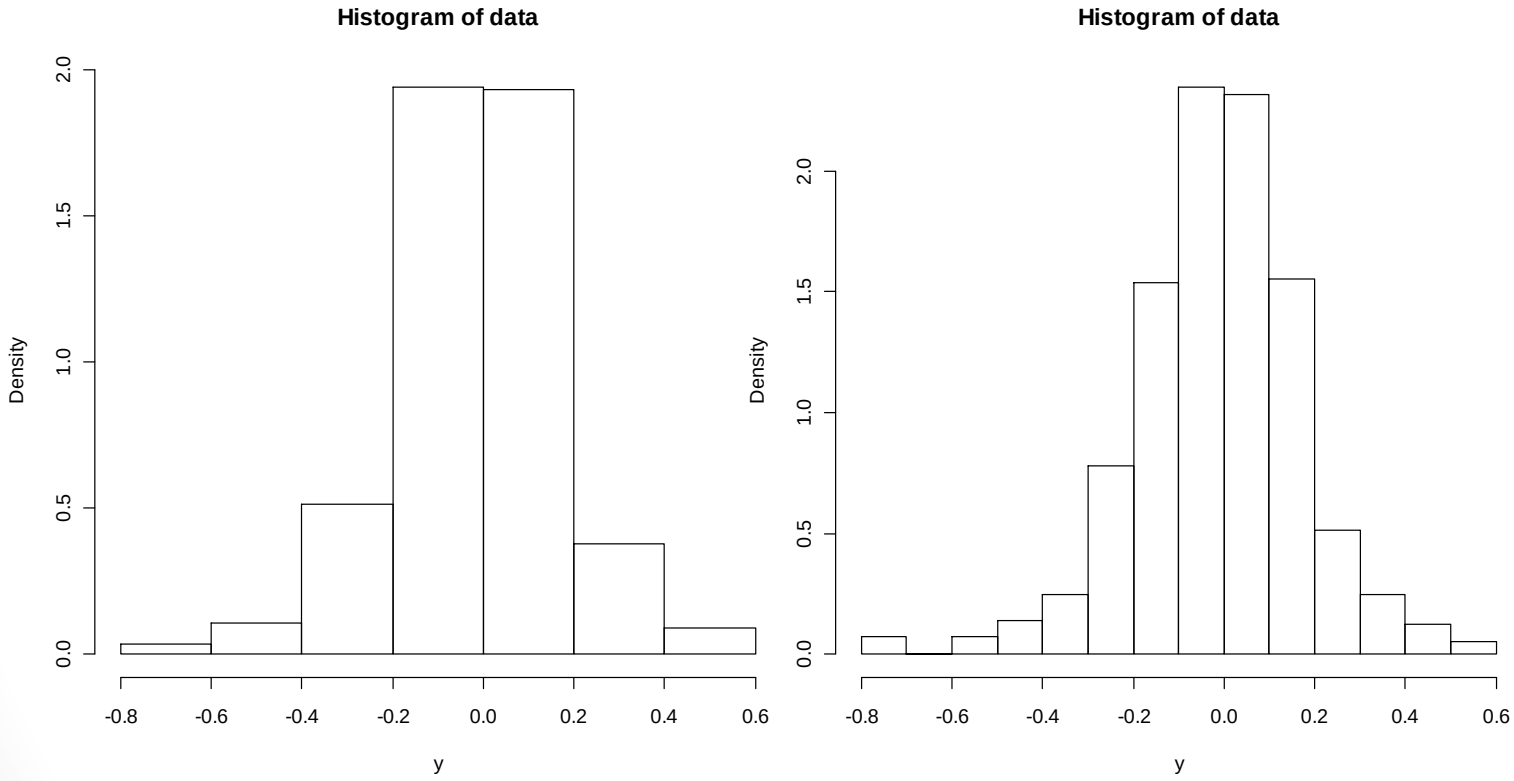
$$\hat{f}(y) = \frac{1}{n} \cdot \frac{\text{число наблюдений в интервале } y}{\text{длина интервала}} \quad (1)$$

Для построения гистограммы нужно определить:

1. Границы области определения;
2. Длину (количество) интервалов

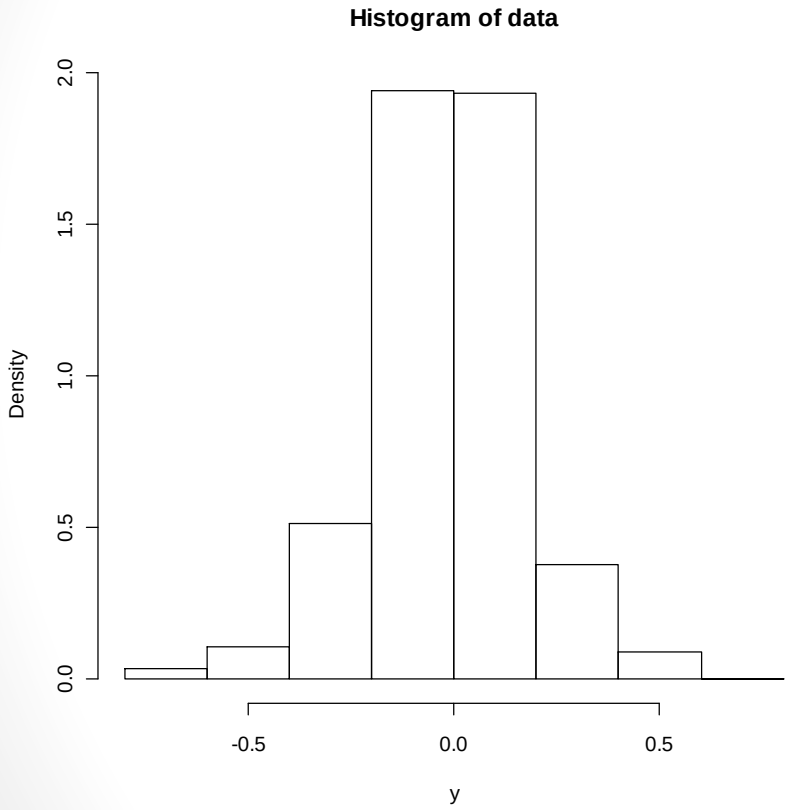
Параметры гистограммы

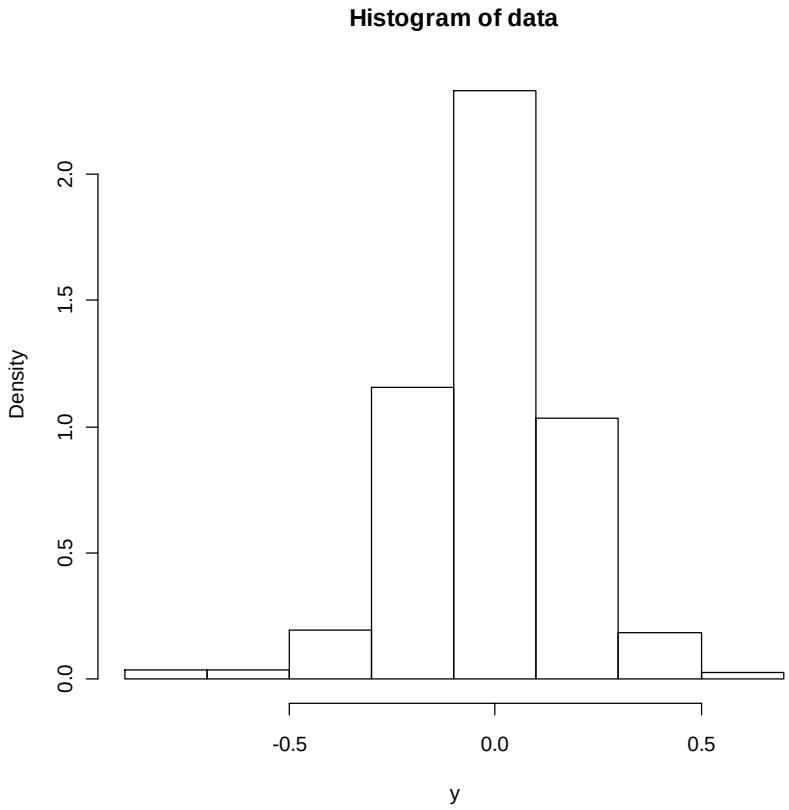
Длина интервалов влияет на детализацию гистограммы



Параметры гистограммы

Область определения может повлиять на форму





Оценка плотности распределения

Оценку (1) можно записать более формально. Пусть у нас есть m интервалов вида $(z_k; z_{k+1}),\ k \in \{1; \dots; m\},\ m < n.$ Все интервалы одинаковой длины $h = z_{k+1} - z_k$, тогда

$$\hat{f}(y) = \frac{1}{nh}\sum_{i=1}^{n} I(z_k < y_i \leq z_{k+1}),\ z_k < y \leq z_{k+1} \quad (2)$$

Длина интервала *h* должна быть достаточно большой, чтобы в него попало существенное количество наблюдений, и достаточно малой, чтобы не потерять важные детали распределения

Остаётся нерешенной проблема области распределения

Выход — ядерные оценки

Одномерный случай

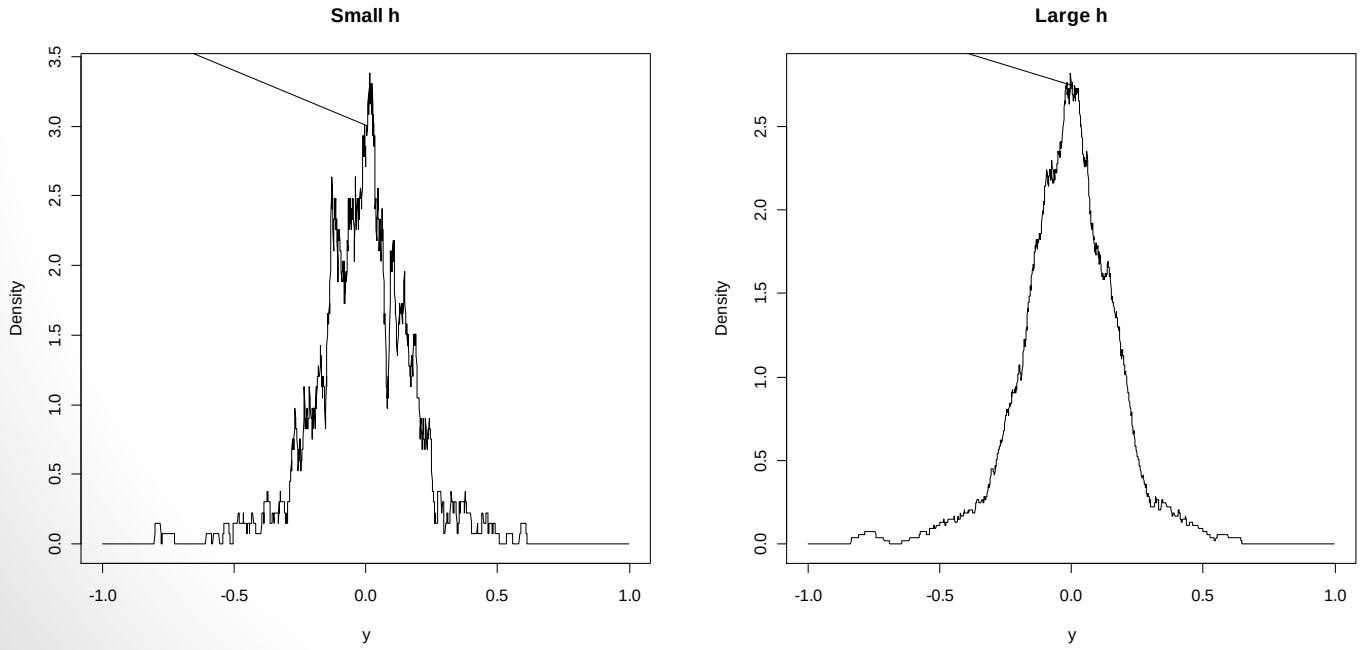
Простая непараметрическая оценка

Принцип построения простой (naïve) непараметрической оценки плотности в точке *y* состоит в подсчёте количества наблюдений, находящихся вблизи неё:

$$\hat{f}(y) = \frac{1}{nh}\sum_{i=1}^{n} I\left(y - \frac{h}{2} < y_i < y + \frac{h}{2}\right) \quad (3),$$

где h — длина интервала

Большие значения *h* дают более гладкие оценки:



Ядерная оценка

Простая оценка мало где дифференцируема. Чтобы понять это перепишем формулу (3) в следующем виде:

$$\hat{f}(y) = \frac{1}{nh}\sum_{i=1}^{n} w\left(\frac{y-y_i}{h}\right), \text{ где } w(x) = I\left(|x| < \frac{1}{2}\right) \quad (4)$$

Проблема заключается в функции *w(x),* которая придаёт наблюдениям дискретные веса (0 или 1)

Проблема решается с помощью замены функции *w(x)* на ядерную функцию *K(x)* с плавно изменяющимися весами:

$$\hat{f}(y) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{y-y_i}{h}\right) \quad (5)$$

Для того, чтобы оценка $\hat{f}(y)$ была функцией плотности, ядро должно удовлетворять условию $\int_{-\infty}^{+\infty} K(x)dx = 1$

Любая функция плотности удовлетворяет этому условию

Ядерные функции

В качестве ядерных функций обычно используются симметричные одномодальные функции плотности

Наиболее часто используемые на практике ядра:

$K_G(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \quad (6)$ — гауссовское ядро

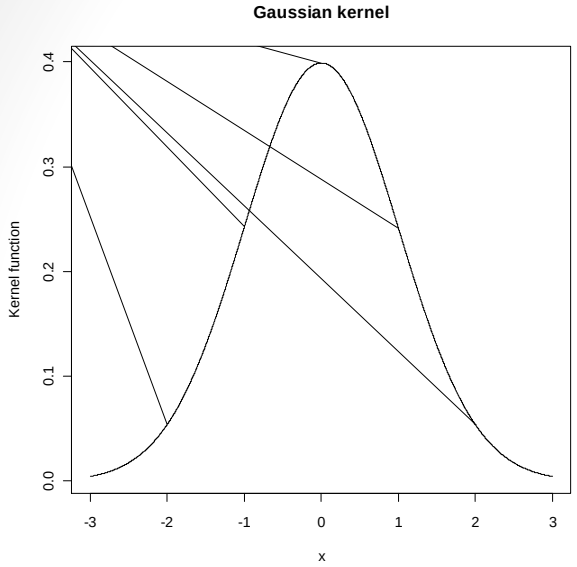
$K_E(x) = \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right) \cdot I\left(|x| < \sqrt{5}\right) \quad (7)$ — ядро Епанечникова

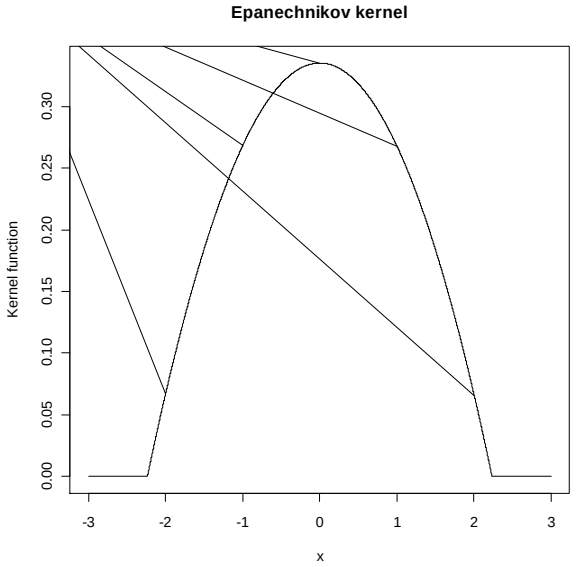
$K_T(x) = (1 - |x|) \cdot I(|x| < 1) \quad (8)$ — треугольное ядро

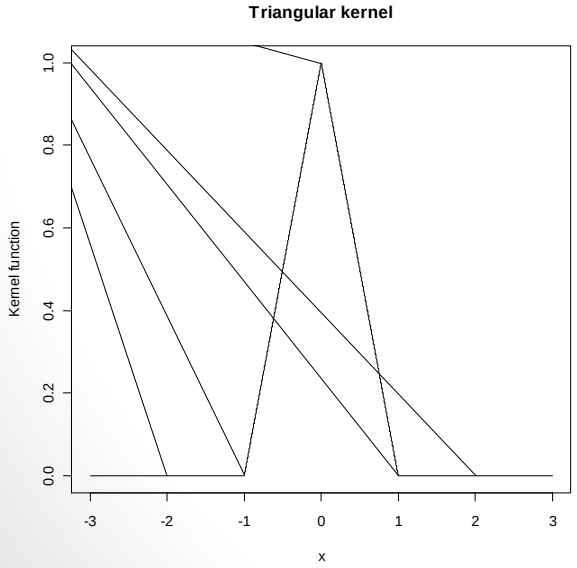
$K_U(x) = \frac{1}{2} I(|x| < 1) \quad (9)$ — прямоугольное (равномерное) ядро

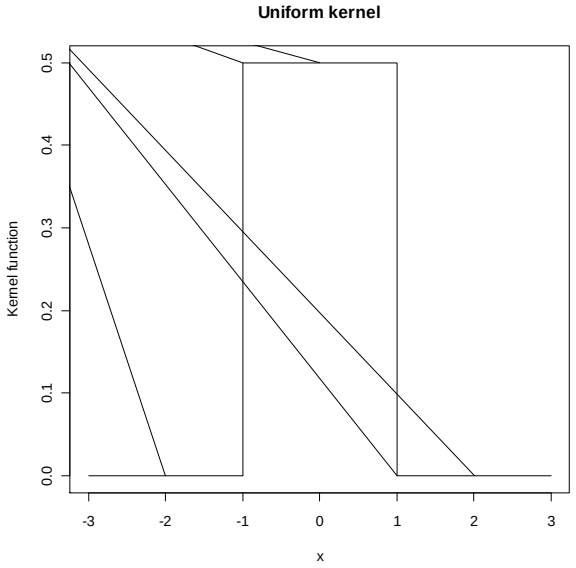
Вид этих функций представлен на следующем слайде

Ядерные функции



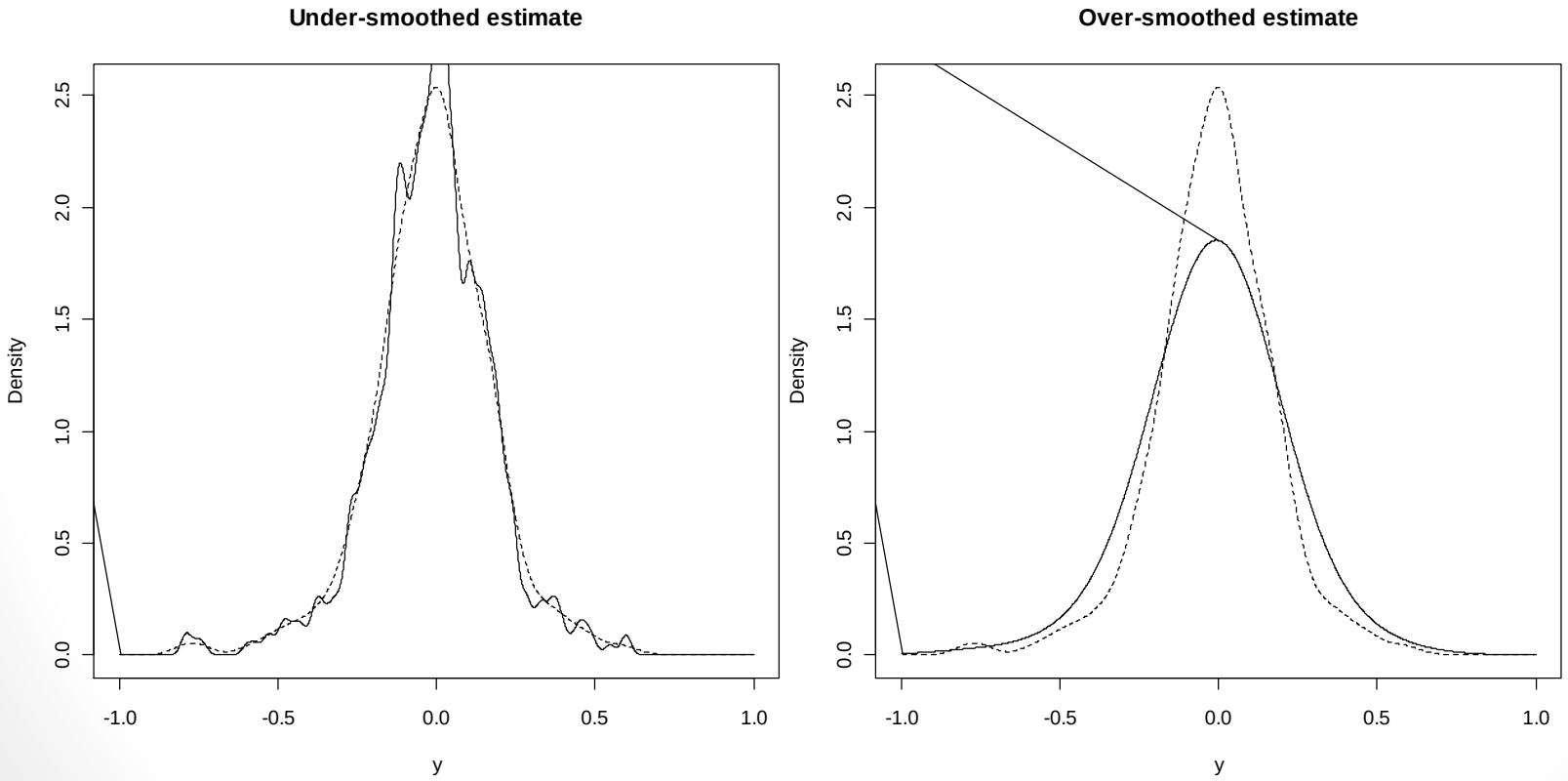






Влияние ширины интервала

Тогда как выбор ядра оказывает незначительное влияние на оценку плотности, выбор ширины интервала имеет решающее значение



Выбор ширины интервала

Существует два основных подхода к определению величины сглаживающего множителя (ширины интервала):

1. Фиксированная ширина интервала на всей выборке. В рамках этого подхода выделяют:
   - правило подстановки (rule of thumb);
   - метод перекрёстной проверки (cross-validation)
2. Ширина интервала меняется в зависимости от локальной концентрации наблюдений. Методы:
   - обобщённый метод ближайших соседей (generalized nearest neighbors);
   - адаптивный метод (adaptive nearest neighbors)

Фиксированная ширина интервала

ОДНОМЕРНЫЙ СЛУЧАЙ

Среднеквадратичная ошибка

Выбирать величину *h* следует так, чтобы оценка была как можно ближе к истинной плотности распределения, т.е. минимизировать разницу между $\hat{f}(y)$ и $f(y)$

Наиболее естественным кандидатом на эту разницу является среднеквадратичная ошибка (Mean Squared Error, MSE), рассчитываемая в конкретной точке *y*:

$$MSE(h, y) = E\left(\left(\hat{f}(y) - f(y)\right)^2\right) \quad (10)$$

Распишем выражение (10) подробнее:

$$MSE(h, y) = E\left(\left(\hat{f}(y) - f(y)\right)^2\right) = E\left(\hat{f}^2(y)\right) - 2f(y)E\left(\hat{f}(y)\right) + f^2(y)$$

$$= \left[E\left(\hat{f}^2(y)\right) - E^2\left(\hat{f}(y)\right)\right] + \left[E^2\left(\hat{f}(y)\right) - 2f(y)E\left(\hat{f}(y)\right) + f^2(y)\right] =$$

$$var\left(\hat{f}(y)\right) + \left(E\left(\hat{f}(y)\right) - f(y)\right)^2 \quad (11)$$

Дисперсия и смещение оценки

$$MSE(h, y) = var\left(\hat{f}(y)\right) + \left(E\left(\hat{f}(y)\right) - f(y)\right)^2 \quad (11)$$

Первое слагаемое выражения (11) соответствует дисперсии оценки, второе — квадрату её смещения

Если ширина интервала слишком большая, то оценка оказывается пересглаженной, и растёт смещение

Если значение *h* слишком маленькое, то это увеличивает дисперсию

Минимальное смещение достигается при максимальной дисперсии $(h = 0)$, а минимальная дисперсия — при максимальном смещении $(h \to +\infty)$

Нужно искать компромисс

Интегральная среднеквадратичная ошибка

Поскольку мы заинтересованы в минимизации отклонения между оценкой $\hat{f}(y)$ и плотностью $f(y)$ не только в конкретной точке *y*, рассмотрим интегральную среднеквадратичную ошибку (Mean Integrated Squared Error, MISE):

$$MISE(h) = E\left(\int_{-\infty}^{+\infty}\left(\hat{f}(y) - f(y)\right)^2 dy\right) \quad (12)^1$$

Мы можем переписать это так:

$$MISE(h) = \int E\left(\left(\hat{f}(y) - f(y)\right)^2\right) dy = \int MSE(h, y)dy \quad (13), —$$

или в следующем виде:

$$MISE(h) = \int var\left(\hat{f}(y)\right) dy + \int\left(E\left(\hat{f}(y)\right) - f(y)\right)^2 dy \quad (14)$$

[1] Далее вместо определённого интеграла по всей числовой оси будет использоваться неопределённый

Оптимальная ширина интервала

Минимизируя аппроксимацию к критерию *MISE*, обозначаемую *AMISE*, можно найти оптимальное значение параметра сглаживания:

$$h_{opt} = \left(\int x^2 K(x)dx\right)^{-\frac{2}{5}}\left(\int K^2(x)dx\right)^{\frac{1}{5}}\left(\int f''^2(y)\,dy\right)^{-\frac{1}{5}} n^{-\frac{1}{5}} \quad (15)$$

Замечания к формуле (15):

- h_{opt} стремится к нулю по мере роста объема выборки, но сравнительно медленно (по степенному закону);
- h_{opt} уменьшается, если $f(y)$ сильно варьируется, и возрастает, если функция плотности варьируется слабо;
- наиболее подходящее ядро $K(x)$ можно определить, исходя из значения критерия *MISE* (14)

Методы оценки оптимальной ширины интервала

В выражении для h_{opt} (15) остаётся неопределённость, связанная с незнанием истинной функции плотности $f(y)$

Мы рассмотрим два способа преодоления этой неопределённости:

1. Правило подстановки (Rule of Thumb);
2. Метод перекрёстной проверки (Cross-Validation)

Правило подстановки

Вместо $f(y)$ в выражение для оптимального интервала (15) подставляется какое-либо известное распределение

Если подставить нормальное распределение $N(\mu, \sigma^2)$ и использовать гауссовское ядро, то получим:

$$\hat{h}_{opt} \approx 1.059\sigma n^{-\frac{1}{5}} \quad (16)$$

В качестве оценки *σ* можно использовать выборочное стандартное отклонение, $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \frac{1}{n}\sum_{j=1}^{n} y_j\right)}$, или межквартильное расстояние, $\frac{\hat{q}_3 - \hat{q}_1}{1.349}$, где $\hat{q}_i$ — выборочное значение *i*-го квартиля, 1.349 — межквартильное расстояние для стандартного нормального распределения

Модифицированное правило подстановки

Правило подстановки хорошо работает тогда, когда истинный закон распределения близок к подставляемому

Существует также модифицированное правило подстановки:

$$\hat{h}_{opt} = 0.9 \min\left(\hat{\sigma}, \frac{\hat{q}_3 - \hat{q}_1}{1.349}\right) n^{-\frac{1}{5}} \quad (17)$$

Модифицированное правило является более устойчивым к отклонениям истинного распределения от нормального закона

Метод перекрёстной проверки

Мы опишем вариацию метода, основанную на наименьших квадратах

Идея состоит в рассмотрении интегральной квадратической ошибки (Integrated Squared Error, ISE), аналогичной критерию *MISE* (12), но без математического ожидания:

$$ISE(h) = \int \left(\hat{f}(y) - f(y)\right)^2 dy =$$

$$\int \hat{f}^2(y)dy - 2\int \hat{f}(y)f(y)dy + \int f^2(y)dy \quad (18)$$

Последнее слагаемое не зависит от *h* не играет роли в оптимизации

Величина $\int \hat{f}(y)f(y)dy$ есть матожидание оценки, которое приближённо равно $E\left(\hat{f}(y)\right) \approx \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{-i}(y_i)$, где $\hat{f}_{-i}(y_i)$ — оценка плотности по всем наблюдениям, кроме y_i

Метод перекрёстной проверки

Таким образом, оптимизационная задача сводится к минимизации выражения

$$CV(h) = \int \hat{f}^2(y)dy - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(y_i) \quad (19)$$

Достоинства методов с фиксированной шириной интервала:

- простота вычислений;
- интуитивная понятность;
- оценки обладают известными статистическими свойствами

Недостатки:

- пересглаженный центр распределения;
- недосглаженные и тонкие хвосты

Меняющаяся ширина интервала

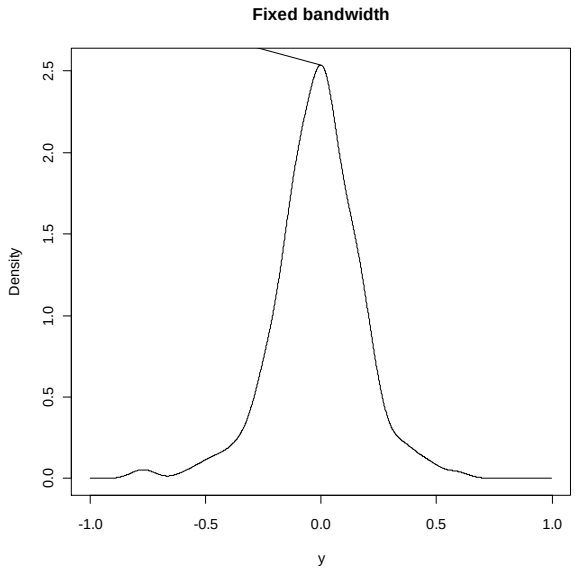
(адаптивные методы)

ОДНОМЕРНЫЙ СЛУЧАЙ

Адаптивные методы

Распределение данных может иметь различную концентрацию в центре и на хвостах, поэтому логично использовать широкий интервал *h* там, где они расположены редко (на хвостах), и меньший — в зонах высоких концентраций (в центре)

Ядерные оценки с постоянной шириной интервала в случае гетерогенной концентрации данных пересглаживают распределение в центре и недосглаживают на хвостах:



Адаптивный метод ближайших соседей

Оценка строится в следующем виде:

$$\hat{f}(y) = \frac{1}{n}\sum_{i=1}^{n} \frac{K\left(\frac{y-y_i}{hd_k(y_i)}\right)}{hd_k(y_i)}, \quad (20)$$

$d_k(y_i)$ — расстояние от точки y_i до k-го ближайшего к ней наблюдения

Сглаживающий параметр разделяется на две части:

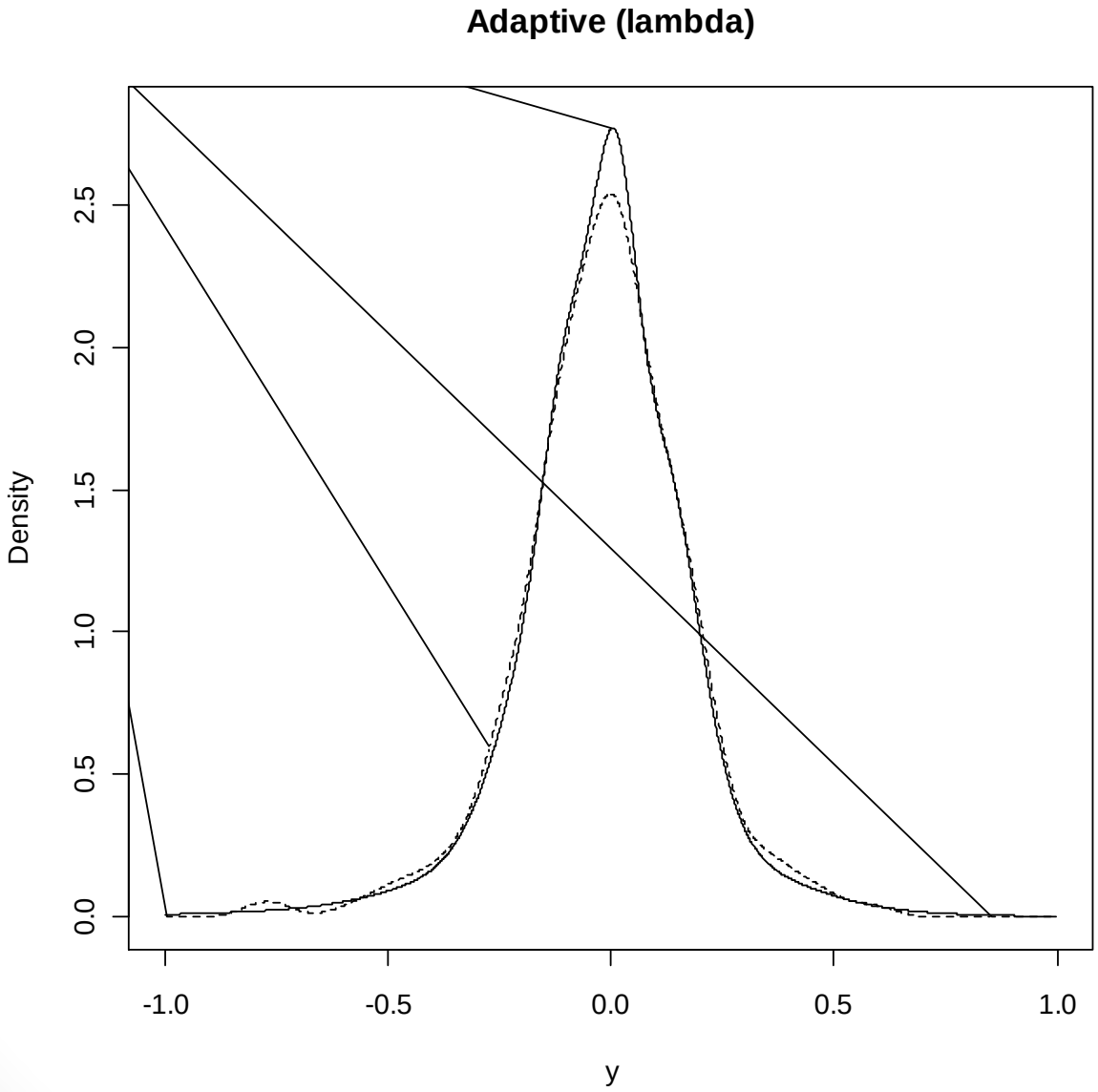
1. глобальная (*h*);
2. локальная концентрация наблюдений ($d_k(y_i)$)

Величину *h* определяют путём построения пилотной оценки плотности $\tilde{f}(y)$ с фиксированной шириной интервала

Часто вместо $d_k(y_i)$ используют показатель

$$\lambda_i = \left(\frac{g}{\tilde{f}(y_i)}\right)^{\alpha}, \; g = \left(\prod_{i=1}^{n} \tilde{f}(y_i)\right)^{\frac{1}{n}}, \; \alpha \in [0;1] \quad (21)$$

Сравнение двух методов



Практическая часть

Построение непараметрических оценок плотности

в программной среде «R»

cran.r-project.org

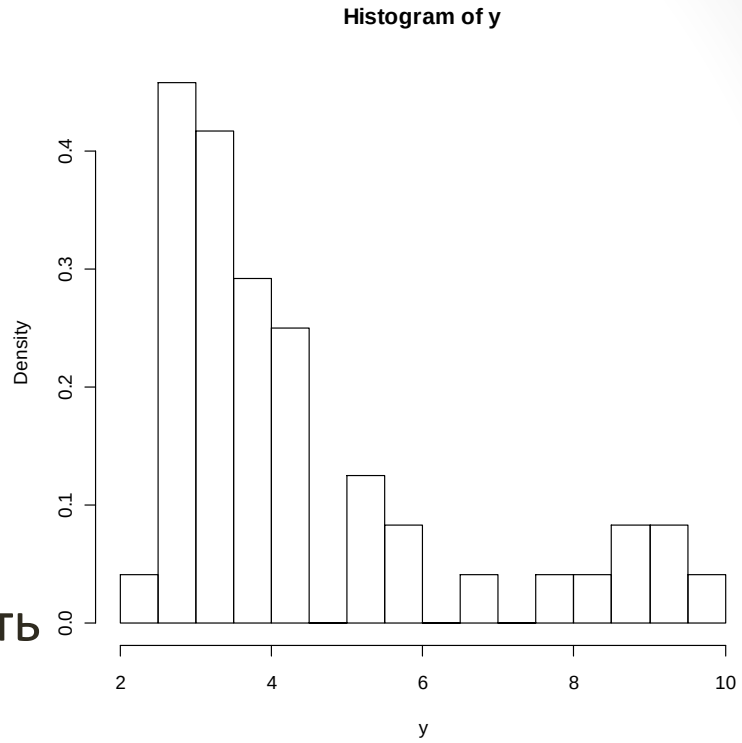
Пример 1. Острова

```
library(datasets)
y <- log(islands)
```

Построение гистограммы

```
hist(y,nclass=12,probability=TRUE)
```

- ***nclass*** определяет количество интервалов
- ***probability*** преобразует количество наблюдений в интервале в плотность распределения



С помощью дополнительного параметра ***breaks=c(y_1,…,y_k)*** задаётся разбиение на интервалы

Пример 1. Острова

Простая непараметрическая оценка плотности

```
L <- 10^4; N <- length(y)
h <- 2   # ширина интервала

# в точках x будет оцениваться плотность
x <- seq(0,12,length=L)   # последовательность 0 – 12 длиной L
f.naive <- numeric()      # нулевой (пока) вектор оценок

# считаем количество элементов в интервалах x_i ± h/2
for (i in 1:L) f.naive[i] <- sum(1*((y>x[i]-h/2)&(y<x[i]+h/2)))
f.naive <- f.naive/(N*h)  # нормируем оценку
```

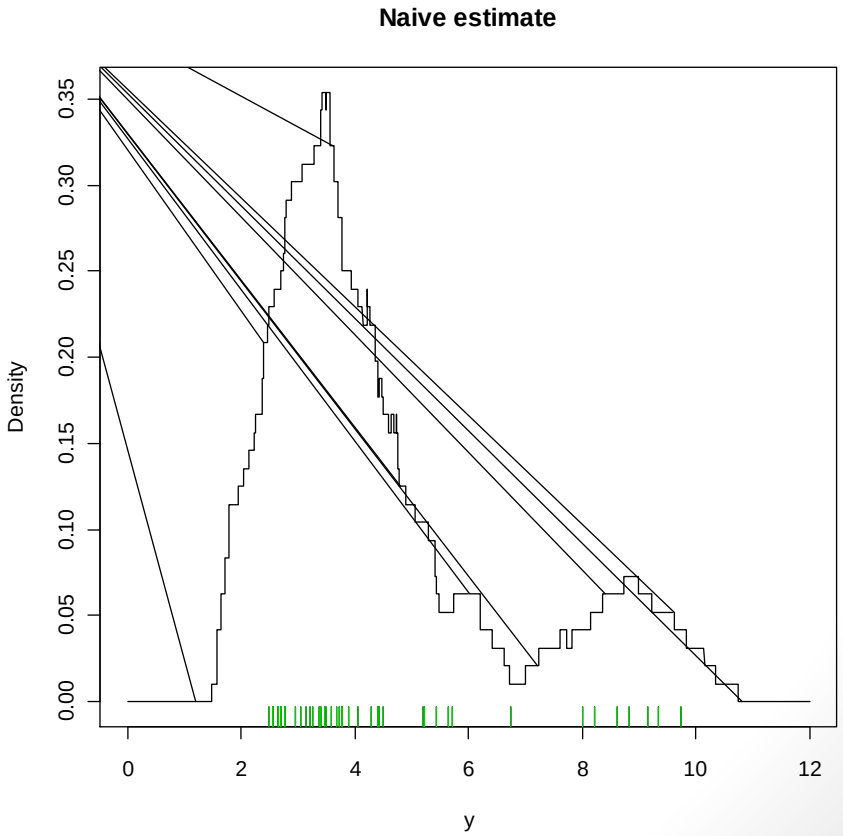
Пример 1. Острова

График простой оценки

```
plot(x,f.naive,type="l",main="Naive estimate",
xlab="y",ylab="Density")
rug(y,col=3)
```

- ***type*** определяет вид графика

  "l" — линии, "p" — точки, …
- ***main*** — заголовок
- ***xlab*** — подпись на оси x
- ***ylab*** — подпись на оси y



Пример 1. Острова

Ядерные оценки

```
library(np)
f.fix <- npudens(tdat=y,edat=x,
  ckertype="gaussian",bwtype="fixed")
```

- ***tdat*** — обучающая выборка
- ***edat*** — точки, в которых рассчитывается оценка
- ***ckertype*** — вид ядерной функции
  "gaussian", "epanechnikov", "uniform"
- ***bwtype*** определяет метод расчёта интервала *h*
  "fixed", "generalized_nn", "adaptive_nn"
- ***f$dens*** — искомые значения оценок

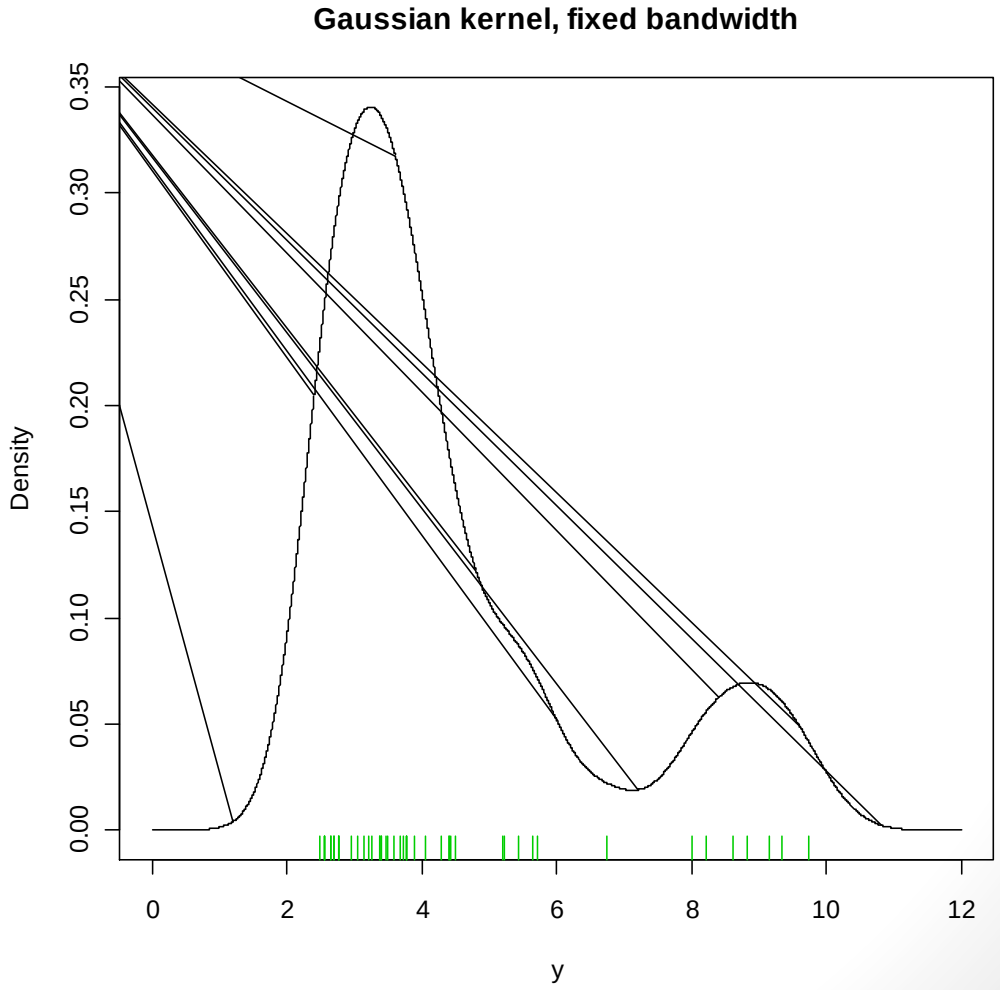
Пример 1. Острова

Адаптивный метод с λ_i

```
pilot <- npudens(tdat=y,ckertype="gaussian",bwtype="fixed")
h <- pilot$bws$bw   # оценка глобальной составляющей интервала

# среднегеометрическое пилотных оценок
g <- 1
for (i in 1:N) g <- g*pilot$dens[i]^(1/N)

# расчёт локальной концентрации наблюдений
alpha <- 0.5
lambda <- (g/pilot$dens)^alpha

kern <- function(u)  exp(-u^2/2)/sqrt(2*pi)   # ядро Гаусса

# расчёт оценок плотности
f <- numeric(L)
for (i in 1:L)  {
  for (j in 1:N)  f[i] <- f[i] + kern((x[i]-
       y[j])/(h*lambda[j]))/(h*lambda[j])
  f[i] <- f[i]/N
}
```

Пример 1. Острова

```
plot(x,f.fix$dens,type="l",
main="Gaussian kernel, fixed bandwidth",
xlab="y",ylab="Density")
```

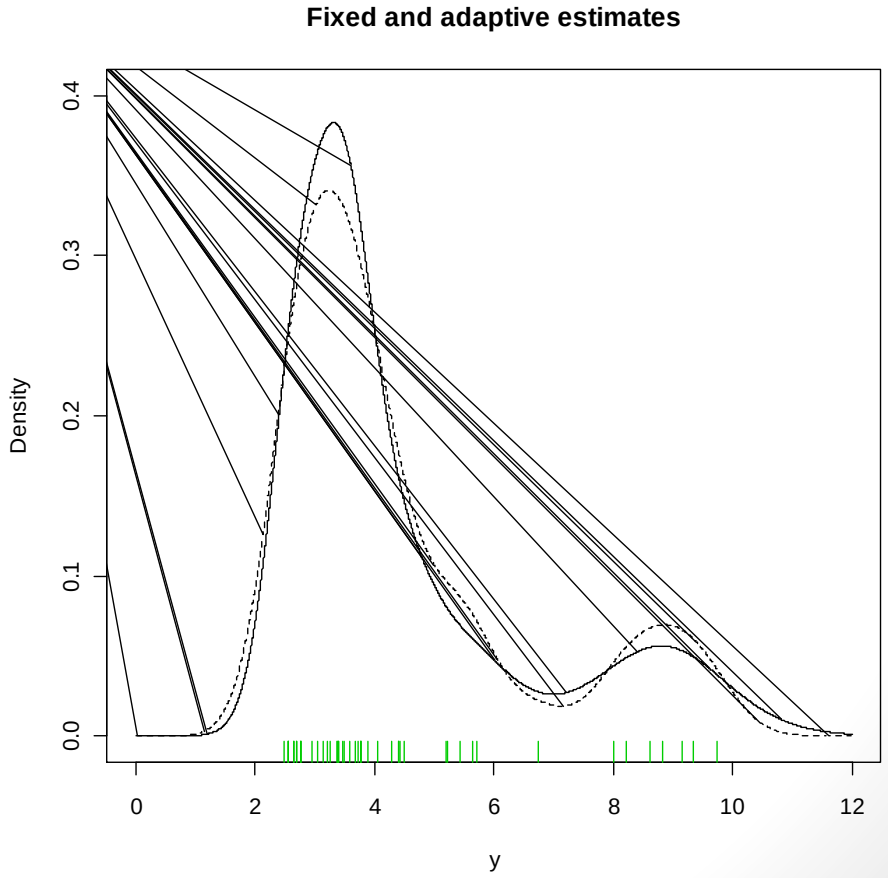


Пример 1. Острова

Сравнение адаптивной и фиксированной оценок

```
plot(x,f.fix$dens,type="l",lty="dashed",ylim=c(0,0.4),
main="Fixed and adaptive estimates",
xlab="y",ylab="Density")

lines(x,f)
```

- ***lty*** — тип линии
  "solid", "dashed", "dotted", "dotdash", "longdash", …
- ***ylim*** — границы по оси ординат
- ***lines*** — добавление кривых на существующий график



Пример 1. Острова

Значения логарифмической функции правдоподобия:

$\ln L = \sum_{i=1}^{n} \ln \hat{f}(y_i)$, — сумма логарифмов оценок в точках y_i

У нас есть оценки в точках x_j — последовательности 0 – 12

Пусть $dx = x_j - x_{j-1}$, найдём такой индекс j, что $x_j = y_i$:

$$x_1 + (j-1)dx = y_i \Rightarrow j = \frac{y_i - x_1}{dx} + 1$$

```
dx <- x[2]-x[1]
llh.fix <- sum(log(f.fix$dens[round((y-x[1])/dx)+1]))
llh.ada <- sum(log(f[round((y-x[1])/dx)+1]))

llh.fix; llh.ada   # вывод результатов на экран
```

llh.fix	-81.17
llh.ada	-81.29

Пример 1. Острова

Нахождение квантилей оценки распределения, $\hat{F}^{-1}(\alpha)$

```
# оценка функции распределения
F.fix <- npudist(tdat=y,edat=x,ckertype="gaussian",bwtype="fixed")

# для адаптивного варианта
F <- rep(0,times=L)
for (i in 1:L)  F[i] <- sum(f[1:i])*dx

# поиск квантиля методом деления пополам
alpha <- 0.99
a <- 1; b <- L; ab <- trunc((a+b)/2)
while ((b-a)>2) {
  if  (F.fix$dist[ab]<=alpha)  a <- ab
  if  (F.fix$dist[ab]>=alpha)  b <- ab
  ab <- trunc((a+b)/2)
}
q.fix <- x[ab]
```

q.fix	10.00
q.ada	10.49

Пример 1. Острова

Генератор случайных чисел

```
# фиксированный интервал
M <- 10^6
y.fix.sim <- sample(x,prob=f.fix$dens,size=M,replace=TRUE)
q.fix <- sort(y.fix.sim)[alpha*M]

# для адаптивного варианта
y.ada.sim <- sample(x,prob=f,size=M,replace=TRUE)
q.ada <- sort(y.ada.sim)[alpha*M]
```

q.fix	10.01
q.ada	10.46

Домашнее задание

В файле «y.csv» содержаться значения некоей случайной величины. Вашей задачей является построение оценок плотности распределения этой величины в точках из файла «x.csv»

Ответы принимаются на

https://kaggle.com/join/cmf_np

Многомерный случай

Оценки плотности

Простая оценка плотности в двумерной точке (y_1, y_2):

$$\hat{f}(y_1, y_2) = \frac{1}{nh^2} \sum_{i=1}^{n} \left(I\left(y_1 - \frac{h}{2} < y_{i,1} < y_1 + \frac{h}{2}\right) \cdot I\left(y_2 - \frac{h}{2} < y_{i,2} < \right.\right.$$

Двумерные ядерные функции

Ядро Гаусса:

$$K_G(x_1, x_2) = \frac{1}{2\pi} \exp\left(-\frac{x_1^2 + x_2^2}{2}\right) \quad (29)$$

Двумерное ядро Гаусса в точности равно произведению двух одномерных:

$$K_G(x_1, x_2) \equiv K_G(x_1) \cdot K_G(x_2)$$

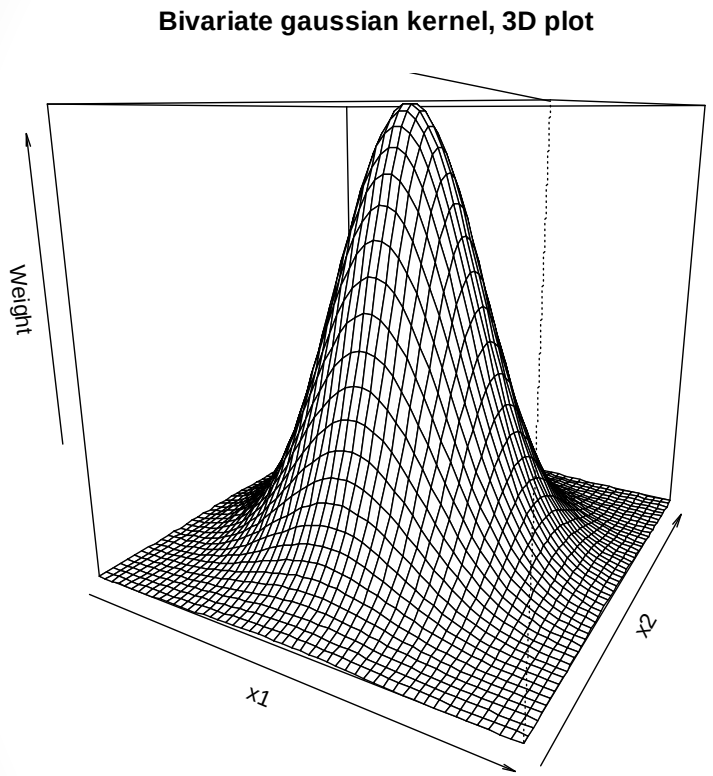
Ядро Епанечникова:

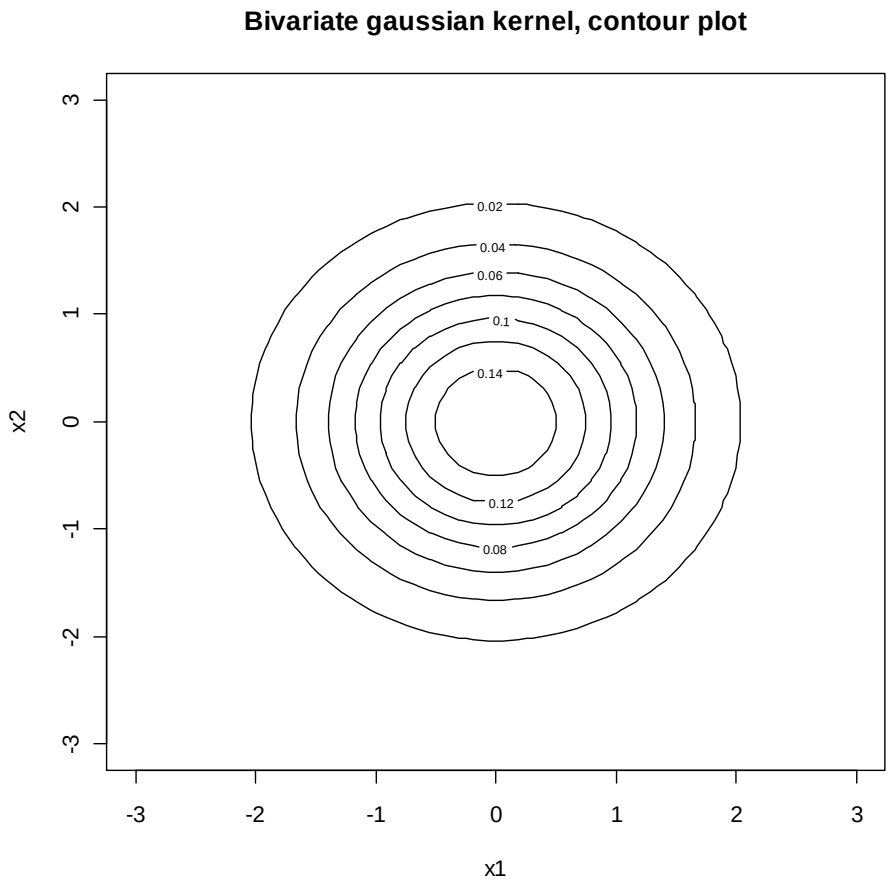
$$K_E(x_1, x_2) = \frac{2}{\pi}(1 - x_1^2 - x_2^2) \cdot I(x_1^2 + x_2^2 < 1) \quad (30)$$

Двумерное ядро не равно произведению двух одномерных:

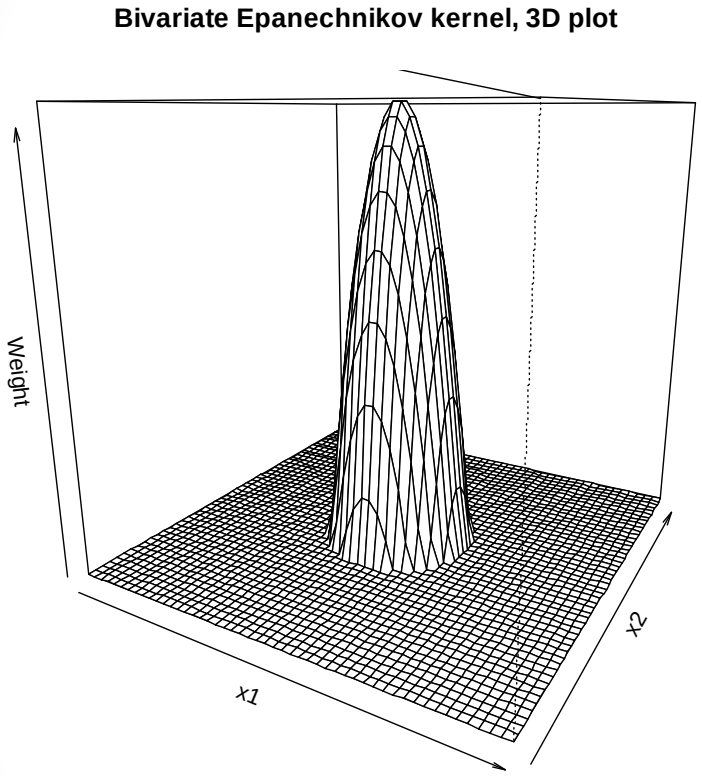
$$K_E(x_1, x_2) \neq K_E(x_1) \cdot K_E(x_2)$$

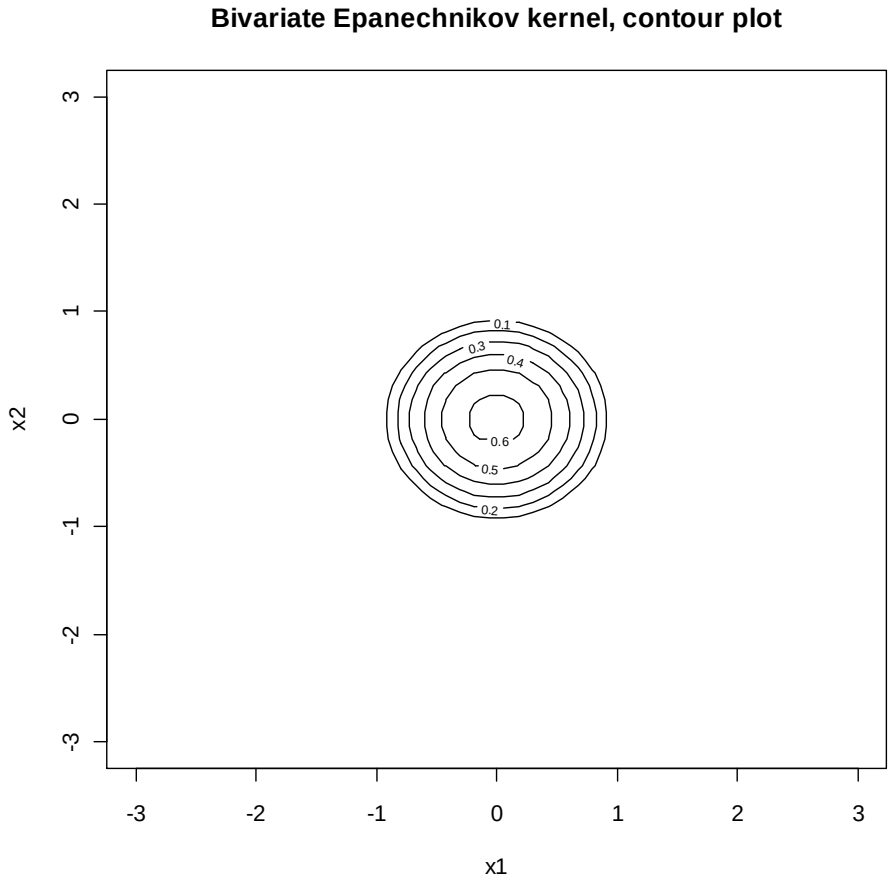
Двумерное гауссовское ядро



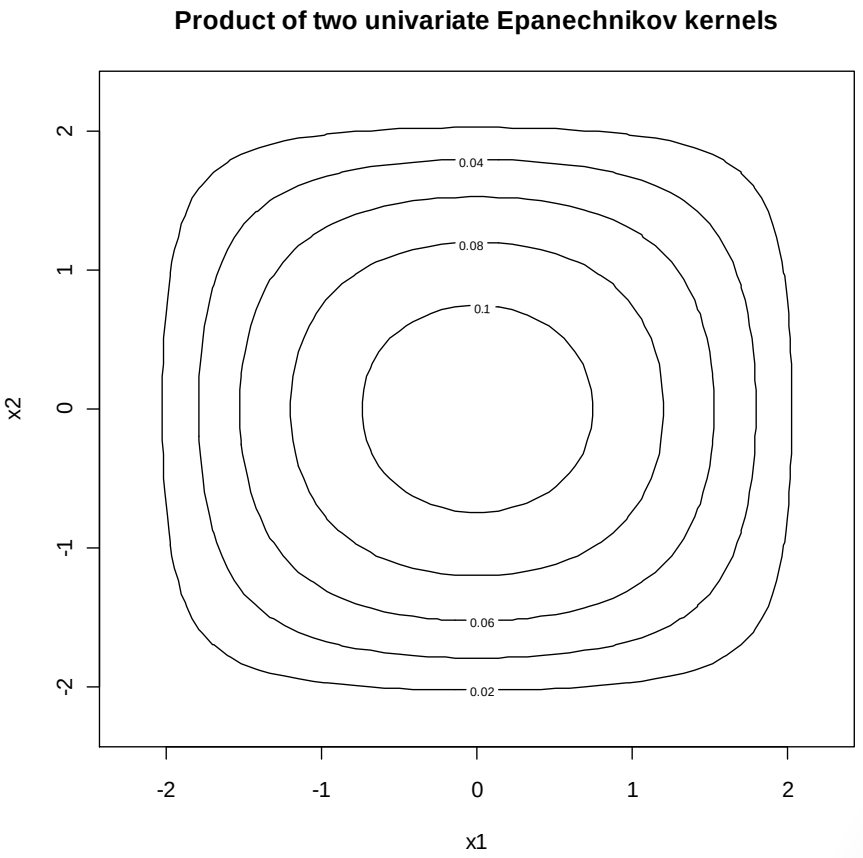


Двумерное ядро Епанечникова





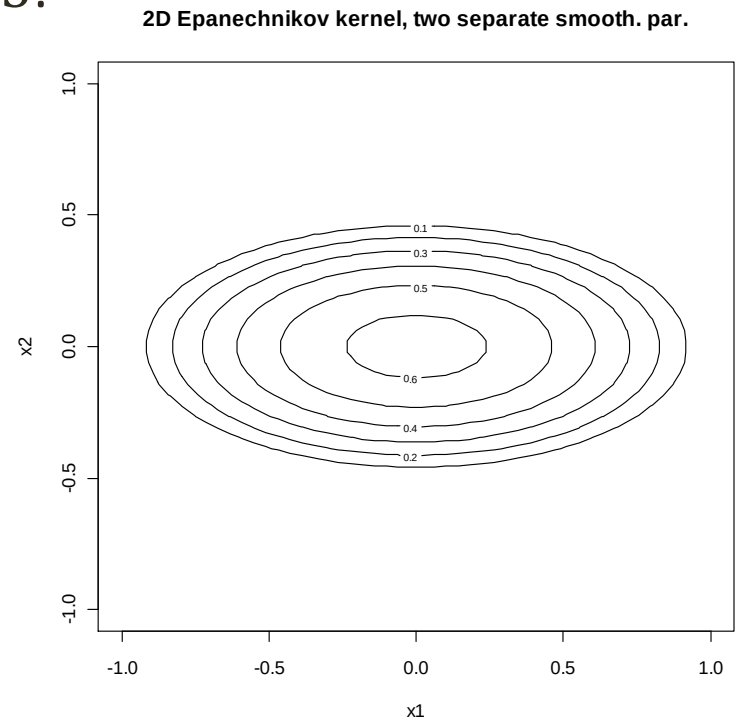
Двумерные ядерные функции



Различные сглаживающие параметры

Если разброс данных в первой и во второй выборке сильно отличаются, то можно использовать для этих выборок разные сглаживающие параметры $h_1 \neq h_2$

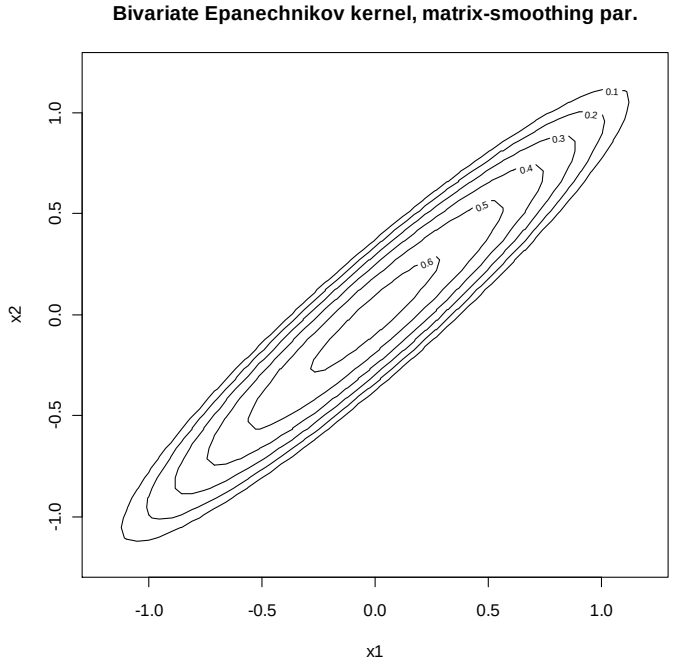
На рисунке ниже представлено двумерное ядро Епанечникова $K\left(\frac{x_1}{h_1}, \frac{x_2}{h_2}\right)$ со сглаживающими параметрами $h_1 = 1,\ h_2 = 0.5$:



Сглаживающая матрица

Если рассматриваемые величины коррелируют, это учитывается с помощью симметричной положительно определённой сглаживающей матрицы (matrix-smoothing parameter) *H*, которая в двумерном случае состоит их 4-х элементов: $H = \begin{pmatrix} h_1 & h_{12} \\ h_{21} & h_2 \end{pmatrix},\ h_{12} = h_{21}$

Для $h_1 = h_2 = 1,\ h_{12} = h_{21} = \sqrt{0.5}$, получим:



Общий случай

Ядерная оценка *d*-мерной плотности:

$$\hat{f}(\vec{y}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{|H|} K\left(H^{-1}(\vec{y} - \vec{y}_i)\right), \quad (31)$$

$\vec{y} = (y_1, \dots, y_d)$, $|H|$ — определитель матрицы H

Пусть $\vec{x} = (x_1, \dots, x_d)$, тогда *d*-мерные ядра Гаусса и Епанечникова запишутся как

$$K_G(\vec{x}) = (2\pi)^{-\frac{d}{2}} \exp\left(-\frac{\vec{x}\vec{x}'}{2}\right), \quad (32)$$

$$K_E(\vec{x}) = (2c_d)^{-1}(d+2)(1 - \vec{x}\vec{x}') \cdot I(\vec{x}\vec{x}' < 1), \quad (33)$$

c_d — объём единичного *d*-мерного шара

Правило подстановки

Если в качестве подставляемого распределения использовать нормальное $N(\mu, \Sigma)$, $\Sigma = \left(\sigma_1^2, \dots, \sigma_d^2\right) \cdot Ъ$, $Ъ_{[d \times d]}$ — единичная матрица, то по критерию *MISE*, оптимальная диагональная матрица *H* состоит из элементов

$$h_j = \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} n^{-\frac{1}{d+4}} \sigma_j \quad (34)$$

Так как первый множитель при любых *d* приблизительно равен единице на практике используют правило

$$\hat{h} = n^{-\frac{1}{d+4}} \hat{\sigma}_j \quad (35)$$

Обобщённое правило подстановки:

$$\hat{H} = n^{-\frac{1}{d+4}} \hat{\Sigma}^{\frac{1}{2}} \quad (36)$$

Правило подстановки

На практике обобщённое правило подстановки (36) используют следующим образом:

- данные преобразуются так, чтобы они имели единичную ковариационную матрицу $\Sigma = \text{ב}$;
- строится оценка плотности с единственным сглаживающим параметром, $\widehat{H} = \hat{h} \cdot \text{ב},\ \hat{h} = n^{-\frac{1}{d+4}}$;
- выполняется обратное преобразование для полученной оценки

Метод перекрёстной проверки

Метод перекрёстной проверки также может быть обобщён на многомерный случай, однако при этом он становится достаточно сложным, требующим ресурсоёмких вычислений

Алгоритм практического применения метода аналогичен правилу подстановки, но в этом случае, при предположении, что $K(\vec{x})$ — симметричная функция, оценка $\hat{h}$ находится путём минимизации выражения

$$CV(h) = \frac{1}{n^2 h^d} \sum_{i=1}^{n} \sum_{j=1}^{n} K^* \left(\frac{\vec{y}_i - \vec{y}_j}{h} \right) + \frac{2}{nh^d} K\left(0_{[1\times d]}\right), \quad (37)$$

$$K^*(\vec{x}) = \int_{R^d} K\left(\vec{t}\right) K\left(\vec{x} - \vec{t}\right) d\vec{t} - 2K(\vec{x}),$$

$$\vec{t} = (t_1, \dots, t_d)$$

Обобщённый метод ближайших соседей

Пусть $d_k(\vec{y})$ — евклидово расстояние от точки $\vec{y}$ до *k*-го ближайшего наблюдения в выборке, $V_k(\vec{y})$ — объём *d*-мерного шара радиусом $d_k(\vec{y})$, $V_k(\vec{y}) = c_d d_k^d(\vec{y})$

В этом случае простая оценка равна

$$\hat{f}(\vec{y}) = \frac{k}{nV_k(\vec{y})} = \frac{k}{nc_d d_k^d(\vec{y})}, \quad (38)$$

что аналогично одномерной оценке (22), так как $c_1 = 2$

Оценка (38) может быть обобщена с помощью ядер:

$$\hat{f}(\vec{y}) = \frac{1}{c_d n d_k^d(\vec{y})} \sum_{i=1}^{n} K\left(\frac{\vec{y}-\vec{y}_i}{c_d d_k(\vec{y})}\right), \quad (39)$$

что аналогично одномерной оценке (23)

Адаптивный метод ближайших соседей

Рассмотрим $d_k(\vec{y}_i)$ — расстояние от элемента $\vec{y}_i$ до *k*-го ближайшего элемента выборки

Оценка плотности по адаптивному методу равна

$$\hat{f}(\vec{y}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^d d_k^d(\vec{y}_i)} K\left(\frac{\vec{y}-\vec{y}_i}{h d_k(\vec{y}_i)}\right), \quad (40)$$

что аналогично одномерной оценке (24)

Как и в одномерном случае показатель локальной концентрации наблюдений $d_k(\vec{y}_i)$ часто заменяется на величину

$$\lambda_i = \left(\frac{g}{\tilde{f}(\vec{y}_i)}\right)^{\alpha}, \ \alpha \in [0;1], \quad (41)$$

$g = \left(\prod_{i=1}^{n} \tilde{f}(y_i)\right)^{\frac{1}{n}}$ — геометрическое среднее пилотных оценок плотности

Практическая часть

Построение непараметрических оценок плотности

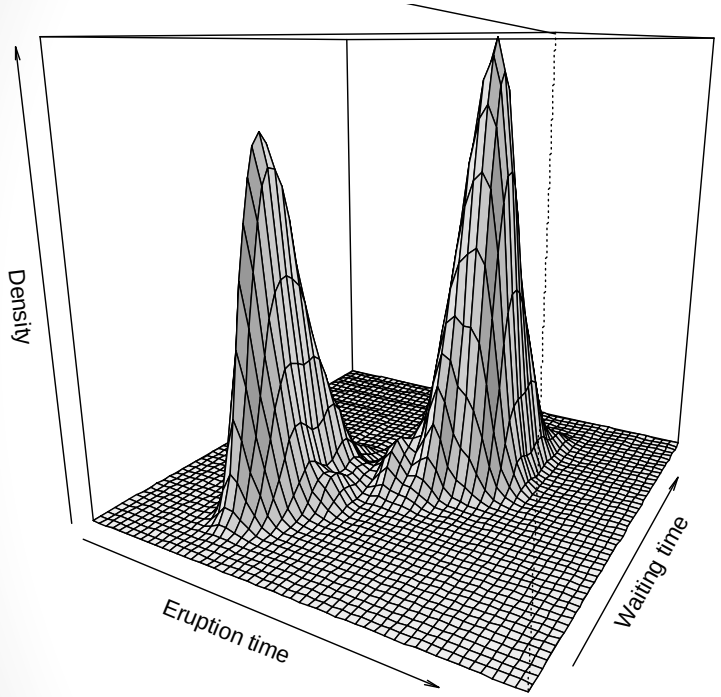
в программной среде «R»

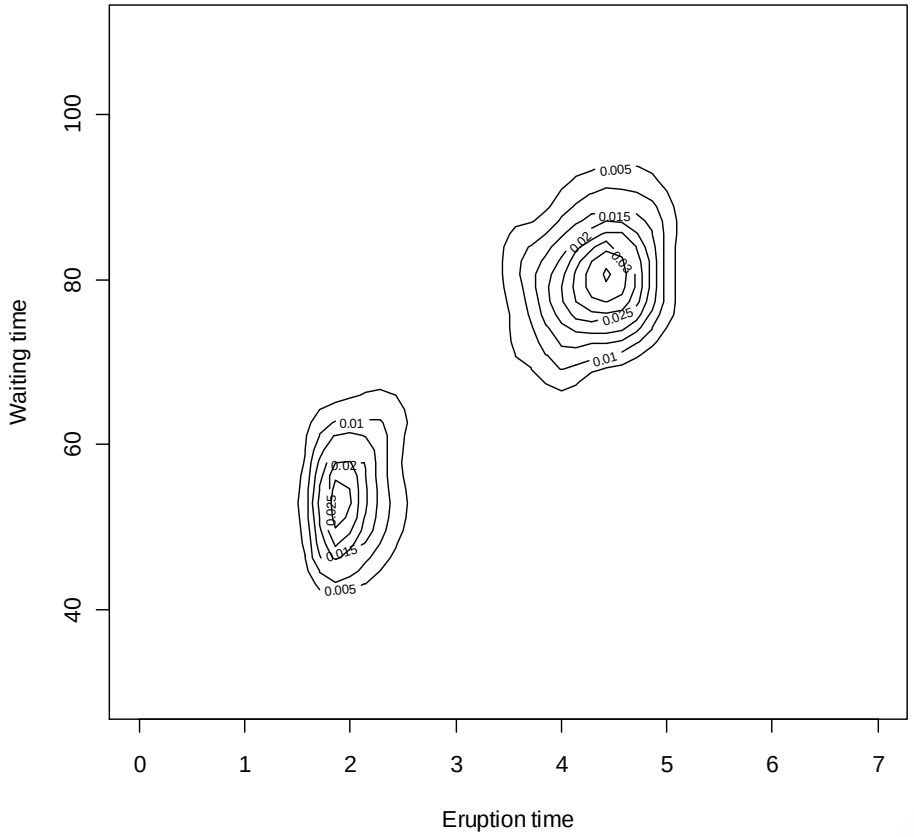
cran.r-project.org

Пример 2. Старый служака

```
y <- faithful; N <- nrow(y)

# сетка для расчёта оценок плотности
L <- 50; u <- seq(0,7,length=L); v <- seq(30,110,length=L)
uv <- expand.grid(u,v)

# оценка плотности
f.fix <- npudens(tdat=y,edat=uv,ckertype="gaussian",bwtype="fixed")

# графики оценки
w <- f.fix$dens; dim(w) <- c(L,L)

persp(u,v,w,theta=30,main="Bivariate kernel estimate, 3D plot",
xlab="Eruption time",ylab="Waiting time",zlab="Density")

contour(u,v,w,nlevel=7,
main="Bivariate kernel estimate, contour plot",
xlab="Eruption time",ylab="Waiting time")
```

Пример 2. Старый служака



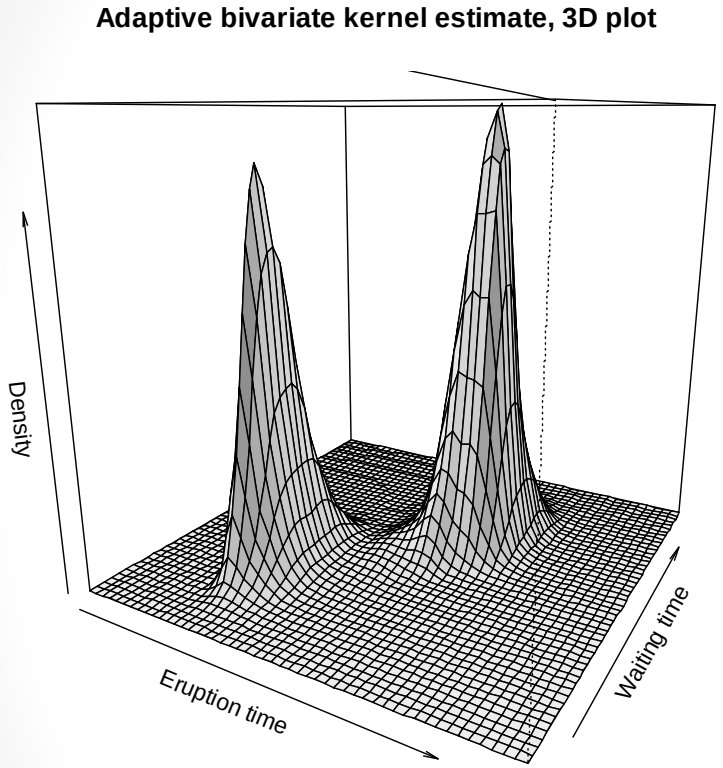


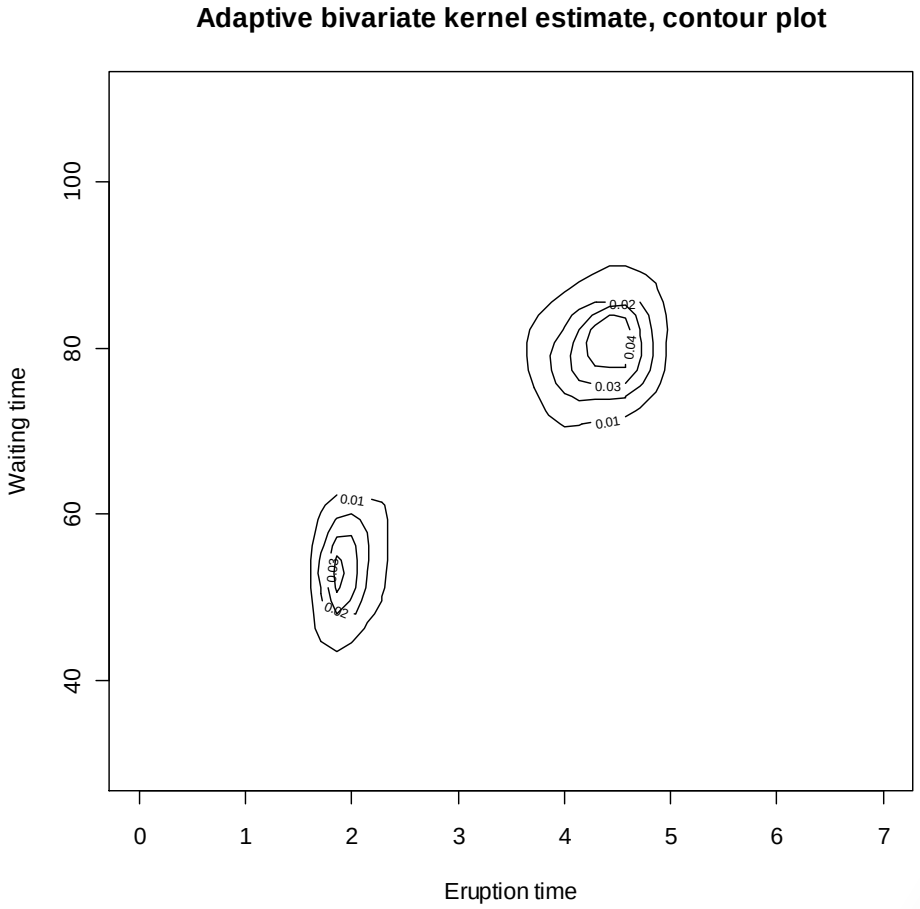
Пример 2. Старый служака

Адаптивный метод с λ_i, аналогично одномерному случаю

```
pilot <- npudens(tdat=y,ckertype="gaussian",bwtype="fixed")
h <- pilot$bws$bw

g <- 1
for (i in 1:N) g <- g*pilot$dens[i]^(1/N)

alpha <- 0.5
lmbd <- (g/pilot$dens)^alpha

kern <- function(x) exp(-(x[1]^2+x[2]^2)/2)/(2*pi)

f <- rep(0,times=L^2)
for (i in 1:(L^2)) {
  for (j in 1:N) f[i] <- f[i]+kern((uv[i,]-
    y[j,])/(h*lmbd[j]))/lmbd[j]^2
  f[i] <- f[i]/(N*h[1]*h[2])
}
```

Пример 2. Старый служака





Пример 2. Старый служака

Значения логарифмической функции правдоподобия

оценки плотности в точках y_i

```
f.fix.llh <- npudens(tdat=y,ckertype="gaussian",bwtype="fixed")
llh.fix <- sum(log(f.fix.llh$dens))
```

для адаптивного метода

```
f.llh <- rep(0,times=N)
for (i in 1:N) {
  for (j in 1:N) f.llh[i] <- f.llh[i]+kern((y[i,]-
    y[j,])/(h*lmbd[j]))/lmbd[j]^2
  f.llh[i] <- f.llh[i]/(N*h[1]*h[2])
}
llh.ada <- sum(log(f.llh))
```

llh.fix	**-1106**
llh.ada	**-1114**

Пример 2. Старый служака

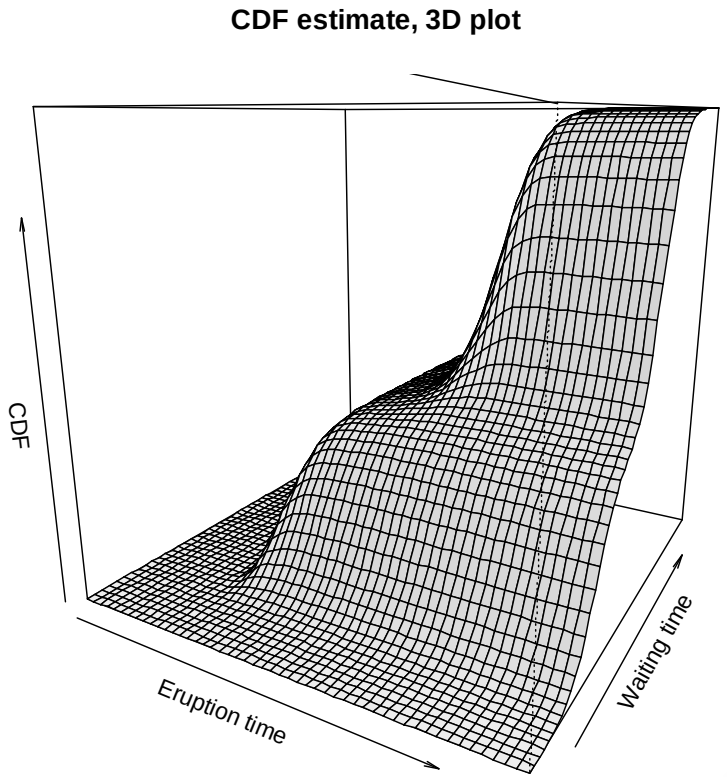
Расчёт функций распределения

фиксированный метод

```
F.fix <- npudist(tdat=y,edat=uv,ckertype="gaussian",bwtype="fixed")
```

адаптивный метод

```
du <- u[2]-u[1]; dv <- v[2]-v[1]
w <- f; dim(w) <- c(L,L)
F <- rep(0,times=L^2)
for (i in 1:L) {
  for (j in 1:L) F[j+(i-1)*L] <-
  sum(w[1:j,1:i])*du*dv
}
```

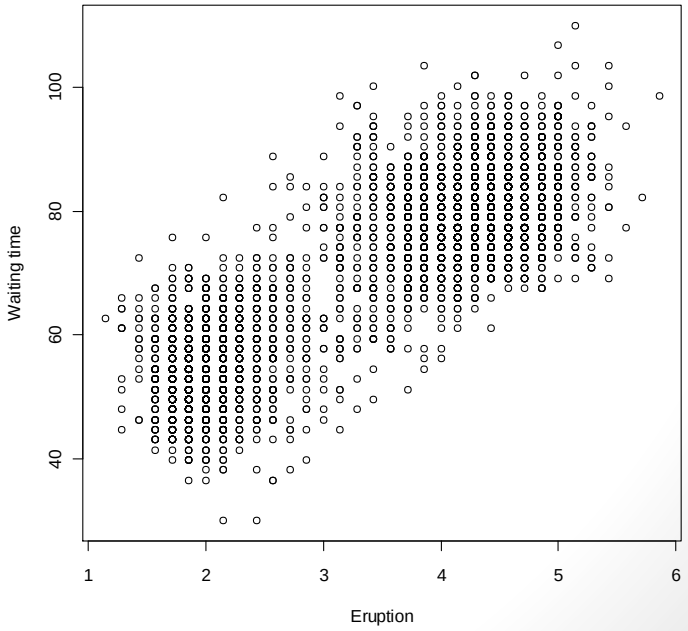


Пример 2. Старый служака

Генератор случайных чисел

для адаптивного метода

```
alpha <- 0.99
M <- 5000
smpl.ind <- sample(1:(L^2),prob=f,size=M,replace=TRUE)
y.ada.sim <- uv[smpl.ind,]
plot(y.ada.sim,xlab="Eruption",ylab="Waiting time")
```



Пример 2. Старый служака

Рисование графиков с перекрывающими друг друга точками

```
plot(y.ada.sim,col=rgb(0,0,1,alpha=0.2))
smoothScatter(y.ada.sim)
```