# 概率密度估计与非参数回归

曾焰

版本 1.0，最后修订于 2017-11-05

**摘要**

陈希孺等 [1] 第六章的内容摘要。

## 1 概率密度估计

### 1.1 几种重要的密度估计方法

1. 直方图法。这个方法可描述如下：假设随机变量 $X$ 有密度 $f$，并有 $X$ 的独立同分布样本 $X_1$，$\cdots$，$X_n$。选择一个适当的正数 $h$，把全直线分为一些长为 $h$ 的区间。任取这些区间之一，记为 $I$。对 $x \in I$，我们有

$$f(x) \approx \frac{P(X \in I)}{h} \approx \frac{\sum_{i=1}^{n} 1_{\{X_i \in I\}}}{n} \cdot \frac{1}{h} \tag{1.1}$$

这一方法重要的是 $h$ 的选择。$h$ 太大了，平均化的作用突出了，而淹没了密度的细节部分。太小了，则受随机性影响太大，而产生极不规则的形状。$h$ 的选择无现成规则可循。实际操作中，我们可能需要取一些不等长的区间，这样的直方图估计称为“Data-based”的直方图估计。

直方图估计的优点是简单易行，缺点是它不是连续函数（这可以通过适当地修匀来解决），且从统计角度看一般说效率较低。例如，在这一方法下，每一区间中心部分密度估计较准，而边缘部分则较差。

2. Rosenblatt 法。为克服直方图法的一个缺点——对每个区间边缘部分密度值的估计较差，Rosenblatt 在 1955 年提出了一个简单的改进。指定一个正数 $h$，对每个 $x$，定义 $I_x = \left[x - \frac{h}{2}, x + \frac{h}{2}\right]$，并对密度函数 $f$ 作如下估计

$$f_n(x) \triangleq f_n(x; X_1, \cdots, X_n) = \frac{\sum_{i=1}^{n} 1_{\{X_i \in I_x\}}}{n} \cdot \frac{1}{h} \tag{1.2}$$

Rosenblatt 法与直方图法不同之处仅在于，它事先不把分割区间定下来，而让区间随着要估计之点 $x$ 跑，使 $x$ 始终处在区间之中心位置，而获致较好的效果。理论上可以证明，从估计量与被估计量接近的数量级上看，Rosenblatt 方法确实优于直方图法。

3. Parzen 的核估计。直观上可以设想：为估计 $f(x)$，与 $x$ 靠近的样本，所起作用似应比远离 $x$ 的样本要大些。这些在 Parzen 于 1962 年提出的核估计方法中都得到了体现。为介绍 Parzen 的思想，我们先将 (1.2) 式变换一个形式，引进一个函数

$$W(x) = I_{\left[-\frac{1}{2}, \frac{1}{2}\right]}(x)$$

则 (1.2) 式可改写为

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

$W(\cdot)$ 定义的是 $\mathbb{R}^1$ 上的均匀密度函数。Parzen 的推广即在于去掉这一特殊性，而容许 $W$ 为一般的密度函数。

**定义 1.1.** 设 $K(\cdot)$ 是 $\mathbb{R}^1$ 上的一个给定的概率密度函数，$h_n > 0$ 是一个同 $n$ 有关的常数，定义

$$f_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) \tag{1.3}$$

称 $f_n$ 为总体未知密度 $f$ 的一个**核估计**，$K$ 为**核函数**，$h_n$ 为**窗宽**。[1]

在给定样本之后，一个核估计性能的好坏，取决于核及窗宽的选取是否适当。当 $h_n$ 选得过大，由于 $x$ 经过压缩变换 $\frac{x-X_i}{h_n}$ 之后使分布的主要部分的某些特征（如多峰性）被掩盖起来了，估计量有较大偏差。如 $h_n$ 太小，整个估计特别是尾部出现较大的干扰，从而有增大方差的趋势。因而在实际使用核估计时，如何选取适当的宽度是一项很细致的工作。选择核 $K$ 是否适当，同样要影响估计的精度。原则上，我们可以对核 $K$ 施加一定的限制，使得估计量与待估函数的偏差在一定意义下尽可能地小。例如可以要求 $K$ 有对称性，其一阶矩（关于密度 $K$）为零，具有有界性、连续性等等。在文献中，核估计已成为密度估计的主要方法。

4. 最近邻估计。这一方法较适合于密度的局部估计。其要旨如下：设 $X_1$，$\cdots$，$X_n$ 是来自未知密度 $f$ 的样本。先选定一个同 $n$ 有关的整数 $k = k_n$，合于 $1 \le k < n$，对固定的 $x \in \mathbb{R}^1$，记 $a_n(x)$ 为最小的正数 $a$ 使得 $[x-a, x+a]$ 中至少包含 $X_1$，$\cdots$，$X_n$ 中的 $k$ 个。定义

$$\hat{f}_n(x) = \frac{k_n}{2a_n(x)n} \tag{1.4}$$

为 $f(x)$ 的估计，称 $\hat{f}_n$ 为 $f$ 的**最近邻估计**（简记为 **N.N. 估计**）。下面的引理说明：从整体看，N.N. 估计的性质与核估计有很大的不同。

**引理 1.1.** *(1)* 对固定 $n$ 及 $X_1$，$\cdots$，$X_n$，$\hat{f}_n(x)$ 作为变元 $x$ 的函数是处处连续的。

*(2)*$\hat{f}_n(x)$ 作为变元 $x$ 的函数非概率密度，并且

$$\hat{f}_n(x) = O\left(\frac{1}{n}\right), \text{当 } |x| \to \infty。$$

特别地，我们有

$$\int \hat{f}_n(x)dx = \infty$$

引理1.1的性质（2）与待估 $f$ 的尾部特征无关，因而对相当一类待估密度，估计 $\hat{f}_n(x)$ 的尾部衰减得太慢，从而 $\hat{f}_n$ 不适宜用作 $f$ 的整体估计。下面的引理给出了 $\hat{f}_n(x)$ 的分布。

**引理 1.2.** 对固定 $x \in \mathbb{R}^1$，$n \ge 1$，有

$$P(a_n(x) \le y) = \sum_{i=k}^{n} C_n^i p^i(y)(1-p(y))^{n-i} = nC_{n-1}^{k-1} \int_0^{p(y)} t^{k-1}(1-t)^{n-k}dt, y > 0,$$

[1]这一定义考虑的是 $X$ 为一维的情况。若 $X$ 为 $d$ 维，只须将 (1.3) 式中分母 $nh_n$ 改为 $nh_n^d$。

其中

$$p(y)=\int_{x-y}^{x+y}f(t)dt=P(x-y\le X\le x+y)$$

如果令 $K(x)=\begin{cases}\frac{1}{2}, & |x|<1\\ 0, & |x|\ge 1\end{cases}$，则可将 N.N. 估计改写为

$$\hat{f}_n(x)=\frac{1}{na_n(x)}\sum_{i=1}^{n}K\left(\frac{x-X_i}{a_n(x)}\right)$$

于是在单个点 $x$ 上的 N.N. 估计与核估计差别不大，只有当同时考虑在几个点或者估计整个 $f$ 时，这两种方法才显示出差别。N.N. 估计由于计算上有某种方便之处，这种方法被广泛地用于模式识别及非参数判别分析。

## 1.2 估计精度的度量

我们用 $T_n(x)\triangleq T_n(x;X_1,\cdots,X_n)$ 表示基于样本 $X_1$，$\cdots$，$X_n$ 的、对未知密度 $f(x)$ 的任一估计。由于 $T_n(x)$ 既同样本有关，又是考察点的函数，因而对固定的考察点 $x$，估计精度的一种自然测度为

$$\mathrm{MSE}(T_n(x))=E_f[(T_n(x)-f(x))^2]=(E_f[T_n(x)]-f(x))^2+\mathrm{Var}_f(T_n(x)),\tag{1.5}$$

称 (1.5) 为估计 $T_n$ 的**均方误差**，其中 $E_f$ 表示期望是在真分布为 $f$ 时的计算。（1.5）右端是由两个部分组成：第一项是偏差项，而第二项是估计的方差。要同时减少这两部分是困难的：通常，如降低偏差，则方差有增大的趋向，反之亦然。例如当 $T_n(x)$ 为核估计时，有

$$E_f[T_n(x)]=\int K(y)f(x-h_ny)dy,$$

$$\mathrm{Var}_f[T_n(x)]=\frac{1}{nh_n}\int K^2(y)f(x-h_ny)dy-\frac{1}{n}\left[\int K(y)f(x-h_ny)dy\right]^2$$

因而一个核估计的光滑程度只与光滑参数 $h_n$ 有关（当核 $K$ 已确定时），而与 $n$ 无直接关系。

对于密度估计来说，更有实际意义的精度的度量应是整体性的测度。一个被广泛使用的整体测度是**积分均方误差**（MISE）：

$$\begin{aligned}\mathrm{MISE}(T_n) &= E\left[\int(T_n(x)-f(x))^2dx\right]=\int\mathrm{MSE}(T_n(x))dx\\ &= \int(E_f[T_n(x)]-f(x))^2dx+\int\mathrm{Var}_f(T_n(x))dx\\ &= \text{积分偏差平方和}+\text{积分方差}\end{aligned}$$

我们在前段对均方误差的分析，同样可施用于积分均方误差。对核估计来说，应该选择 $h_n$ 使得相应的核估计其 MISE 达到最小。

为便于计算及理论分析，我们可以通过泰勒展开，得到估计偏差及方差的渐进表达式。为简单计，设 $K$ 是对称密度函数，满足：$\int tK(t)dt=0$，$k_2\triangleq\int t^2K(t)dt\neq 0$，而 $f$ 具有二阶有界连续导数且 $f''\in L_2(\mathbb{R}^1)$，$h\triangleq h_n\to 0$，当 $n\to\infty$。则有如下渐近公式：

$$\int(E_f[T_n(x)]-f(x))^2dx\approx\frac{1}{4}h^4k_2^2\int[f''(x)]^2dx,\ \int\mathrm{Var}_f(T_n(x))dx\approx(nh)^{-1}\int K^2(u)du$$

合并可得 **MISE** 的渐近公式：

$$\text{MISE} \approx \frac{1}{4} h^4 k_2^2 \int [f''(x)]^2 dx + (nh)^{-1} \int K^2(u) du \tag{1.6}$$

再对上式右端关于 $h$ 求极小，得到**渐近最佳窗宽**

$$h_{opt} = k_2^{-2/5} \left[ \int K^2(u) du \right]^{1/5} \left[ \int [f''(x)]^2 dx \right]^{-1/5} n^{-1/5} \tag{1.7}$$

公式（1.7）表明：最佳渐近窗宽随 $n$ 增大以 $n^{-1/5}$ 的速度趋于零。

如将由（1.7）确定的 $h_{opt}$ 代入（1.6），则有

$$\text{MISE} \approx \frac{5}{4} C(K) \left\{ \int [f''(x)]^2 dx \right\}^{1/5} n^{-4/5}$$

其中

$$C(K) = k_2^{2/5} \left[ \int K^2(t) dt \right]^{4/5}$$

然后可依使 $C(K)$ 尽可能小的原则选择 $K$。从上述公式可看出这样一个事实：不论 $h$ 及 $K$ 如何选取，作为核估计来说，其积分均方误差收敛于零的速度，其主要部分的阶不能超过 $4/5$。这在理论分析上是很有意义的。

### 1.3 密度估计的应用

密度估计的重要性，并不在于它的单独使用，而是作为统计推断的中间环节发挥作用。

1. 非参数判别。设有来自总体 $A$ 的样本 $X_1, \cdots, X_n$，及来自总体 $B$ 的样本 $Y_1, \cdots, Y_m$。今有新的观察 $Z$，问 $Z$ 来自 $A$ 还是 $B$？基于极大似然原理，可定出如下的非参数判别法：分别基于 $X_1, \cdots, X_n$ 及 $Y_1, \cdots, Y_m$ 估计 $f_A$ 及 $f_B$，记估计为 $\hat{f}_A$ 及 $\hat{f}_B$，然后视 $\hat{f}_A(Z) \geq \hat{f}_B(Z)$ 抑或 $\hat{f}_A(Z) < \hat{f}_B(Z)$ 确定 $Z$ 所归属的类。

2. 聚类分析。一种常用的聚类方法即是构造某种“树图”，各个个体按“树图”中的等级归并成若干类，而划分等级的规则需使用密度估计。

3. 随机数的模拟。设已有观察 $X_1, \cdots, X_n$，由于随机影响，这些观察渗杂了某些伪造的细节。我们的目的是模拟一组新数据 $Y_1, Y_2, \cdots$，使得 $Y_1, Y_2, \cdots$ 具有原总体的结构，但无这些伪造的细节。当总体具未知密度 $f$ 时，可用其核估计产生模拟数，例如 $\hat{f}$ 是基于 $X_1, \cdots, X_n$ 的具核 $K$ 及窗宽 $h_n$ 的密度估计，可按如下步骤产生新数据 $Y$：

（1）从数字 1，2，$\cdots$，$n$ 中有放回地随机抽取一个，记为 $I$。（2）产生一个与 $X_1, \cdots, X_n$ 独立的具密度 $K$ 的随机变量 $\varepsilon$。（3）定义 $Y = X_I + h\varepsilon$。

以上过程可不断地重复进行，从而产生一串新数据。易知这样的 $Y$ 有分布密度 $\hat{f}$。

## 2 密度估计的大样本性质

### 2.1 有关概念

由于对未知密度的数学形式没有任何假定，指望得出较为深入的小样本性质是不现实的。迄今为止关于密度估计的研究，几乎全集中在大样本方面。一般来说这本是非参数方法的一个特征。

**定义 2.1.** 如果对每一给定 $x$

$$\lim_{n\to\infty} E_f[T_n(x)] = f(x), \text{对所有可能的 } f$$

则称 $T_n$ 为**渐近无偏估计**。

在相当宽泛的条件下，对固定 $n$，密度函数的无偏估计是不存在的。在不太强的限制下，渐近无偏估计总是存在的。

**定义 2.2.** 如果对固定 $x$，有

$$\lim_{n\to\infty} E[(T_n(x) - f(x))^2] = 0,$$

则称 $T_n$ 为$f$ **的（在 $x$ 处）均方相合估计**。简记为 $T_n(x) \Rightarrow f(x)$。

类似可定义对固定 $x$，$T_n(x)$ 依概率收敛于 $f(x)$ 及以概率 1 收敛。这些相合称为**逐点相合性**。与此相关的概念，则是一致相合性。

**定义 2.3.** 如对任给的 $\varepsilon > 0$，

$$\lim_{n\to\infty} P\left(\sup_x |T_n(x) - f(x)| \geq \varepsilon\right) = 0,$$

则称 $T_n$ 是$f$ **的一致相合估计**，并简记为

$$\sup_x |T_n(x) - f(x)| \xrightarrow{P} 0, \text{ 当 } n \to \infty\text{。}$$

**定义 2.4.** 如果

$$P\left(\lim_{n\to\infty} \sup_x |T_n(x) - f(x)|\right) = 1$$

则称 $T_n$ 为$f$ **的一致强相合估计**，并简记为

$$\sup_x |T_n(x) - f(x)| \to 0, \ a.s. \text{ 当 } n \to \infty\text{。}$$

通常证明一致相合性或一致强相合性是分两步进行的。其一，是证明

$$\lim_{n\to\infty} \sup_x |E[T_n(x)] - f(x)| = 0;$$

其二，是断定当 $n \to \infty$ 时，

$$\sup_n |T_n(x) - E[T_n(x)]| \to 0$$

这里的收敛或者是依概率或者是 a.s.。这第一部分无随机性可言，完全由 $f$ 及估计量的光滑性所确定，因而较容易。主要困难在第二部分。在某些情况下，可将 $\sup_n |T_n(x) - E[T_n(x)]|$ 表成经验过程的适当泛函，然后使用经验过程的有关性质得以证明。

## 2.2 核估计的大样本性质

下面的引理可以说是核估计的一个基本引理，最先由 Parzen 给出。

**引理 2.1.** 设 $K(\cdot)$ 及 $g(\cdot)$ 均为 $\mathbb{R}^1$ 上的 *Borel* 可测函数，满足下述条件：

*(1)* $K$ 有界，

*(2)* $\int |K(u)|du < \infty$，

*(3)* $\lim_{|u|\to\infty} uK(u) = 0$ 或 $g$ 有界，

*(4)* $\int |g(u)|du < \infty$。

常数序列 $\{h_n\}$ 满足 $\lim_{n\to\infty} h_n = 0$。令

$$g_n(x) = \frac{1}{h_n}\int K\left(\frac{u}{h_n}\right) g(x-u)du,$$

则

$$\lim_{n\to\infty} g_n(x) = g(x)\int K(u)du,\ \forall x \in c(g) \tag{2.1}$$

其中 $c(g)$ 是 $g$ 的连续点集。又若 $g$ 有界且一致连续，则（*2.1*）关于 $x$ 一致成立。

对于核估计的逐点相合性，我们有如下定理：

**定理 2.1.** 设核 $K$ 是 $\mathbb{R}^1$ 上的概率密度，且满足引理*2.1*之条件（*1*）、（*2*）。若 $\lim_{n\to\infty} h_n = 0$，则有

$$\lim_{n\to\infty} E[f_n(x)] = f(x),\ x \in c(f)$$

又若 $f$ 一致连续，则上式关于 $x$ 一致成立。

**定理 2.2.** 设核 $K$ 满足定理*2.1*的条件，且

$$\lim_{n\to\infty} h_n = 0,\ \lim_{n\to\infty} nh_n = \infty$$

则

$$f_n(x) \Rightarrow f(x),\ x \in c(f).$$

**定理 2.3.** 设 $f$ 一致连续，$K$ 为概率密度，且

*(1)* $K(u)$ 有可积的特征函数 $k(u)$，

*(2)* $\lim_{n\to\infty} h_n = 0$，$\lim_{n\to\infty} nh_n^2 = \infty$。

则

$$\sup_x |f_n(x) - f(x)| \xrightarrow{P} 0,\ \text{当}\ n \to \infty。$$

为了讨论强相合性，需要一个关于经验分布的概率不等式。

**引理 2.2.** 设 $X_1$，$\cdots$，$X_n$ 是来自连续分布函数 $F(x)$ 的独立同分布样本。$F_n(x)$ 是其经验分布函数，则存在绝对常数 $c > 0$ 及 $0 < \alpha \le 2$，使得对任给 $\varepsilon > 0$，

$$P(\sup_x |F_n(x) - F(x)| \ge \varepsilon n^{-1/2}) \le c exp(-a\varepsilon^2)$$

**定理 2.4.** 设 $K$ 是有界变差的概率密度，$f$ 一致连续，若

$$\lim_{n\to\infty} h_n = 0,\ \lim_{n\to\infty} nh_n^2/\log(n) = \infty,$$

则

$$\lim_{n\to\infty}\sup_x |f_n(x) - f(x)| = 0,\ a.s.$$

## 2.3 N.N. 估计的大样本性质

我们用 $\hat{f}_n(x)$ 表示由（1.4）所定义的 N.N. 估计。

**定理 2.5.** 设 $k = k_n$ 满足

$$k_n \to \infty,\ k_n/n \to 0,\ \text{当}\ n \to \infty,$$

则当 $n \to \infty$ 时，

$$\hat{f}_n(x) \xrightarrow{P} f(x),\ x \in c(f)$$

**定理 2.6.** 设 $k_n$ 满足

$$k_n \to \infty,\ k_n/n \to 0,\ k_n/\log(n) \to \infty\ \text{当}\ n \to \infty,$$

则当 $n \to \infty$ 时，

$$\hat{f}_n(x) \to f(x),\ a.s.\ x \in c(f)$$

**定理 2.7.** 设 $k \triangleq k_n$ 满足

$$k_n \to \infty,\ k_n/n \to 0,\ k_n/\sqrt{(}n\log(n)) \to \infty$$

若 $f$ 一致连续，则有

$$\lim_{n\to\infty}\sup_x |\hat{f}_n(x) - f(x)| = 0,\ a.s.$$

## 2.4 高维情形

此节均设 $X_1$，$\cdots$，$X_n$ 是来自未知 $d$ 维密度 $f(x)$ 的独立同分布样本。

1. 光滑参数的设计。我们可将一元核估计的定义推广为：

$$f_n(x) = \frac{1}{nh_n^d}\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) \tag{2.2}$$

其中 $K(\cdot)$ 是 $\mathbb{R}^d$ 中的密度函数，$h_n > 0$ 是窗宽。在这一定义中，对数据的每一分量用同一刻度因子 $h_n$ 加以光滑。当数据点在某一方向上的变异比其它方向要显著地大时，这一定义就不合适了。这时不如使用一个常向量或常数矩阵作为光滑参数来得好。另一种方法是先将数据作刻度变换，以降低数据点的各向变异，再对经过处理的数据使用定义（2.2)。例如 Fukunaga 曾提出如下变换方法：记 $S$ 为 $X_1$，$\cdots$，$X_n$ 的样本协差阵，作变换 $\widetilde{X}_i = S^{-1/2}X_i$，$i = 1, 2, \cdots, n$。然后使用一个径向对称核 $K$ 加以光滑，最后变回原数据。如此得到的估计可以表成

$$f_n(x) = \frac{(\det S)^{-1/2}}{hh_n^d}\sum_{i=1}^n k\left[h_n^{-2}(x - X_i)'S^{-1}(x - X_i)\right],$$

其中，$k(x'x) = K(x)$。这样做的好处是：变换后的数据 $\widetilde{X}_1$，$\cdots$，$\widetilde{X}_n$，其样本协差阵是单位阵，因而消除了数据的各向变异的差别。其缺点是上述公式的计算量较大。

2. 尾部估计。一般说来，在低维情形 $f$ 的尾部估计失当影响不大。这是因为落在尾部区域中的数据很少。故而绝大部分样本可看成来自截尾分布。然而当维数 $d$ 增大时，情况就有明显的差别。与低维情形相反，低密度区域是高维分布的非常重要部分。因而在高维情形，对 $f$ 的尾部估计需要十分小心。

3. 对给定估计精度，维数对最低限度的样本容量的影响。随着维数的增大，最低样本容量的增大是非常之快。

# 3 非参数回归

## 3.1 引言

设在一实际问题中，我们感兴趣的自变量 $X$ 与因 $Y$（均可为多维）有某种相关关系。在经典回归分析中，常假定 $(X', Y')'$ 有多元正态分布 $N(\mu, \Sigma)$，其中

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix},\ \Sigma = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda 22 \end{pmatrix}$$

在此假定下，当给定 $X = x$ 时，$Y$ 的条件分布仍为多元正态。$Y$ 的条件期望为

$$m(x) \triangleq E[Y|X = x] = E[Y|x] = \mu_2 + \Lambda_{21}\Lambda_{11}^{-1}(x - \mu_1)$$

函数 $m(x)$ 常称为（$Y$ 对 $X$ 的）**回归函数**。

理论和实践都证明了在上述正态回归模型下，最小二乘估计有种种优良性质。然后在很多实际问题中，正态性不一定成立，必须寻找一种普遍适用的估计条件分布的方法。我们可以将问题一般地表述为：设有因变量 $Y$（为一维）与自变量 $X$（$d$ 维）配对，$(X, Y)$ 的分布未知，只假定 $E[|Y|] < \infty$。今有来自 $(X, Y)$ 的随机样本 $(X_i, Y_i)$，$i = 1, 2, \cdots, n$，要求基于该样本估计回归函数 $m(x) = E[Y|x]$，即构造估计 $m_n(x) = m_n(x; X_1, Y_1, \cdots, X_n, Y_n)$，使得对每一个 $x \in \mathbb{R}^d$，用 $m_n(x)$ 作 $m(x)$ 的估计。

## 3.2 Stone 权函数法

**定义 3.1.** 以 $n$ 记样本大小，则 $n$ 个形如 $W_{ni}(x) = W_{ni}(x; X_1, \cdots, X_n)$（$i = 1, 2, \cdots, n$）的函数，称为**权函数**（权函数可以指这 $n$ 个的整体，或者其中任一个）。又若

$$W_{ni}(x) \geq 0,\ 1 \leq i \leq n;\ \sum_{i=1}^{n} W_{ni}(x) = 1,$$

则称 $\{W_{ni}\}$ 为**概率权函数**。对给定的权函数 $\{W_{ni}\}$，定义回归函数 $m(x)$ 的估计为

$$m_n(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i,$$

并称 $m_n(x)$ 为 $m(x)$ 的一个**权函数估计**。

直观上看，权 $W_{ni}(x)$ 表示在估计 $m(x)$ 时，样本 $(X_i, Y_i)$ 所起的作用的“大小”。一个常见的例子是对给定的 $x \in \mathbb{R}^d$，将 $X_1$，$\cdots$，$X_n$ 中恰好等于 $x$ 的那些样本挑选出来，并给予它们一样的权重：

$$W_{ni}(x) = \frac{1_{\{X_i=x\}}}{\#\text{ of } X_j\text{'s such that } X_j = x}, m_n(x) = \sum_{i=1}^{n} \frac{Y_i 1_{\{X_i=x\}}}{\#\text{ of } X_j\text{'s such that } X_j = x}$$

1. 近邻权方法。其直观想法是，对给定的样本 $X_1$，$\cdots$，$X_n$ 及 $x \in \mathbb{R}^d$，虽然可能没有一个 $X_i$ 恰好等于 $x$，但可将“等于 $x$”的要求降低为“与 $x$ 接近”。依每个 $X_i$ 对给定 $x$ 的距离重新排序，与 $x$ 距离越近的其重要程度越大。近邻权方法在理论上已经证明了有诸多优良的大样本性质，但是其计算较复杂。

2. 核函数法。选定 $\mathbb{R}^d$ 上的核函数 $K(\cdot)$ 及窗宽 $h_n$，然后定义

$$W_{ni}(x) = \frac{K\left(\frac{x-X_i}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{h_n}\right)}, \ i = 1, \cdots, n$$

称此 $\{W_{ni}\}$ 为**核权函数**。相应的权函数估计为

$$m_n(x) = \sum_{i=1}^{n} \frac{K\left(\frac{x-X_i}{h_n}\right) Y_i}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{h_n}\right)} \tag{3.1}$$

估计（3.1）的合理性可作如下解释：设 $(X, Y)$ 有联合密度 $f(x, y)$ 则有

$$m(x) = E[Y|x] = \int y f(x, y) dy / \int f(x, y) dy \triangleq \int y f(x, y) dy / f_X(x)$$

边缘密度 $f_X(x)$ 的核估计为 $\frac{1}{nh_n^d} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)$，而 $\int y f(x, y) dy$ 可用 $\frac{1}{nh_n^d} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right) Y_i$ 去估计。分别以这两个估计作分母和分子即得（3.1）。核权函数的优点是有一个明确的关于 $x$ 的统一表达式，从而便于计算。但由于（3.1）的分母是随机变量，给理论处理带来一定的困难。注意近邻权方法可用看作是核函数法的一个特例。

## 3.3 权函数估计的相合性

定义 $d_n(x) \triangleq |m_n(x) - m(x)|$。权函数估计的**逐点相合性**及**逐点强相合性**分别指

$$d_n(x) \xrightarrow{P} 0, \ n \to \infty$$

及

$$d_n(x) \longrightarrow 0 \ a.s., \ n \to \infty$$

另外一种途径则是考虑整体精度，即 $d_n(X)$ 的平均。这一想法导致了“矩相合”的概念。

**定义 3.2.** 设 $\{W_{ni}\}$ 是给定的权函数，若对任意的 $r \geq 1$ 及任一满足

$$E[|Y|^r] < \infty$$

的 $Y$，都有

$$\lim_{n \to \infty} E[(d_n(X))^r] = 0,$$

则称 $\{W_{ni}\}$ 为**矩相合**的。

**定理 3.1.** 设 $\{W_{ni}\}$ 为给定的概率权函数，则其矩相合的充要条件是

(1) 存在有限常数 $C$，使得对任一非负函数 $f$ 都有

$$E[\sum_{i=1}^{n} W_{ni}(X)f(X_i)] \leq CE[f(X)],$$

(2) 对任给 $\varepsilon > 0$，当 $n \to \infty$ 时有

$$\sum_{i=1}^{n} W_{ni}(X)I_{\{|X_i-X|>\varepsilon\}} \xrightarrow{P} 0,$$

(3) $\max_{1\leq i\leq n} W_{ni}(X) \xrightarrow{P} 0$。

条件（2）可理解为对于与 $x$ 距离超过某种限度的那些样本 $X_i$，其权的总和很小，因而在估计 $m(x)$ 时，主要依据最接近 $x$ 的那些样本。条件（3）意味着，作为单独的一个样本点 $X_i$，不论它与 $x$ 的距离多么接近，所起的作用总是很小的。

**定理 3.2.** 近邻权在一定条件下是矩相合的。

**定理 3.3.** 设 $\{W_{ni}\}$ 为以 $K$ 为核的核权函数，而 $K$ 为 $\mathbb{R}^d$ 上具有紧支撑的有界概率密度。若

$$h_n \to 0,\ nh_n^d \to \infty,\ \text{当}\ n \to \infty,$$

则 $\{W_{ni}\}$ 为矩相合的。

## 3.4 应用

1. 条件二阶矩估计。设有 $q$ 维变量 $Y = (Y^{(1)}, \cdots, Y^{(q)})$ 与 $X$ 配对，而 $(X_i, Y_i)$（$i = 1, \cdots, n$）是来自 $(X, Y)$ 的随机样本，且已给定了权函数 $\{W_{ni}\}$，要求估计给定 $X = x$ 时，$Y$ 的条件二阶矩。

**定理 3.4.** 设 $\{W_{ni}\}$ 为矩相合的概率权函数，且 $E[\|Y\|^2] < \infty$，则有

$$\lim_{n\to\infty} E\left[|Cov_n(Y^{(i)}, Y^{(j)}|X) - Cov(Y^{(i)}, Y^{(j)}|X)|\right] = 0$$

又若以概率 1 有 $Var(Y^{(i)}|X) > 0$，$Var(Y^{(j)}|X) > 0$，则对任给 $r > 0$，有

$$\lim_{n\to\infty} E\left[|\rho_n(Y^{(i)}, Y^{(j)}|X) - \rho(Y^{(i)}, Y^{(j)}|X)|^r\right] = 0$$

2. 条件分位数法。设有随机变量 $Y$ 与 $X$ 配对（$X$ 仍设为 $d$ 维向量），以 $F(\cdot|x)$ 记给定 $X = x$ 时，$Y$ 的条件分布函数，记其 $p$ 分位数为 $\xi(p|x)$。$(X_i, Y_i)$，$i = 1, \cdots, n$ 是来自 $(X, Y)$ 的随机样本，$\{W_{ni}\}$ 为基于 $X_1$，$\cdots$，$X_n$ 的一个给定的权函数。则 $F(\cdot|x)$ 的权函数估计为

$$F_n(y|x) = \sum_{i=1}^{n} W_{ni}(x)I_{\{Y_i\leq y\}}$$

以 $F_n(\cdot|x)$ 的任一 $p$ 分位数 $\xi_n(p|x)$ 作为 $\xi(p|x)$ 的估计。显然 $\xi_n(p|x)$ 由权函数 $\{W_{ni}\}$ 所确定，但并不唯一。然而出乎意料的是：在某种限制下，当 $\xi(p|x)$ 唯一时，不论如何选择 $\xi_n(p|x)$，其大样本极限是唯一的。

**定理 3.5.** 设 $\{W_{ni}\}$ 为矩相合的概率权函数，若 $F(\cdot|X)$ 以概率 *1* 有唯一的 $p$ 分位数 $\xi(p|X)$，则不论如何选择 $\xi_n(p|x)$，都有

$$\xi_n(p|X) \xrightarrow{P} \xi(p|X)$$

又若对某个 $r>0$ 有 $E[|Y|^r]<\infty$，则

$$\lim_{n\to\infty} E\left[|\xi_n(p|x)-\xi(p|x)|^r\right]=0$$

3. 预测。令 $L(y,a)$ 表示当 $Y$ 实际取值为 $y$，而预测为 $a$ 时的损失。通常 $L$ 取平方损失或绝对值损失两种形式。设 $\delta(x)$ 为某个预测规则，即当 $X$ 取值 $x$ 时用 $\delta(x)$ 预测 $Y$ 之值。同时称 $E[L(Y,\delta(X))]$ 为$\delta$ **的（在 $L$ 下的）风险**。若有规则 $\delta^*(\cdot)$，使得

$$E[L(Y,\delta^*(X)]=\inf_{\delta} E[L(Y,\delta(X))] \triangleq R^*,$$

则称 $\delta^*$ 为**（在损失 $L$ 下的）Bayes 预测**，而称 $R^*$ 为 **Bayes 预测风险**。

不难求得：当 $L(Y,a)=(Y-a)^2$ 时，$\delta^*(x)=E[Y|x]$。当 $L(Y,a)=|Y-a|$ 时，$\delta^*(x)=\xi\left(\frac{1}{2}|x\right)$。因而当 $(X,Y)$ 的分布已知时，可以求得 Bayes 预测 $\delta^*$。但在实际问题中 $(X,Y)$ 的分布是未知的，只有来自 $(X,Y)$ 的历史样本 $(X_i,Y_i)$，$i=1,2,\cdots,n$。

**定义 3.3.** 设 $L$ 为给定的损失，如一个估计 $\delta_n^*$ 使得

$$\lim_{n\to\infty} E[L(Y,\delta_n*(X))]=R^*,$$

则称$\delta_n^*$ **有（在 $L$ 下的）Bayes 相合性**。又如 $\delta_n^*$ 是由权函数 $\{W_{ni}\}$ 所确定，则称$\{W_{ni}\}$ **具有（在 $L$ 下的）Bayes 相合性**。

**定理 3.6.** 设 $\{W_{ni}\}$ 为矩相合的概率权函数，且 $E[|Y|]<\infty$，若 $\xi\left(\frac{1}{2}|X\right)$ 以概率 *1* 唯一，则在绝对值损失下 $\{W_{ni}\}$ 有 *Bayes* 相合性。

**定理 3.7.** 设 $\{W_{ni}\}$ 为矩相合的概率权函数，且 $E[|Y|^2]<\infty$，则在平方损失下 $\{W_{ni}\}$ 有 *Bayes* 相合性。

**扩展阅读**：沃塞曼 [2] 对现代非参数统计作了一个包罗万象的概述。Härdle [3] 第 10 章对多维非参数回归作了简介。

# 参考文献

[1] 陈希孺，柴根象：《非参数统计教程》。上海：华东师范大学出版社，1993.3。

[2] L. 沃塞曼著，吴喜之译：《现代非参数统计》。北京：科学出版社，2008。

[3] Wolfgang Härdle. *Applied Nonparametric Regression*. Cambridge University Press, 1992.