Chap8 非参数密度估计技术

参考:王星2009《 非参数统计》
清华大学出版社

主讲：王 星
助教：范 超
中国人民大学统计学院
办公地点：明德主楼1019
办公电话：82500167
课程网站：https://dm.ruc.edu.cn

2014年12月24日

基本概念

- 想一想：什么是分布密度？分布密度有什么用？

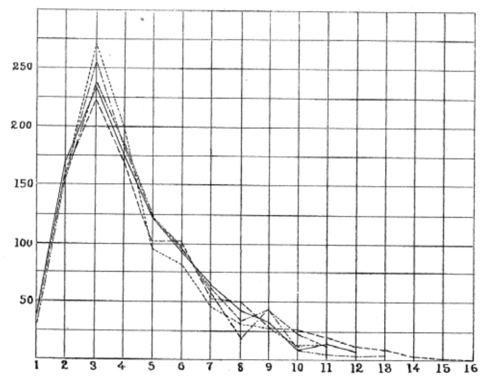


FIG. 2.—SHOWING FIVE GROUPS, OF ONE THOUSAND WORDS EACH, FROM 'OLIVER TWIST.'

Zipf齐普夫定律:在自然语言的语料库里，一个单词出现的频率与它在频率表里的排名成反比

色泽不均衡可能是催熟西瓜

分布密度和一个随机变量取值分布的均衡性有关系，不均衡常常是世界的常态，语言学中重要的词一定被使用的频次高、食品安全监测中的分布异常可能是风险的一个标志？

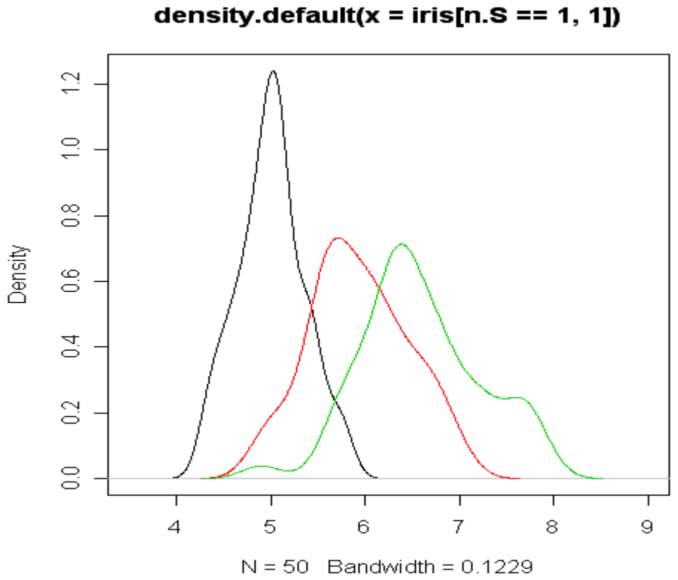
通过数据估计分布密度通常都有什么方法？

非参数密度估计

- 直方图
- Parzen Windows窗
- Kernel density estimator
- 多元密度估计
- 判别分析

Introduction

- 大部分的参数密度都是单峰的 (have a single local maximum), 很多实际问题会涉及多峰问题
- 非参数统计过程将涉及假定宽松的数据结构.
- 有两种常见的非参数密度估计问题:
  - 估计似然函数 P(x|ωj)
  - 直接估计后验概率



密度估计

– Basic idea:
  Probability that a vector x will fall in region R is:

$$P = \int_{\Re} p(x')dx' \qquad (1)$$

Therefore, the ratio k/n is a good estimate for the probability P and hence for the density function p.

$$\int_{\Re} p(x')dx' \cong p(x)V \qquad (4)$$

p(x) is continuous and that the region $\mathcal{R}$ is so small that p does not vary significantly within it, we can write:

$$\hat{p}_n(x) \cong \frac{k/n}{V}$$

where x is a point within $\mathcal{R}$ and V the volume enclosed by $\mathcal{R}$.

equation (1) and (4) yields histogram:

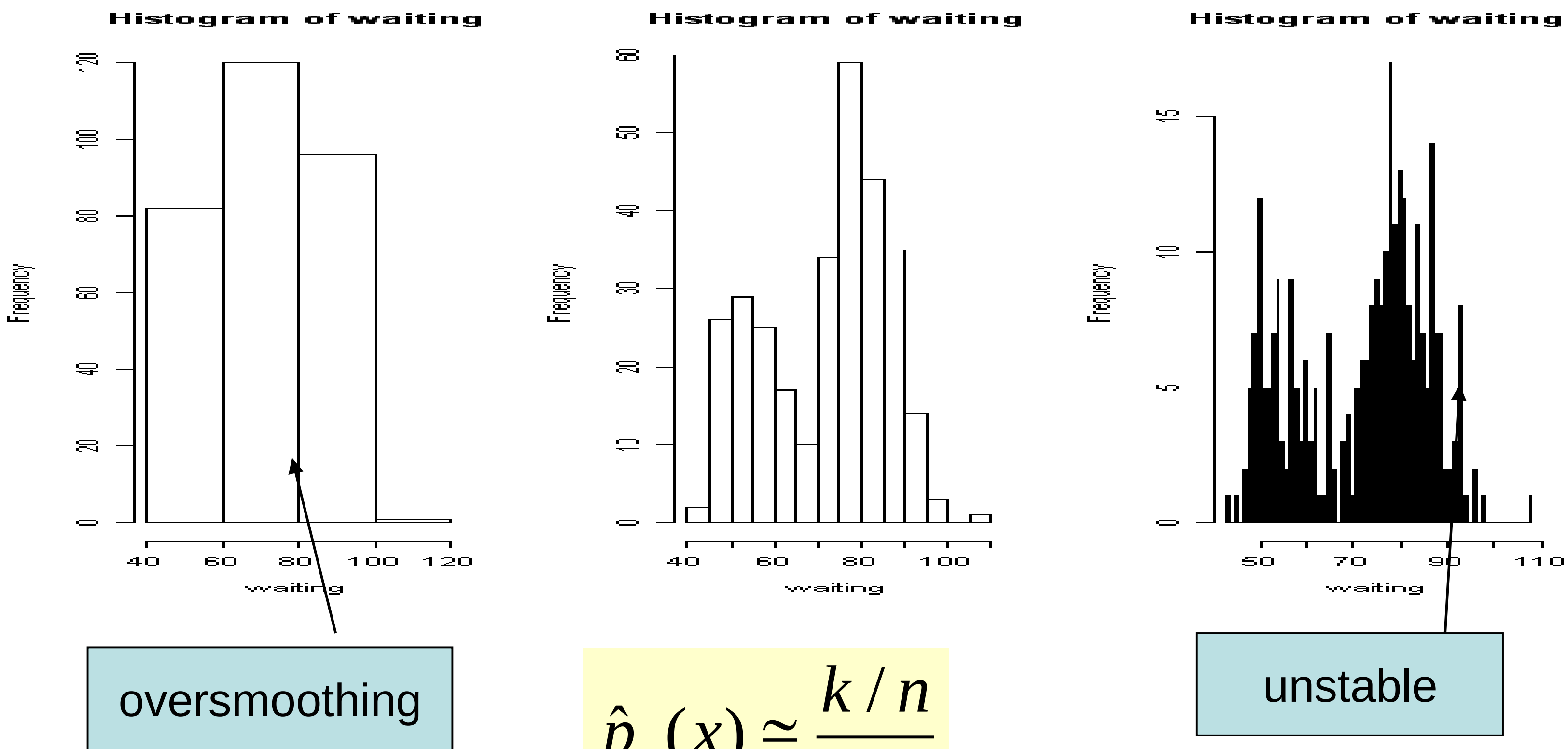
直方图

$$\hat{p}(x) = \begin{cases} \frac{n_i}{nh}, & \text{当} x \in I_i, i = 1, 2, \cdots, k; \\ 0, & \text{其他}. \end{cases}$$

h既是归一化参数，又表示每一组的组距，称为带宽或窗宽.

- Dissects the range of the data into bins of equal width along the horizontal axis
- Vertical axis represents the frequency counts (or percents, proportions)—Bars represent the counts
- Fewer bins, smoother histogram, but less detail about the distribution
- Trade-off between smoothness and detail: We want to preserve as much detail as possible but we do not want the graph to be too rough (difficult to discern shape)

最佳窗宽选择



$$\hat{p}_n(x) \cong \frac{k/n}{V}$$

如果这个体积和所有的样本体积相比很小，就会得到一个很不稳定的估计，这时，密度值局部变化很大，呈现多峰不稳定的特点；反之，如果这个体积太大，则会圈进大量样本，从而使估计过于平滑，不稳定与过度光滑之间寻找平衡就引导出下面两种可能的解决方法：

最优理论窗宽 Histogram

定理: $\int (f'(u))^2 du < +\infty$ **则L^2损失下的最优风险为:**

$$R(\hat{f}_n(x), f) \approx \frac{h^3}{12}\int (f'(u))^2 du + \frac{1}{nh}$$

极小化上面的式子,可以得到理想的窗宽:

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

在这个窗宽的选择下

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

定理8.1 固定x和h,令估计的密度是$p(x)$, 如果$x \in I_j, p_j = \int_{I_j} p(x)\, \mathrm{d}x$, 有

$$E\hat{p}(x) = p_j/h, \quad \mathrm{var}\hat{p}(x) = \frac{p_j(1-p_j)}{nh^2}.$$

证明提示：注意到$E\hat{p}_j = n_j/n = \int_{I_j} p(x)\, \mathrm{d}x, \mathrm{var}\hat{p}_j = p_j(1-p_j)/n.$

考察平方损失风险：

$$\begin{aligned}
R(\hat{p}, p) &= EL(\hat{p}(x), p(x)) \\
&= \int (\hat{p}(x) - p(x))^2 \, \mathrm{d}x \\
&= \int (E\hat{p}(x) - p(x))^2 \, \mathrm{d}x + \int (\hat{p}(x) - E\hat{p}(x))^2 \, \mathrm{d}x \\
&= \int \mathrm{Bias}^2(x) \, \mathrm{d}x + \int V(x) \, \mathrm{d}x.
\end{aligned}$$

积分均方误 (Mean Integral Square Error, 简称： MISE)

$$\text{MISE} = \mathrm{E}\left[\int (\hat{p}_n(x) - p(x))^2 \,\mathrm{d}x\right].$$

$$\text{AMISE} = \int [(\text{Bias}(\hat{f}))^2 + \text{Var}(\hat{f})] \,\mathrm{d}x.$$

$$\begin{aligned}\mathrm{Bias}(x) &= E\hat{p}(x) - p(x) = \frac{p_j}{h} - p(x) \\ &= \frac{p(x)h + hp'(x)[h(j-1/2)-x]}{h} - p(x) \\ &= p'(x)[h(j-1/2)-x].\end{aligned}$$

$$\begin{aligned}\int_{I_j} \mathrm{Bias}^2(x)\mathrm{d}x &= \int_{I_j} (p'(x))^2 \, [h(j-1/2)-x]^2 \; \mathrm{d}x \\ &\approx (p'(\xi_j))^2 \frac{h^3}{12},\end{aligned}$$

$$
\begin{aligned}
V(x) &= \frac{p_j}{nh^2} \\
&= \frac{p(x)h + hp'(x)[h(j-1/2)-x]}{nh^2} \\
&\approx p(x)/nh.
\end{aligned}
$$

$$R(\hat{p}, p) \approx \frac{h^3}{12} \int (p'(u))^2 \, \mathrm{d}u + \frac{1}{nh}.$$

极小化上式, 得到理想带宽为

$$h^* = \frac{1}{n^{1/3}} \left(\frac{6}{\int p'(x)^2 \, \mathrm{d}x} \right)^{1/3}.$$

$$h = Cn^{-1/3}.$$

选择箱量（等价于窗宽）

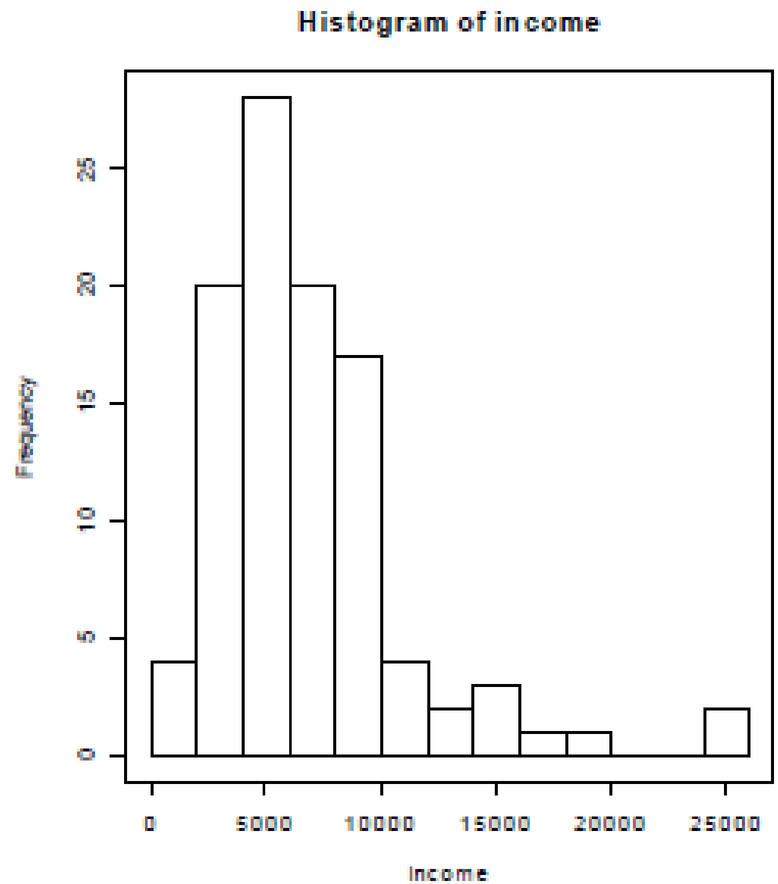
- Simple rule of thumb for small datasets (approx. 100 or less) is:

$$\# \text{ of bins} = 2\sqrt{n}$$

- For larger samples, the `car` package for **R** implements Freedman and Diaconis (1981) recommended formula from the `n.bins` function:

$$\# \text{ of bins} = \left[\frac{n^{1/3}(\max - \min)}{2(Q_3 - Q_1)} \right]$$

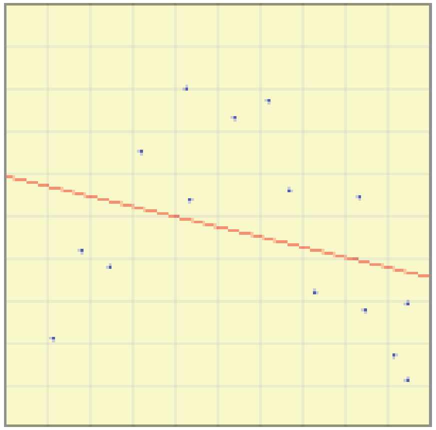
```
>hist(income, nclass=n.bins(income))
```

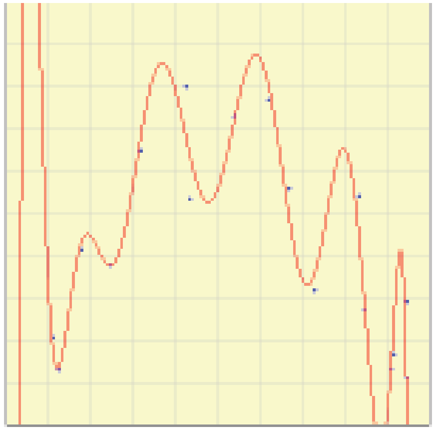


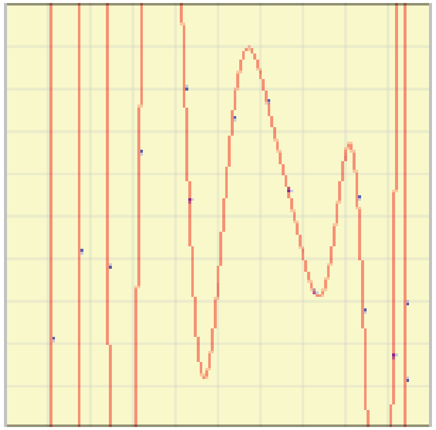
偏差与方差分解

Choice of hypothesis class introduces learning bias

- More complex class → less bias
- More complex class → more variance







模型偏差太大

模型方差太大

bias-variance偏差和方差分解

For any estimator $\tilde{\theta}$:

$$\begin{aligned}
\mathrm{MSE}(\tilde{\theta}) &= E(\tilde{\theta}-\theta)^2 \\
&= E(\tilde{\theta}-E(\tilde{\theta})+E(\tilde{\theta})-\theta)^2 \\
&= E(\tilde{\theta}-E(\tilde{\theta}))^2+E(E(\tilde{\theta})-\theta)^2 \\
&= Var(\tilde{\theta})+(E(\tilde{\theta})-\theta)^2
\end{aligned}$$

bias

Note MSE closely related to prediction error:

$$E(Y_0-x_0^T\tilde{\beta})^2 = E(Y_0-x_0^T\beta)^2+E(x_0^T\tilde{\beta}-x_0^T\beta)^2 = \sigma^2+MSE(x_0^T\tilde{\beta})$$

The practical approximate bandwidth from Cross Validation

$$\hat{J}(h) = \int (\hat{f}_n)^2 dx - \frac{2}{n} \sum_{i=1} \hat{f}_{(-i)}(x_i)$$

一般当h未知的时候，可以用更实用的方式选择窗宽，

$$\begin{aligned} R(h) &= \int (\hat{p} - p(x))^2 \mathrm{d}x \\ &= \int \hat{p}^2 \mathrm{d}x - 2\int \hat{p}p \mathrm{d}x + \int p^2(x) \mathrm{d}x \\ &= J(h) + \int p^2(x) \mathrm{d}x. \end{aligned}$$

注意到后面一项与h无关，第一项可以用交叉验证方法估计：

$$\hat{J}(h) = \int (\hat{p})^2 \mathrm{d}x - \frac{2}{n} \sum_{i=1} \hat{p}_{(-i)}(x_i).$$

其中，$\hat{p}_{(-i)}(x_i)$是去掉第i个观测值后对直方图的估计，$\hat{J}(h)$称为交叉验证得分.

Parzen Windows（固定V）

– Parzen-window approach to estimate densities assume that the region $\mathcal{R}_n$ is a d-dimensional hypercube

$$V_n = h_n^d \ (h_n : \text{length of the edge of } \Re_n)$$

Let $\varphi(u)$ be the following window function :

$$\varphi(u) = \begin{cases} 1 & |u_j| \le \frac{1}{2} \quad j = 1, \ldots, d \\ 0 & \text{otherwise} \end{cases}$$

$\varphi((x-x_i)/h_n)$ is equal to unity if x_i falls within the hypercube of volume V_n centered at x and equal to zero otherwise.

– The number of samples in this hypercube is:

$$k_n = \sum_{i=1}^{i=n} \varphi\left(\frac{x - x_i}{h_n}\right)$$

By substituting k_n in equation (7), we obtain the following estimate:

$$\mathbf{p_n(x) = \frac{1}{n}\sum_{i=1}^{i=n} \frac{1}{V_n}\, \varphi\left(\frac{x - x_i}{h_n}\right)}$$

$P_n(x)$ estimates $p(x)$ as an average of functions of x and the samples (x_i) $(i = 1, \ldots, n)$. These functions φ can be general!

– 举例:

The behavior of the Parzen-window method

Case where $p(x) \rightarrow N(0,1)$

Let $\varphi(u) = (1/\sqrt{2\pi}) \exp(-u^2/2)$ and $h_n = h_1/\sqrt{n}$ $(n>1)$

(h_1: known parameter)

Thus:

$$p_n(x) = \frac{1}{n} \sum_{i=1}^{i=n} \frac{1}{h_n} \varphi\left(\frac{x - x_i}{h_n} \right)$$

is an average of normal densities centered at the samples x_i.

- Essentially a sophisticated form of ***locally weighted averaging*** of the distribution
- Use a weight function (kernel) that ensures the enclosed area of the curve equals 1
  - Probability density functions (such as the ***standard normal density function***) are good choices because they are smooth and symmetric
- The ***kernel density estimate*** is calculated as follows:

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

where K is the kernel density function

x is the point where the density is estimated

X_i is the centre of the interval

h is the bandwidth (or window half-width)

$$K(x) \geqslant 0, \quad \int K(x)\,\mathrm{d}x = 1.$$

$$\begin{aligned}
&\int \hat{p}(x)\mathrm{d}x \\
&= \int \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-x_i}{h}\right)\,\mathrm{d}x = \frac{1}{n}\sum_{i=1}^{n}\int \frac{1}{h}K\left(\frac{x-x_i}{h}\right)\,\mathrm{d}x \\
&= \frac{1}{n}\sum_{i=1}^{n}\int K(u)\,\mathrm{d}u = \frac{1}{n}\cdot n = 1 \ \left(\text{其中 } u = \frac{x-x_i}{h}\right).
\end{aligned}$$

R中常用的核函数

表 8.1. 常用核函数

核函数名称	核函数 $K(u)$	S-Plus 中
Parzen 窗 (Uniform)	$\frac{1}{2}I(\|u\| \le 1)$	√
三角 (Triangle)	$(1-\|u\|)I(\|u\| \le 1)$	√
Epanechikov	$\frac{3}{4}(1-u^2)I(\|u\| \le 1)$	
四次 (Quartic)	$\frac{15}{16}(1-u^2)I(\|u\| \le 1)$	
三权 (Triweight)	$\frac{35}{32}(1-u^2)^3I(\|u\| \le 1)$	
高斯 (Gauss)	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)$	√
余弦 (Cosinus)	$\frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right)I(\|u\| \le 1)$	√
指数 (Exponent)	$\exp\{\|u\|\}$	

核估计的性质

与直方图类似, 也可以得到大样本情况下核估计的如下一些基本结论.

我们先来估计$Bias(\hat{p})$, 首先，令$(x-x_i)/h=t$ and $x_i=x-ht$，计算可得

$$
\begin{aligned}
\int h^{-1}K\left(\frac{x-x_i}{h}\right)p(x_i)dx_i &= \int h^{-1}K(u)p(x-ht)d(x-ht)\\
&= \int h^{-1}K(u)p(x-ht)|-h|dt\\
&= \int K(u)p(x-ht)dt
\end{aligned}
$$

使用泰勒展开$p(x-ht)-p(x)=-htp'(x)+\frac{1}{2}h^2t^2p''(x)+O(h^3)$ 因此，我们得到

$$
\begin{aligned}
&\int h^{-1}K\left(\frac{x-x_i}{h}\right)p(x_i)dx_i - p(x)\\
&= \int K(u)\{p(x-ht)-p(x)\}dt\\
&= -hp'(x)\int tK(t)dt + \frac{1}{2}h^2p^{(2)}(x)\int t^2K(t)dt + O(h^3)\\
&= \frac{h^2}{2}\mu_2(K)p^{(2)}(x) + O(h^3)
\end{aligned}
$$

$$\mu_2(K)=\int t^2K(t)dt.$$

核估计的性质

定理 8.2: 假设 $\hat{p}_n(x)$ 定义如式 (8.3) 是 $p(x)$ 的核估计，令 $\mathrm{supp}(p) = \{x : p(x) > 0\}$ 是密度 p 的支撑。设 $x \in \mathrm{supp}(p) \subset R$ 为 $\mathrm{supp}(p)$ 的内点 (非边界点)，当 $n \to +\infty$ 时， $h_n \to 0$， $nh_n \to +\infty$，核估计有如下性质：

$$\mathrm{Bias}(\hat{p}_n(x)) = \frac{h_n^2}{2}\mu_2(K)p^{(2)}(x) + o(h^3);$$

带宽h越小，核估计的偏差越小

$$\mathrm{Var}(\hat{p}_n(x)) = (nh_n)^{-1}p(x)R(K) + 0((nh_n)^{-1}) + O(n^{-1});$$

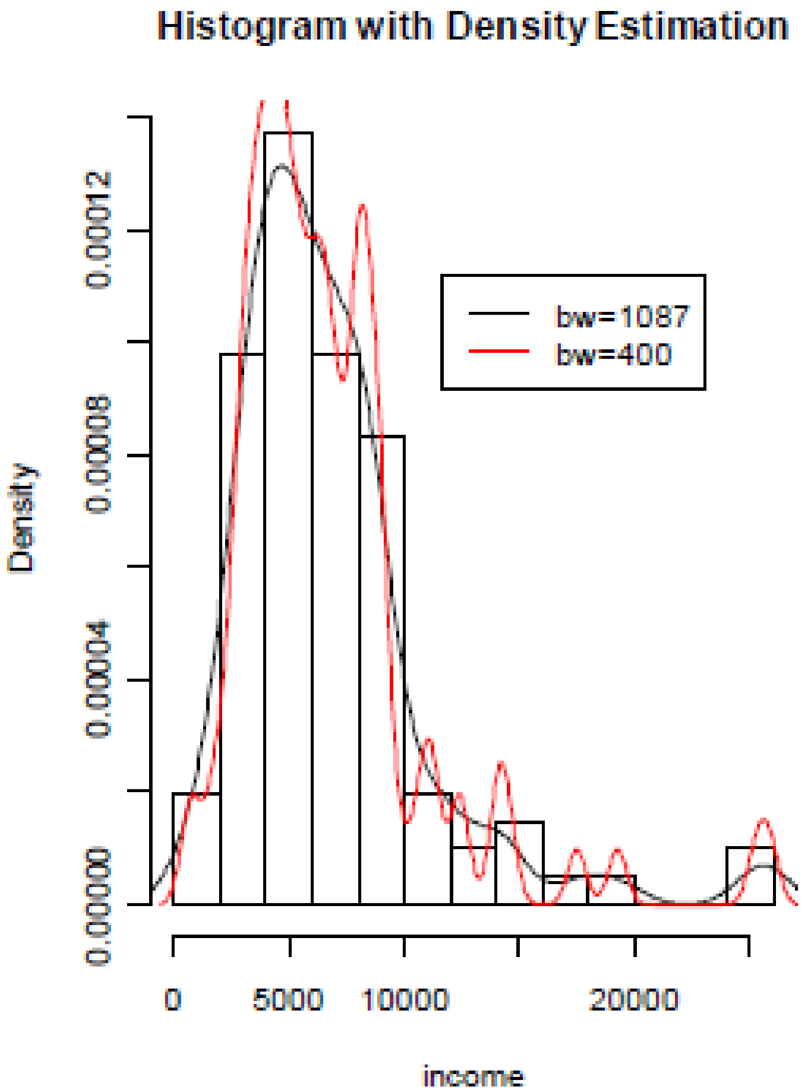
若 $\sqrt{(nh_n)}\, h_n^2 \longrightarrow 0$, 则

$$\sqrt{(nh_n)}(\hat{p}_n(x) - p(x)) \longrightarrow N(0, p(x)R(K))$$

其中 $R(g(x)) = \int g(x)^2 \mathrm{d}x$.

$$h_{opt} = \mu_2(K)^{-2/5}\left\{\int K(x)^2 dx\right\}^{1/5}\left\{\int p^{(2)}(x)^2 dx\right\}^{-1/5} n^{-1/5}$$

- If the underlying density distribution is substantially nonnormal, $h = 0.9An^{-1/5}$ produces a window width $2h$ that is too wide (*i.e.*, the line is too rough), but it is good as a starting point
- As the bandwidth increases, the density curve becomes smoother
  - Ideally we want a smooth curve like the black line to the right (bw=1087)




```
hist(income, nclass=n.bins(income), probability=T,
     main='Histogram with Density Estimation', ylab='Density')
lines(density(income), col='red', lwd=2)
lines(density(income, bw = 400,
      kernel = c("rectangular")), lty=1, lwd=1)
legend(locator(1), lty=1:1, lwd=2:1, col=2:1,
      legend=c('bw=1087', 'bw=400'))
```

- Unlike histograms we no longer set the number of bins; instead we must select the bandwidth h. We can do this visually, but statistical theory provides some help:

$$h = 0.9\sigma n^{-1/5}$$

- The population standard deviation σ is unknown so we replace it with an adaptive estimator of spread (The sample standard deviation S can be inflated if the underlying density isn't normal):

$$A = \min\left(S, \frac{\text{hinge spread}}{1.349}\right)$$

  - Hinge spread is the inter-quartile range; 1.349 is the hinge spread of the standard normal distribution.
- The formula for the bandwidth is then:

$$h = 0.9An^{-1/5}$$

核估计的性质

与直方图类似, 也可以得到大样本情况下核估计的如下一些基本结论.

我们先来估计$Bias(\hat{p})$, 首先，令$(x-x_i)/h=t$ and $x_i=x-ht$，计算可得

$$\begin{aligned}\int h^{-1}K\left(\frac{x-x_i}{h}\right)p(x_i)dx_i &= \int h^{-1}K(u)p(x-ht)d(x-ht)\\ &= \int h^{-1}K(u)p(x-ht)|-h|dt\\ &= \int K(u)p(x-ht)dt\end{aligned}$$

使用泰勒展开$p(x-ht)-p(x)=-htp'(x)+\frac{1}{2}h^2t^2p''(x)+O(h^3)$ 因此，我们得到

$$\begin{aligned}&\int h^{-1}K\left(\frac{x-x_i}{h}\right)p(x_i)dx_i-p(x)\\ &= \int K(u)\{p(x-ht)-p(x)\}dt\\ &= -hp'(x)\int tK(t)dt+\frac{1}{2}h^2p^{(2)}(x)\int t^2K(t)dt+O(h^3)\\ &= \frac{h^2}{2}\mu_2(K)p^{(2)}(x)+O(h^3)\end{aligned}$$

$$\mu_2(K)=\int t^2K(t)dt.$$

应用：分位回归的参数分布估计

- 给出一个分位回归模型fit=rq(y~x)后，命令summary(fit,se='…')可以查看参数估计的结果
- se选项用于选择参数估计的不同方法，se='ker':核函数数估计法

定理 2. $\sqrt{n}(\hat{\beta}(\tau)-\beta(\tau)) \sim N(0,\tau(1-\tau)H_n(\tau)^{-1}Q_nH_n(\tau))$，其中：

$$H_n(\tau)=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}x_ix_i'f_i(\varepsilon_i(\tau))$$

$f_i(\varepsilon_i(\tau))$ 表示第i个残差 ε_i 在分位点 τ 处的分布密度；

```
library(quantreg)
fit1=rq(foodexp~income,data=engel)
summary(fit1,se="ker")
summary(fit1,se="boot")
summary(fit1,se="nid")
```

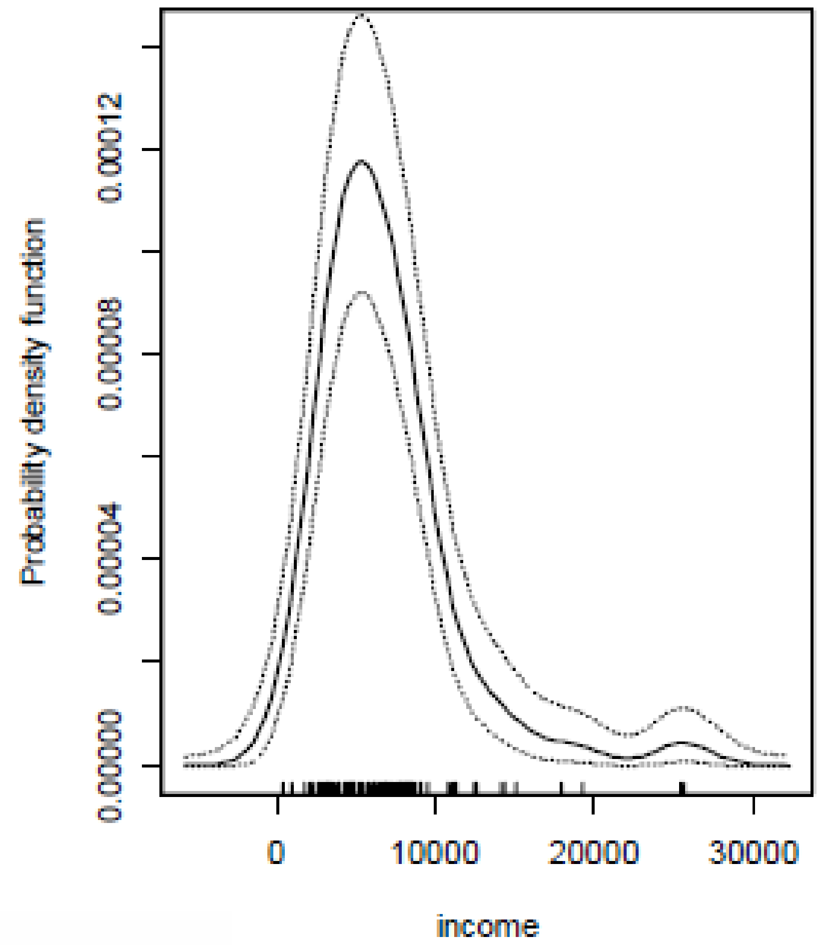
$$Q_0=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}x_ix_i'$$

- 因为残差分布未知，无法直接求出 $f_i(\varepsilon_i(\tau))$ $H_n(\tau)$
- Powell给出如下估计方法：

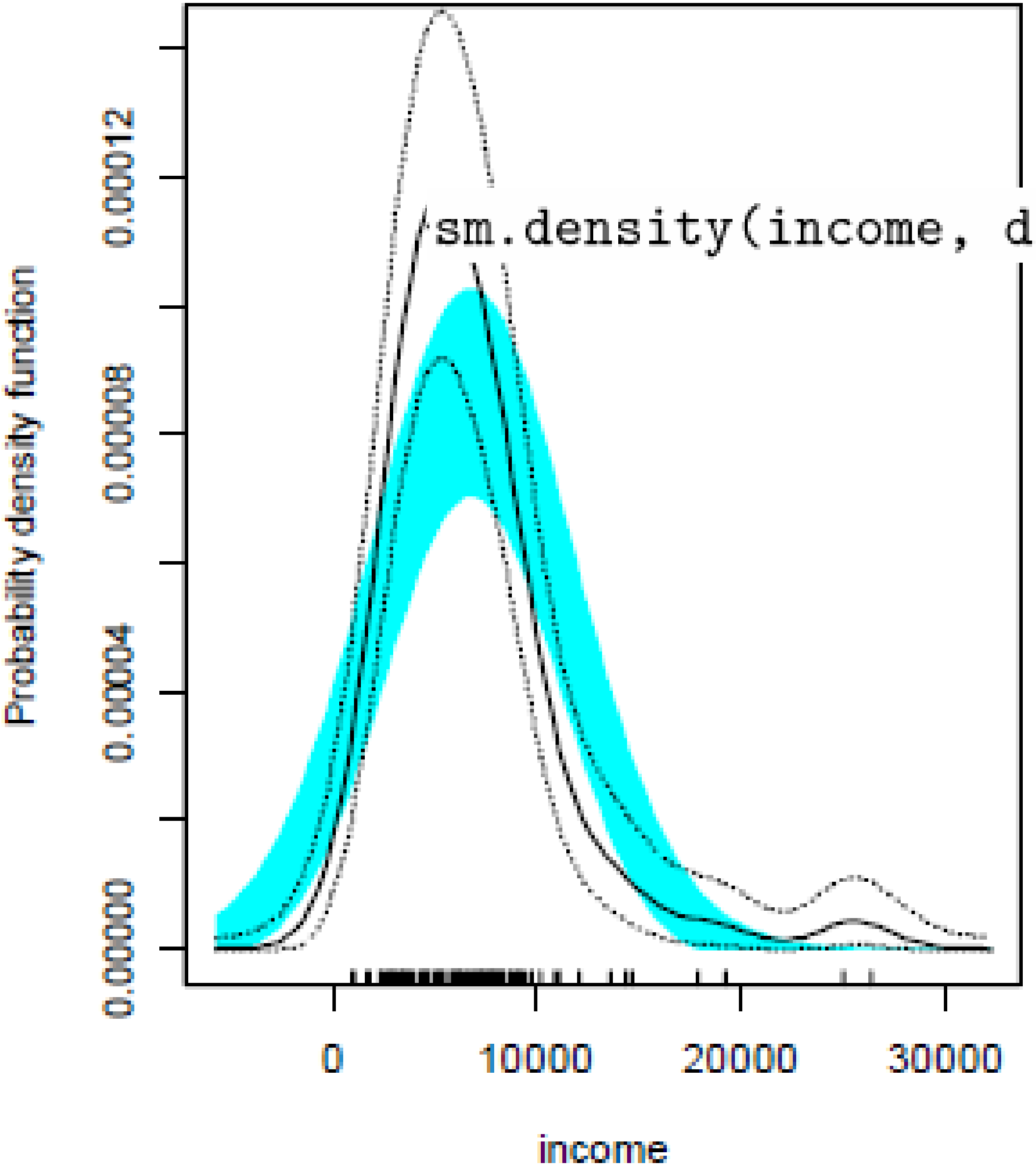
$$\hat{H}=\frac{1}{2c_nn}\sum_{i=1}^{n}I(|u_i|<c_n)x_ix_i'$$

sm包 confidence envelope

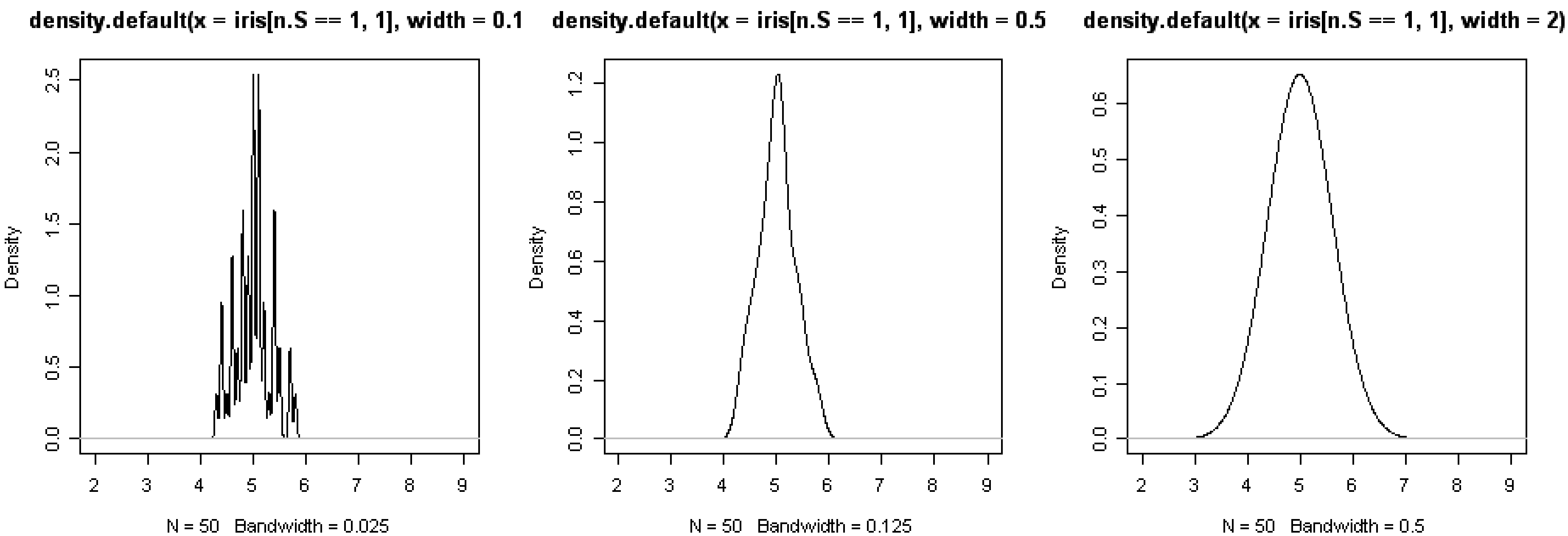
- The `sm` package for **R** allows you to plot variability bands that are a width of two standard errors
- These bands can be especially useful for assessing modality
- More details are in Bowman and Azzalini (1997: Chapter 2)




```
>library(sm)
>sm.density(income, display="se")
```




```
sm.density(income, display="se",model="normal")
```


density.default(x = iris[n.S == 1, 1], width = 0.1
Density
N = 50 Bandwidth = 0.025
density.default(x = iris[n.S == 1, 1], width = 0.5
Density
N = 50 Bandwidth = 0.125
density.default(x = iris[n.S == 1, 1], width = 2)
Density
N = 50 Bandwidth = 0.5

多维密度估计（h一致，h不一致）

定义8.2 假设数据$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$ 是d维向量，并取自一个连续分布$p(\boldsymbol{x})$，在任意点$\boldsymbol{x}$处的一种核密度估计定义为

$$\hat{p}(\boldsymbol{x}) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right), \tag{9.7}$$

注意到这里$p(\boldsymbol{x})$是一个d维随机变量的密度函数。$K(\cdot)$ 是定义在d维空间上的核函数，即$K:\mathbb{R}^d \to \mathbb{R}$，并满足如下条件：

$$K(\boldsymbol{x}) \geqslant 0, \quad \int K(\boldsymbol{x})\,\mathrm{d}\boldsymbol{u} = 1.$$

$$K_n(\boldsymbol{x}) = (2\pi)^{-d/2}exp(-\boldsymbol{x}^T\boldsymbol{x}/2)$$

$$K_2(\boldsymbol{x}) = 3\pi^{-1}(1-\boldsymbol{x}^T\boldsymbol{x})^2 I(\boldsymbol{x}^T\boldsymbol{x}<1)$$

$$K_3(\boldsymbol{x}) = 4\pi^{-1}(1-\boldsymbol{x}^T\boldsymbol{x})^3 I(\boldsymbol{x}^T\boldsymbol{x}<1)$$

$$K_e(\boldsymbol{x}) = \frac{1}{2}c_d^{-1}(d+2)(1-\boldsymbol{x}^T\boldsymbol{x}) I(\boldsymbol{x}^T\boldsymbol{x}<1)$$

K_e被称为多维Epanechinikow核函数，其中c_d是一个和维度有关的常数，$c_1 = 2, c_2 = \pi, c_3 = 4\pi/3$.

$$\hat{p}(\boldsymbol{x}) = \frac{1}{nh_1\cdots h_d}\sum_{i=1}^{n} K\left(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{\boldsymbol{h}}\right)$$

二元密度估计

- The kernel smoothing method used for histograms can be easily extended to the joint distribution of two random continuous variables
- The bivariate density function takes the following form:

$$\hat{p}(x_1, x_2) = \frac{1}{nh_1h_2} \sum_{i=1}^{n} K\left(\frac{x_1 - X_{1i}}{h_1}\right) K\left(\frac{x_2 - X_{2i}}{h_2}\right)$$

- Where K is the kernel function and (h_1 and h_2) are the joint smoothing parameters
- For univariate densities, probabilities are associated with ***area*** under the density curve. For a bivariate density curve, probabilities are associated with ***volume*** under the density, where the total volume equals one

关于最优带宽的选择，我们也有类似一维情况下的结论。对于多维核密度估计，利用多维泰勒展开，我们有

$$Bias(\boldsymbol{x}) \approx \frac{1}{2}h^2\alpha\nabla^2 p(\boldsymbol{x}),$$
$$V(\widehat{p}(\boldsymbol{x})) \approx n^{-1}h^{-d}\beta p(\boldsymbol{x}).$$

其中，$\alpha = \int x^2 K(x)dx$, $\beta = \int K(x)^2 dx$.

因此我们可以得到渐进积分均分误

$$AMISE = \frac{1}{4}h^4\alpha^2\int \nabla^2 p(\boldsymbol{x})d\boldsymbol{x} + n^{-1}h^{-d}\beta.$$

由此可得最优带宽为

$$h_{opt} = \left\{d\beta\alpha^{-2}(\int \nabla^2 p(\boldsymbol{x})d\boldsymbol{x})\right\}^{1/(d+4)} n^{-1/(d+4)}$$

在上述的最优带宽中，真实密度$p(\boldsymbol{x})$是未知的，因此我们可以采用多维正态密度$\phi(\boldsymbol{x})$来代替，进而得到

$$h_{opt} = A(K)n^{-1/(d+4)},$$

$$\text{其中} A(K) = \left\{d\beta\alpha^{-2}(\int \nabla^2 \phi(\boldsymbol{x})d\boldsymbol{x})\right\}^{1/(d+4)}$$

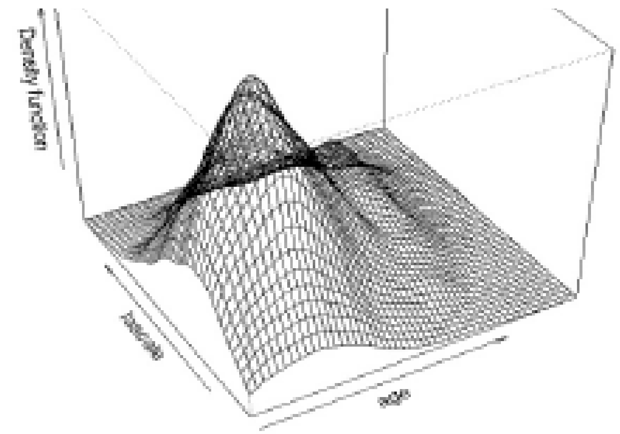
对于$A(K)$，在知道估计中的核函数类型后，可以计算出来，并进而得到最优带宽h_{opt}. 以下是不同核函数的$A(K)$，

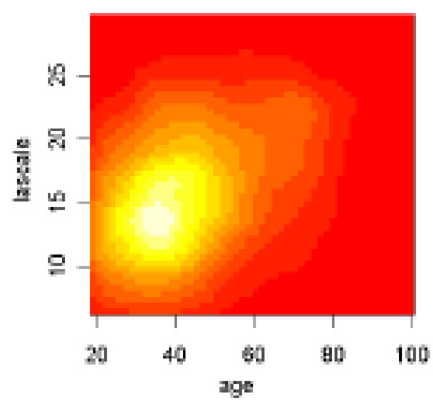
Kernel	Dimensionality	$A(K)$
K_n	2	1
K_n	d	$\{4/(d+2)\}^{1/(d+4)}$
K_e	2	2.40
K_e	3	2.49
K_e	d	$\{8c_d^{-1}(d+4)(2\sqrt{\pi})\}^{1/(d+4)}$
K_2	2	2.78
K_3	2	3.12

- *Perspective plots*: the joint distribution is shown in a 3D plot—height is used to show level of density
- *Imageplots*: different intensities of colour or shading denote density levels
- *Contour plots* or *slice plots*: lines trace paths of constant levels of density (similar to the depiction of elevation in a geographical contour map)

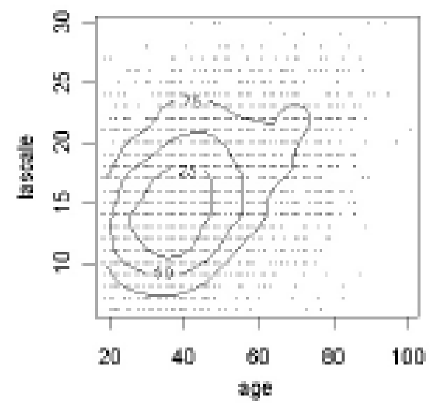
```
#sm library is needed for bivariate density plots below
>library(sm)
>data<-cbind(age,lascale)
#perspective plot is the default
>sm.density(data)
>sm.density(data, display="image")
>sm.density(data, display="slice")
```

Perspective Plot

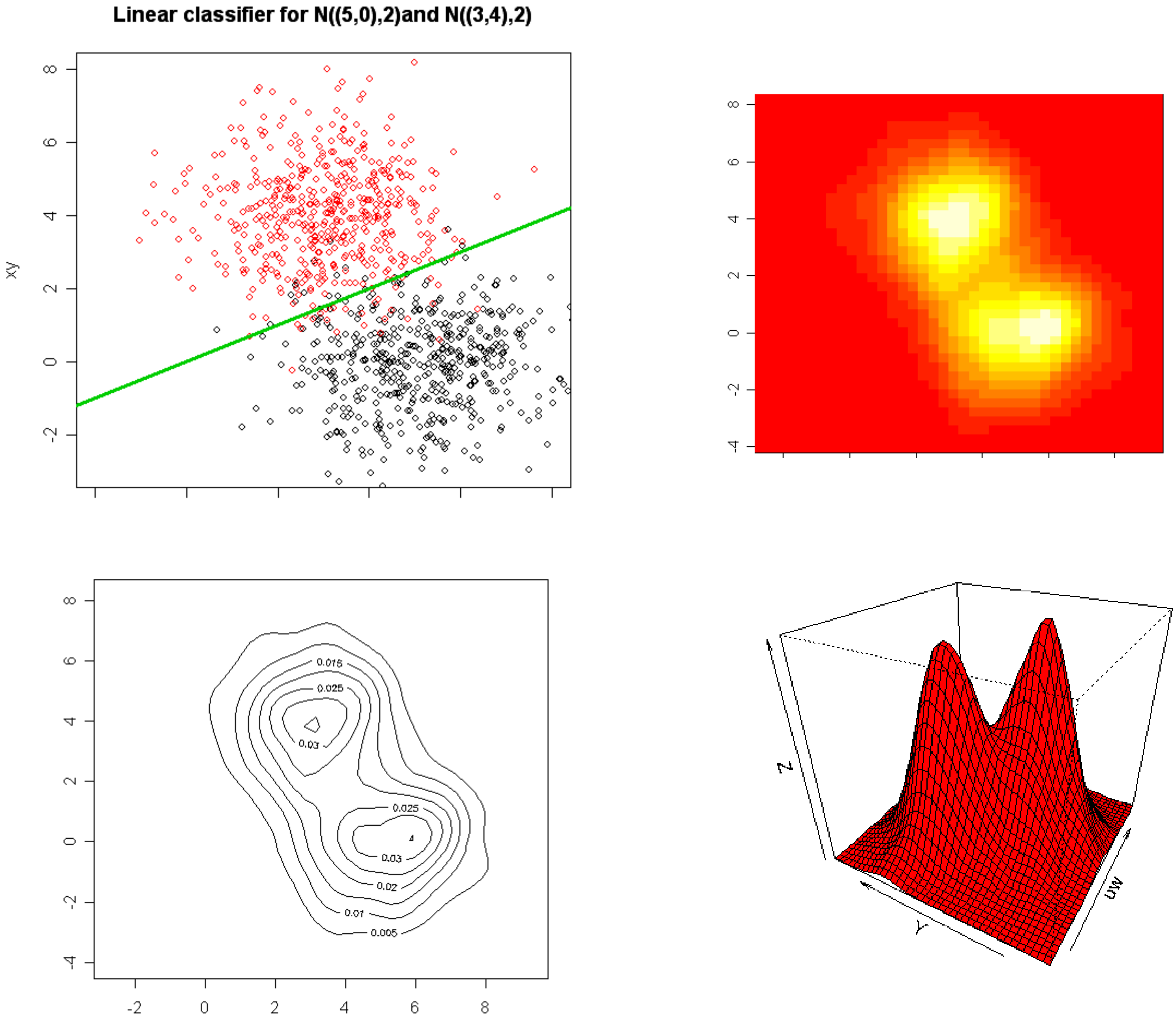




Imageplot

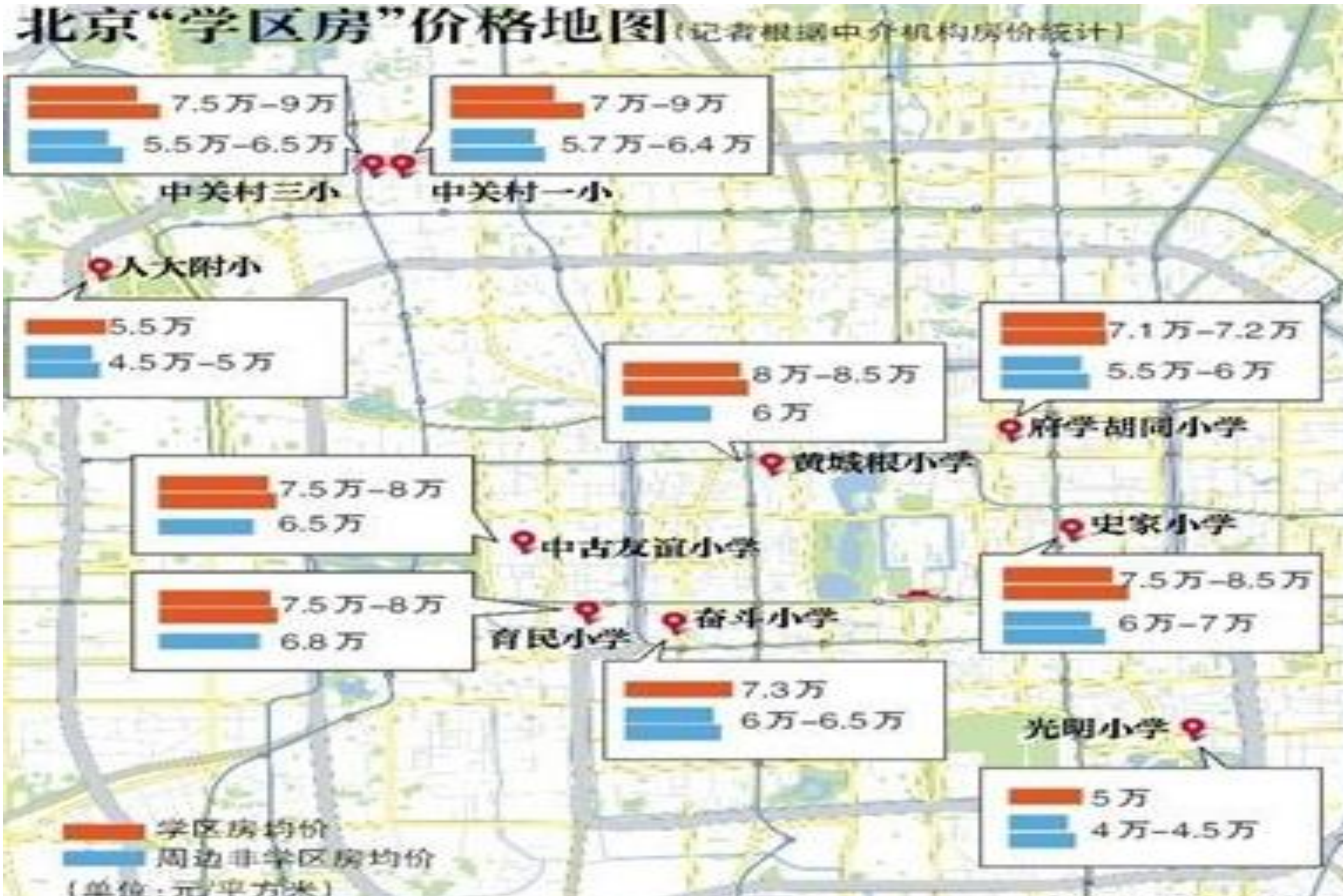


Contour plot



Linear classifier for N((5,0),2)and N((3,4),2)
xy
8
6
4
2
0
-2
-4
0.005
0.01
0.015
0.02
0.025
0.03
-2
0
2
4
6
8
Z
Y
uw

课堂作业和讨论：北京市学区房价格分布与周边价格密度估计



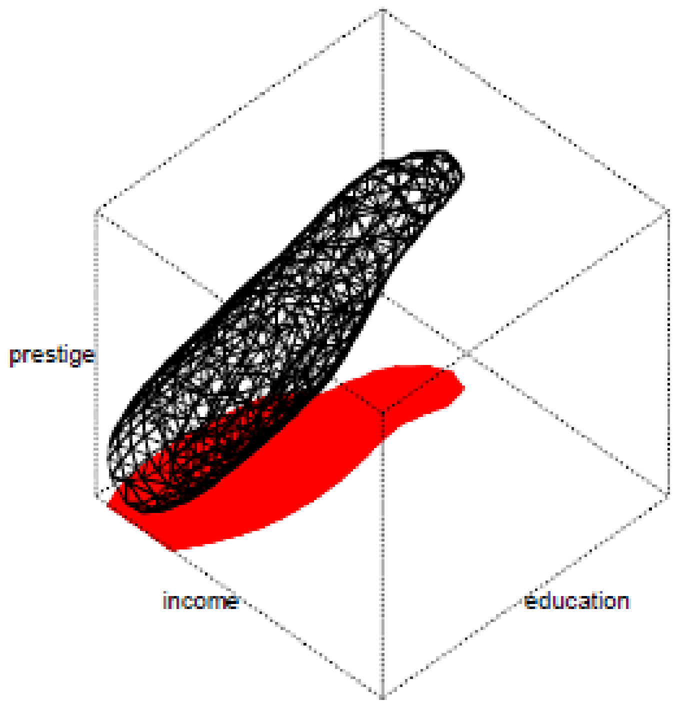
三维密度估计

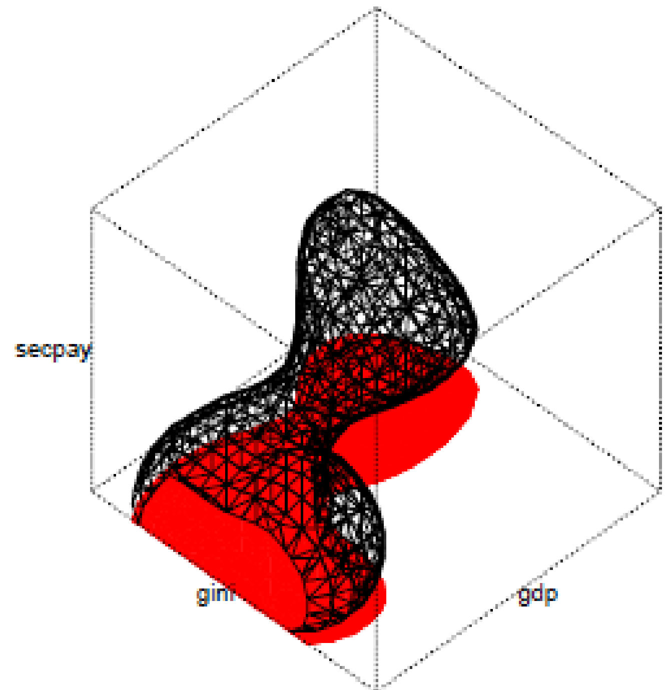
- The three dimensional density estimate also extends simply from the bivariate case:

$$\hat{p}(x_1, x_2, x_3) = \frac{1}{nh_1h_2h_3} \sum_{i=1}^{n} K\left(\frac{x_1 - X_{1i}}{h_1}\right) K\left(\frac{x_2 - X_{2i}}{h_2}\right) K\left(\frac{x_3 - X_{3i}}{h_3}\right)$$

- Where K is the kernel function and (h_1, h_2, and h_3) are the joint smoothing parameter
- In these plots contours represent ***closed surfaces***
- Like the other density estimates, these are helpful for assessing clustering of the data

Some examples of three dimensional density estimates

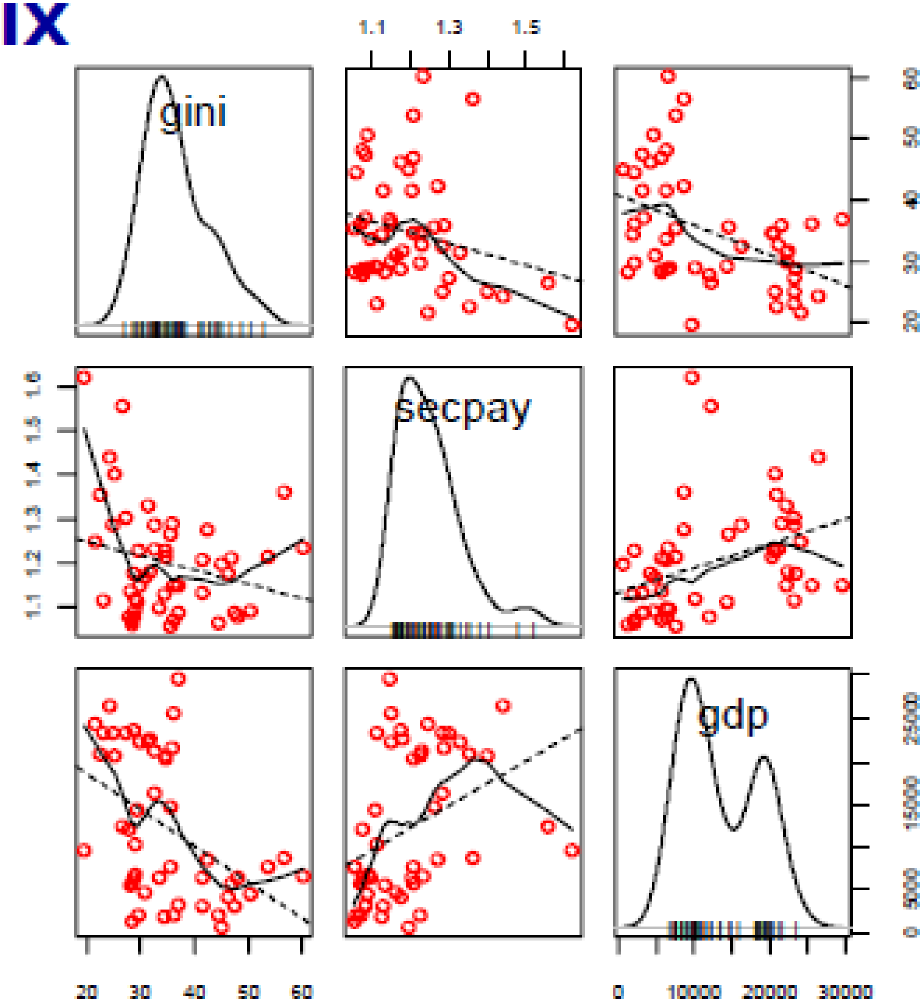





```
>y <- cbind(income, prestige, education)
>sm.density(y)
```

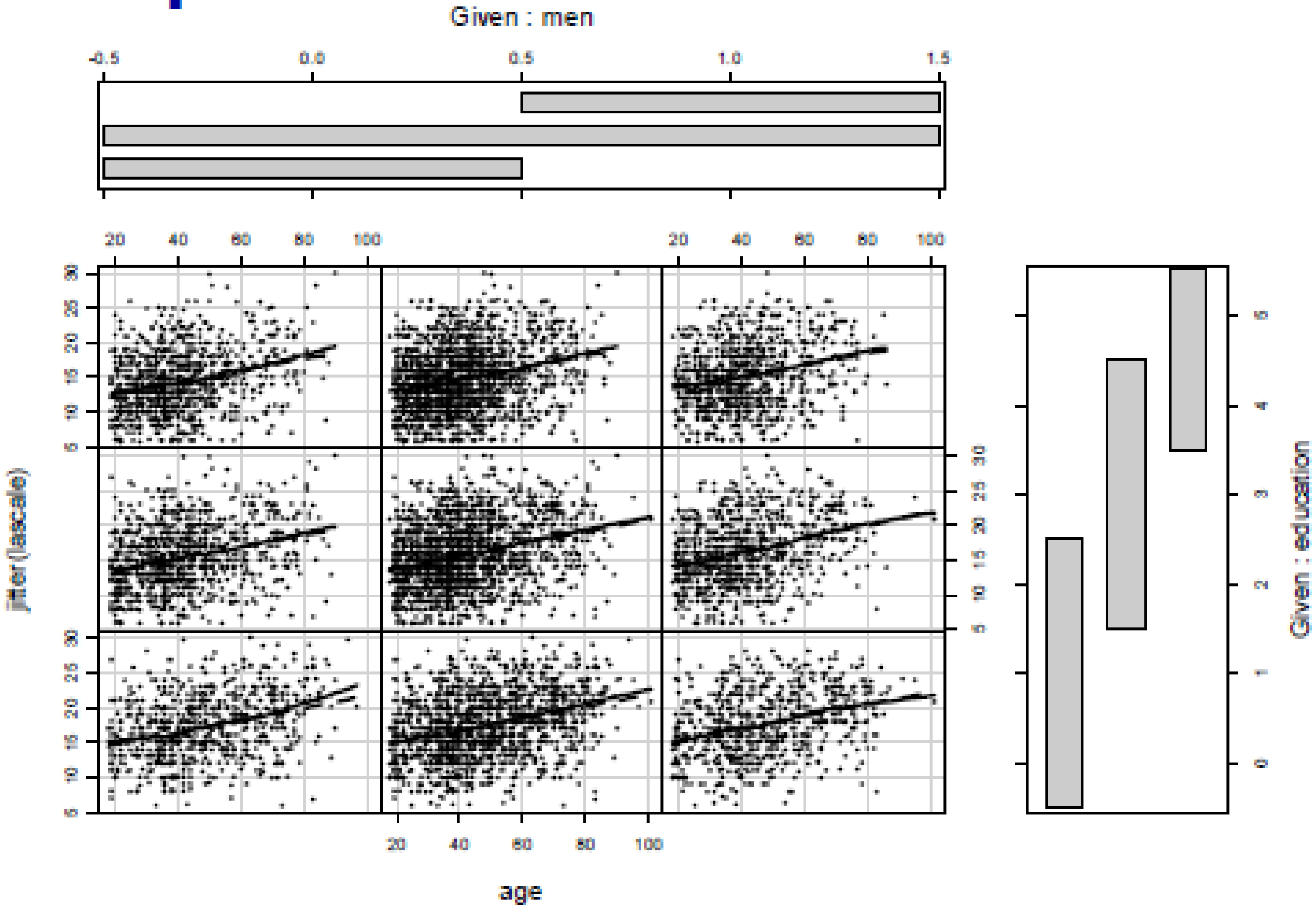
Scatterplotmatrix

- Plots individual scatterplots for all possible bivariate relationships at one time
- Can be enhanced by adding density estimates for each variable on the diagonal
- ***Note:*** Only *marginal relationships* are depicted (*i.e.*, no control for other variables)




```
>library(car)
>scatterplot.matrix(cbind(gini, secpay, gdp))
```

Conditioning plots:
An example from the CES




```
>library(car)
>coplot(jitter(lascale)~age|men+education,
      panel = panel.car, lwd=3, cex=0.4)
```

判别分析

– Classification example

In classifiers based on Parzen-window estimation:

- We estimate the densities for each category and classify a test point by the label corresponding to the maximum posterior
- The decision region for a Parzen-window classifier depends upon the choice of window function as illustrated in the following figure.

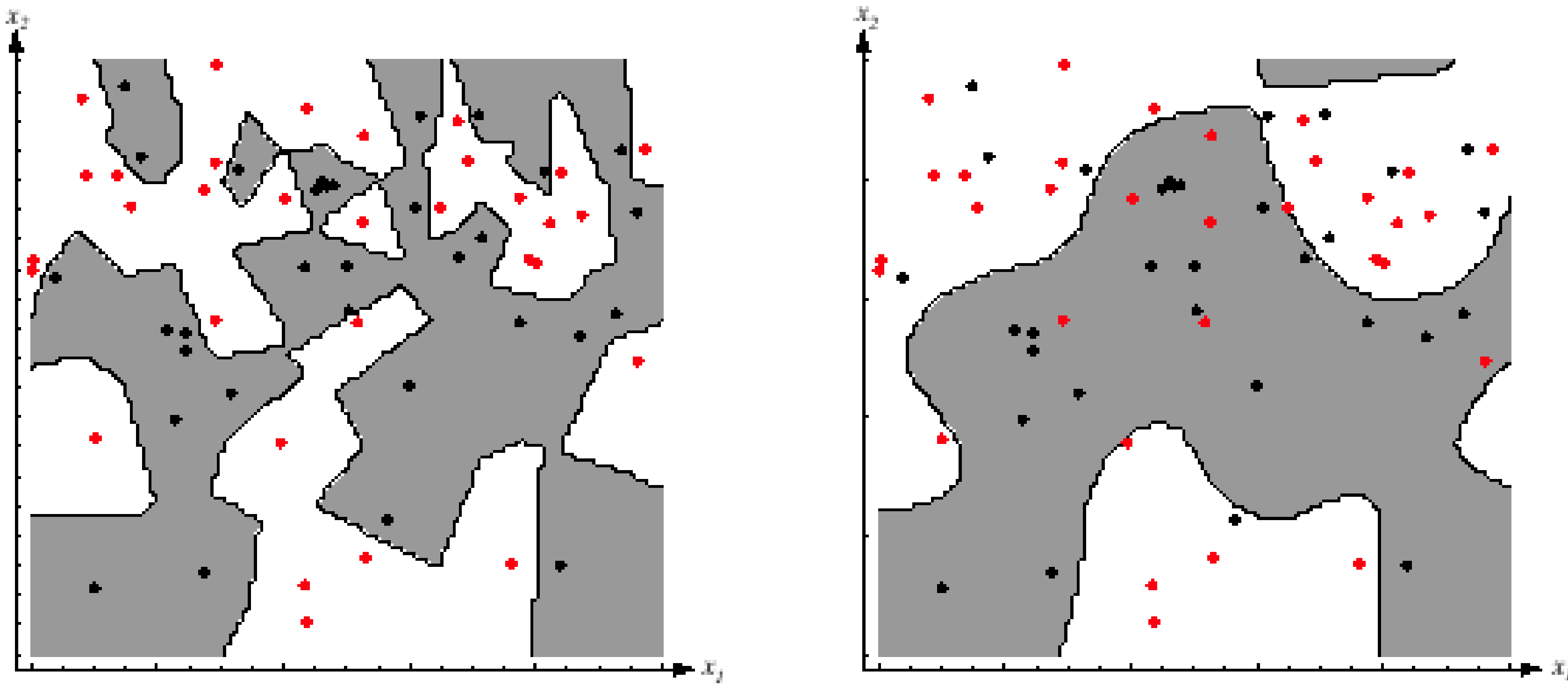
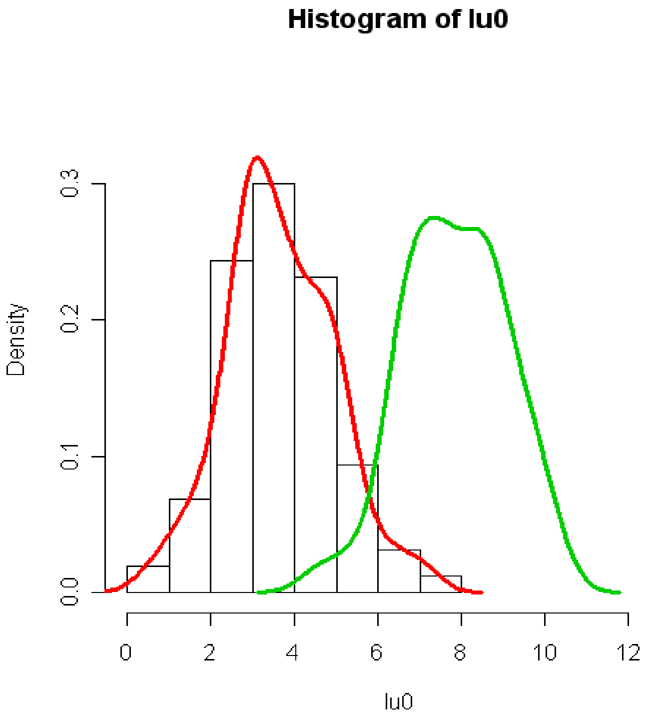


FIGURE 4.8. The decision boundaries in a two-dimensional Parzen-window dichotomizer depend on the window width h. At the left a small h leads to boundaries that are more complicated than for large h on same data set, shown at the right. Apparently, for these data a small h would be appropriate for the upper region, while a large h would be appropriate for the lower region; no single window width is ideal overall. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.

- The sea bass/salmon example
- Decision rule with only the prior information
  - Decide ω_1 if $P(\omega_1) > P(\omega_2)$ otherwise decide ω_2

- $P(x \mid \omega_1)$ and $P(x \mid \omega_2)$ describe the difference in lightness between populations of sea bass and salmon



例：基于非参数密度估计下的判别计算(二分类问题求解步骤)

- 1. 先验密度, 损失矩阵→计算域值.
- 2. 非参数似然密度估计→生成判别决策.
- 3. 给出新的点,比较判别决策的的判定.

Bayes' Rule

- Posterior, likelihood, evidence

posterior *likelihood* *prior*

  - $P(\omega_j \mid x) = P(x \mid \omega_j) \,.\, P(\omega_j) \,/\, P(x)$

evidence

  - Where in case of two categories

$$P(x) = \sum_{j=1}^{j=2} P(x \mid \omega_j) P(\omega_j)$$

  - Posterior = (Likelihood. Prior) / Evidence

更一般的Bayes公式的解释

假设空间: H={H_1, …, H_n}　　　　样本和数据: E

$$P(H_i \mid E) = \frac{P(E \mid H_i)P(H_i)}{P(E)}$$

If we want to pick the most likely hypothesis H*, we can drop P(E)

Posterior probability of H_i ↓　　　　**Prior probability of H_i** ↓

$$P(H_i \mid E) \propto P(E \mid H_i)P(H_i)$$

↑ **Likelihood of data/evidence if H_i is true**

- Decision given the posterior probabilities

  X is an observation for which:

  if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$ $\Longrightarrow$ True state of nature = ω_1

  if $P(\omega_1 \mid x) < P(\omega_2 \mid x)$ $\Longrightarrow$ True state of nature = ω_2

因此:

当观察到某个 x, 我们各种决定可能的错误是:

$P(\text{判错} \mid x) = P(\omega_1 \mid x)$ *如果决策是* ω_2

$P(\text{判错} \mid x) = P(\omega_2 \mid x)$ *if we decide* ω_1

- Minimizing the probability of error

- Decide $\omega 1$ *if* $P(\omega 1 \mid x) > P(\omega 2 \mid x)$; otherwise decide $\omega 2$

- 因此有关判错可以有如下的等价表达:

$$P(error \mid x) = P(\omega_1 \mid x) \text{ if we decide } \omega_2$$
$$P(error \mid x) = P(\omega_2 \mid x) \text{ if we decide } \omega_1$$

$$\Updownarrow$$

$$P(error \mid x) = \min [P(\omega 1 \mid x), P(\omega 2 \mid x)]$$

The preceding rule is equivalent to the following rule:

$$if \ \frac{P(x|\omega_1)}{P(x|\omega_2)} > \frac{l_{12}-l_{22}}{l_{21}-l_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

Then take action α_1 (decide ω_1)

Otherwise take action α_2 (decide ω_2)

结论: 贝叶斯决策规则可以解释成如果似然比超过某个不依赖于观测值x的阈值,那么判断为ω_1.

例：基于非参数密度估计下的判别计算

- State:{ω_1, ω_2},
- Action :

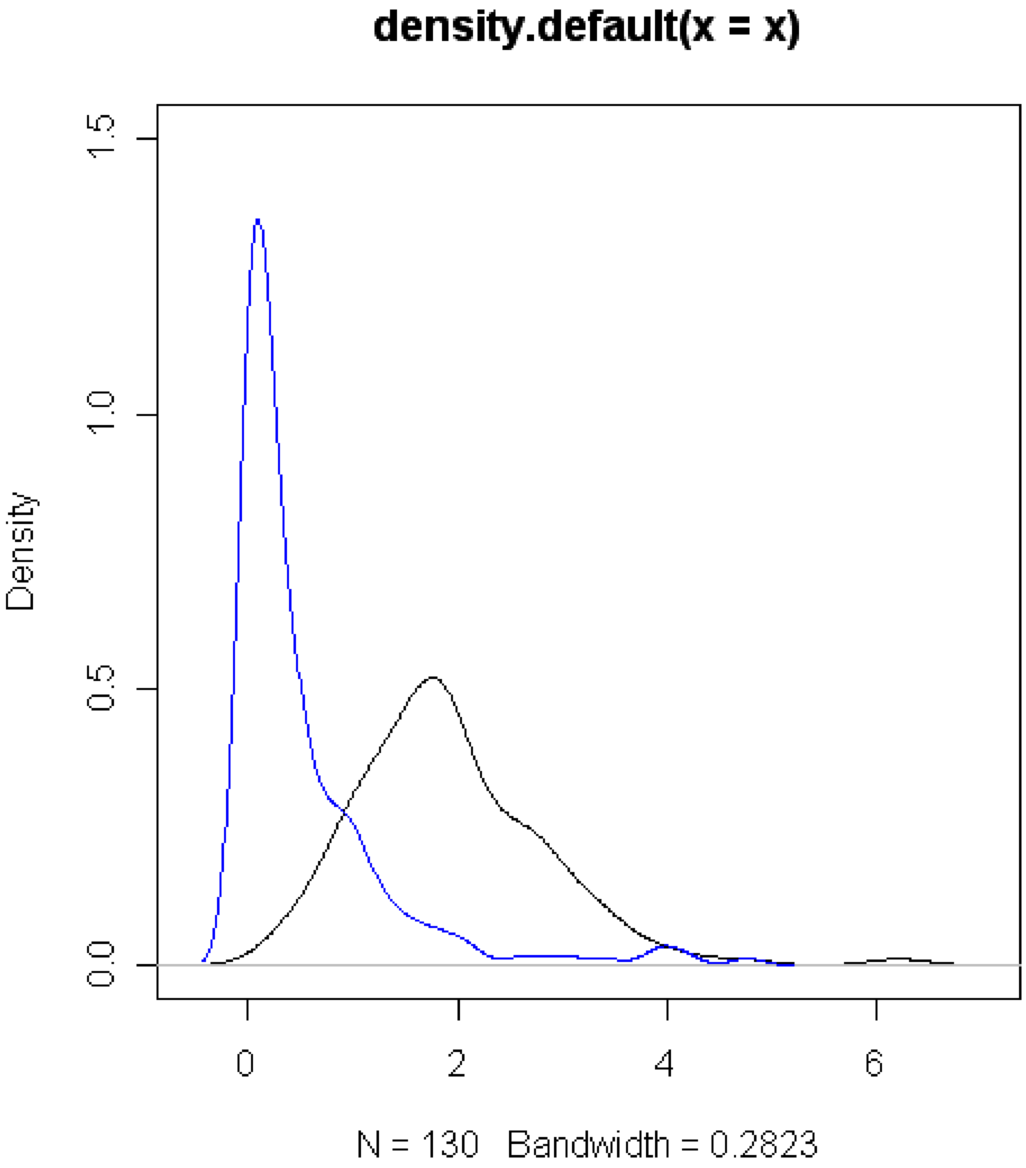
  α_1 : *deciding* ω_1

  α_2 : *deciding* ω_2

- The preceding rule is equivalent to the following rule:

$$if \ \frac{P(x \mid \omega_1)}{P(x \mid \omega_2)} > \frac{l_{12} - l_{22}}{l_{21} - l_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

- Then take action $\alpha 1$ (decide $\omega 1$)
- Otherwise take action $\alpha 2$ (decide $\omega 2$)



两类不同鱼光泽度的分布密度:

$$L = \begin{matrix} 0 & 1 \\ 2 & 0 \end{matrix}$$

newpoint=2 class=1

newpoint=0.1 class=2

本章要求

- 掌握密度估计基本原理；
- 掌握几种多维可视化的建模方法
- 了解密度估计的应用