贝叶斯学习（为了反映样本数目）是一个过程

样本集重新定义 $X^N=\{x_1, x_2, \ldots x_N\}$ 不喜欢这种不一致定义

贝叶斯估计 $\theta^* = \int_{\Theta} \theta p(\theta|X^N)\, d\theta$

当 $N>1$，有 $p(X^N|\theta) = p(x_N|\theta) \cdot p(X^{N-1}|\theta)$ IID 假设

$$p(\theta|X^N) = \frac{p(X^N|\theta)\, p(\theta)}{\int_{\Theta} p(X^N|\theta)\, p(\theta)\, d\theta}$$ 分子分母相同~~形~~式

为了用 X^N 求出

$$p(X^N|\theta) \cdot p(\theta) = p(x_N|\theta) \cdot p(X^{N-1}|\theta) \cdot p(\theta)$$

$$= p(x_N|\theta) \cdot \frac{p(X^{N-1}, \theta)}{p(\theta)} \cdot p(\theta)$$

$$= p(x_N|\theta) \cdot p(X^{N-1}, \theta)$$

$$= p(x_N|\theta) \cdot p(\theta|X^{N-1}) \cdot p(X^{N-1})$$

X^{N-1} 对 $d\theta$ 不是变量 故分母积分下 X^{N-1} 可提出，与分子约分。

$$\int_\Theta p(X^N,\theta)\,d\theta = p(X^N)$$

直接估计出样本的概率密度

贝叶斯估计把θ看作Random variable

递推的贝叶斯估计：

$p(\theta)$ $p(\theta|x_1)$ $p(\theta|x_1,x_2)$ $\cdots$ $p(\theta|x_1,\cdots)$

$p(x|X^{N\to\infty}) = p(x)$ 真实密度函数.

正态分布时的贝叶斯估计

$x \sim N(\mu, \sigma^2)$

$\mu \sim N(\mu_0, \sigma_0^2)$ 假设 μ 服从先验分布

根据 X 估计 μ，用贝叶斯估计化简

$$P(\mu|X) \sim N(\mu_N, \sigma_N^2)$$

$$\mu_N = \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2} m_N + \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0$$

$$\sigma_N^2 = \frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}$$

$$m_N = \frac{1}{N}\sum_i x_i$$

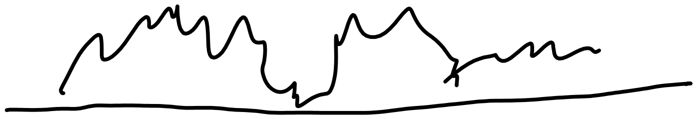
μ_N 由2部分组成 m_N 和 μ_0（先验知识），当 $N \to \infty$

$\mu_N = m_N$

当 $N \to 0$，σ_0^2 很小，即先验知识确定 $\mu_N = \mu_0$

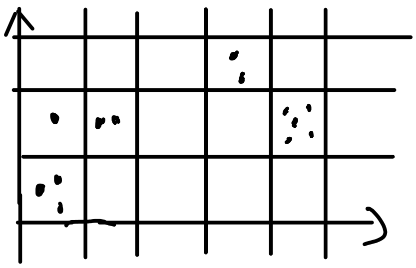
$x|X \sim N(\mu_N, \delta^2+\delta_N^2)$ 用样本估计的 $P(x)$ 方差变大

概率密度估计的非参数方法.



写不出 $p(x)$ 的形式怎么办?

• 直方图法.



(1) 把样本 x 的每个分量在其取值范围内分成 k 个等间隔的小窗。如果 x 是 d 维向量，则这种分割就会得到 k^d 个小体积或者称作小舱，每个小舱的体积记作 V。

(2) 统计落入每个小舱内的样本数目 q_i。

(3) 把每个小舱内的概率密度看作是常数，并用 $q_i/(NV)$ 作为其估计值，其中 N 为样本总数。

分析非参数估计基本原理：

x 落入 R_1 区域 region-1 的概率

$$P = \int_{R_1} p(x)\,dx$$

k 个样本落入 R_1 的概率

$$P_k = C_N^k P^k \cdot (1-P)^{N-k}$$

在 R_1 中 k 的期望

$$E(k) = NP \quad , \quad \hat{P} = \frac{k}{N}$$

在 R_1，k 的众数（最可能的取值）.

$$P_m = \max P_k \quad , \quad m = [(N+1)P] \text{，} [\ ] \text{为取整.}$$

$$\hat{P} = \int_{R_1} p(x)\,dx = p(x)\cdot V.$$

将 $p(x)$ 在这一个小区域看作定值.

$$\hat{p}(x) = \frac{\hat{P}}{V} = \frac{k/N}{V}$$

所以，小舱的选择应该与样本总数相适应。理论上讲，假定样本总数是 n，小舱的体积为 V_n，在 x 附近位置上落入小舱的样本个数是 k_n，那么当样本趋于无穷多时 $\hat{p}(x)$ 收敛于 $p(x)$ 的条件是

$$(1)\ \lim_{n\to\infty} V_n = 0, \quad (2)\ \lim_{n\to\infty} k_n = \infty, \quad (3)\ \lim_{n\to\infty} \frac{k_n}{n} = 0 \tag{3-62}$$

直观的解释是：随着样本数的增加，小舱体积应该尽可能小，同时又必须保证小舱内有充分多的样本，但每个小舱内的样本数又必须是总样本数中很小的一部分。

不足：样本密度大的区域，一个小舱有很多样本，密度小的区域一个小舱只有很少或没有样本，导致密度估计在样本密度不同的区域不一致。

k_n近邻估计方法

$$\hat{p}(x) = \frac{k_N/N}{V}$$

在任一点x，调整V的体积，直到恰好有k_N个样本落入，兼顾在高密度区域的分辨率和低密度区域连续性。

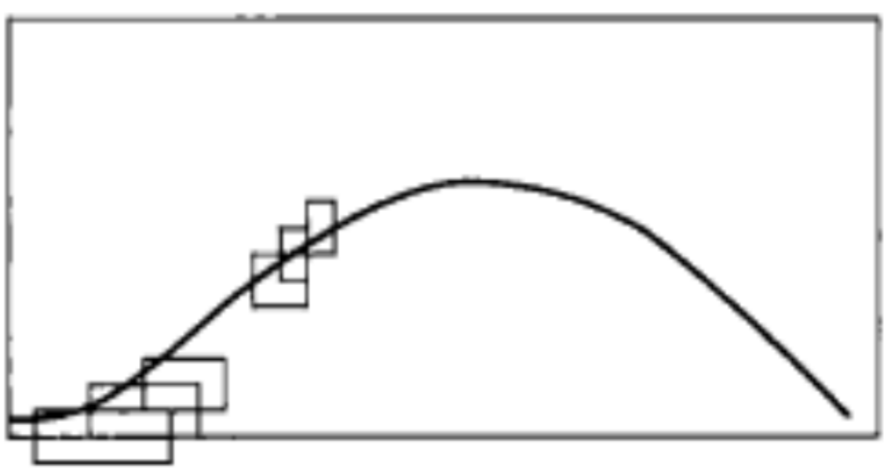
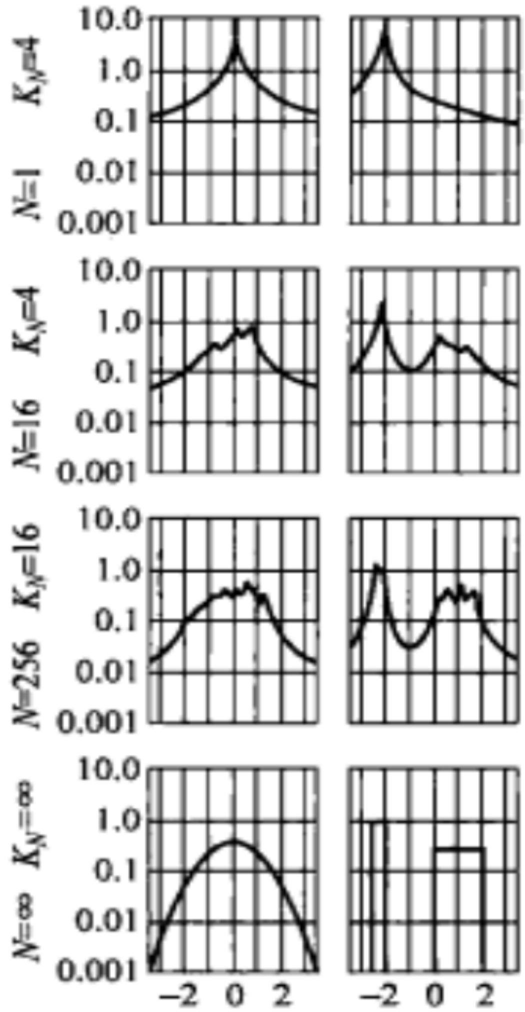


图 3-4 k_N 法的窗口宽度与样本密度的关系示意



N=∞即真实的p(x)

图 3-5 不同样本数和不同参数下 k_N 法估计的效果举例

Parzen 窗法

体积 $V = h^d$

$$\varphi([u_1, u_2 \cdots u_d]) = \begin{cases} 1, & 若 |u_j| \le \frac{1}{2} \\ 0, & 其他 \end{cases}$$

$$k_N = \sum_{i=1}^{N} \varphi\left(\frac{x - x_i}{h}\right)$$

将直方图估计的 k_N 替换为了

自适应 k_N

$$\hat{p}(x) = \frac{1}{NV}\sum_{i=1}^{N}\varphi\left(\frac{x-x_i}{h}\right)$$

从核函数角度

$$K(x, x_i) = \frac{1}{V}\varphi\left(\frac{x-x_i}{h}\right)$$

反映了一个观测样本 x_i 对在 x 处

概率密度估计的贡献，与

x_i 到 x 的距离有关

ATTENTION

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N} K(x, x_i)$$

$K(x, x_i)$ 的条件

$$\int \hat{p}(x)\,dx = \frac{1}{N}\sum_{i=1}^{N}\int K(x, x_i)\,dx = 1$$

$$\sum_{i=1}^{N}\int K(x, x_i)\,dx = N$$

$\int K(x, x_i)\,dx = 1$ 是满足的

$K(x, x_i) \geq 0$. 核函数条件

核函数可定义多种.

图 3-6 给出用高斯窗估计一个概率密度函数的例子。

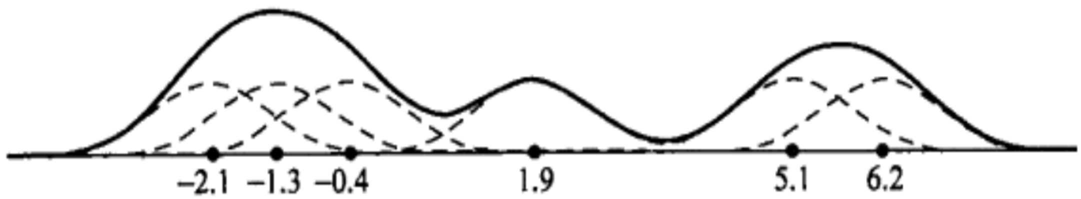


图 3-6 Parzen 窗估计概率密度函数示例

叠加效应

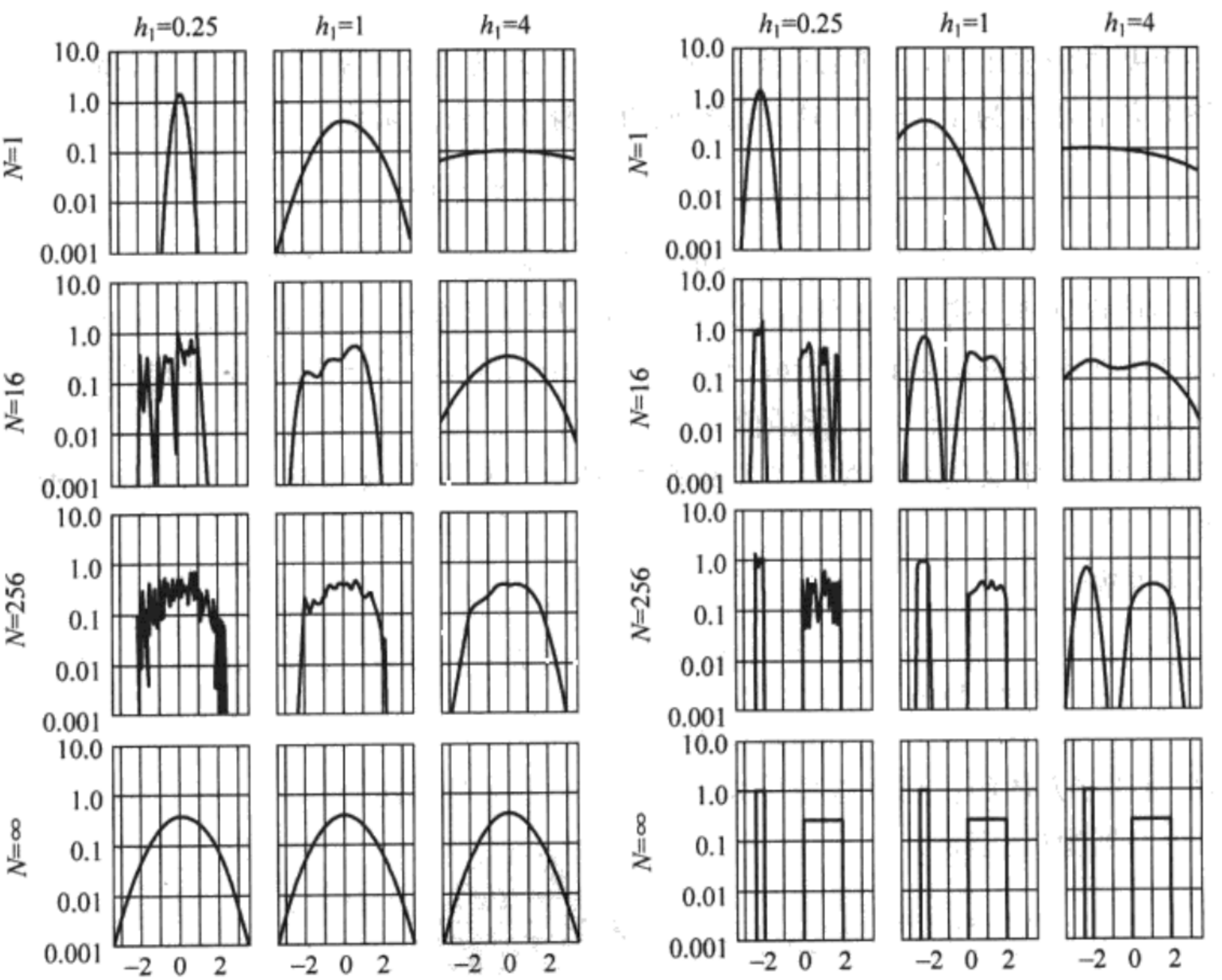


图 3-7 不同样本数和不同参数下 Parzen 窗估计的效果示例

窗口宽度h是平滑参数，h越大越平滑。

Non-parametric的不足

① 对样本数目需求大

② 计算存储量大

优点：能收敛于任何复杂未知密度.