基础数理统计

第二十章 非参数曲线估计

本章节，我们主要研究两个问题

- 给定一组样本 $X_1, \cdots, X_n$, 如何估计其密度函数 $f(x)$?
- 给定一组样本 $(X_1, Y_1), \cdots, (X_n, Y_n)$, 如何估计回归函数 $r(x) = E(Y|X = x)$?

共同特点是都需要去估计一个函数曲线。

1 20.1 偏差-方差平衡
2 20.2 直方图
3 20.3 核密度估计
4 20.4 非参数回归

风险

令 g 表示一个未知函数，$\widehat{g}_n$ 表示 g 的一个估计。

- 积分平方的误差 (ISE):

$$L(g, \widehat{g}_n) = \int (g(u) - \widehat{g}_n(u))^2 du.$$

- 风险或期望积分平方的误差 (MISE):

$$R(g, \widehat{g}_n) = \mathsf{E}\left(L(g, \widehat{g}_n)\right).$$

引理 1

风险可以写为

$$R(g, \widehat{g}_n) = \int b^2(x)dx + \int v(x)dx,$$

其中

$$b(x) = E(\widehat{g}_n(x)) - g(x)$$

为 $\widehat{g}_n(x)$ 在某固定点 x 处的偏差，而

$$v(x) = V(\widehat{g}_n(x)) = E\left(\widehat{g}_n(x) - E(\widehat{g}_n(x))\right)^2$$

为 $\widehat{g}_n(x)$ 在某固定点 x 处的方差。

- 当数据被过光滑化时，偏差项变大而方差项变小。
- 当数据被欠光滑化时，结论相反。

称为偏差–方差平衡，最小化风险相当于寻找偏差与方差的平衡。

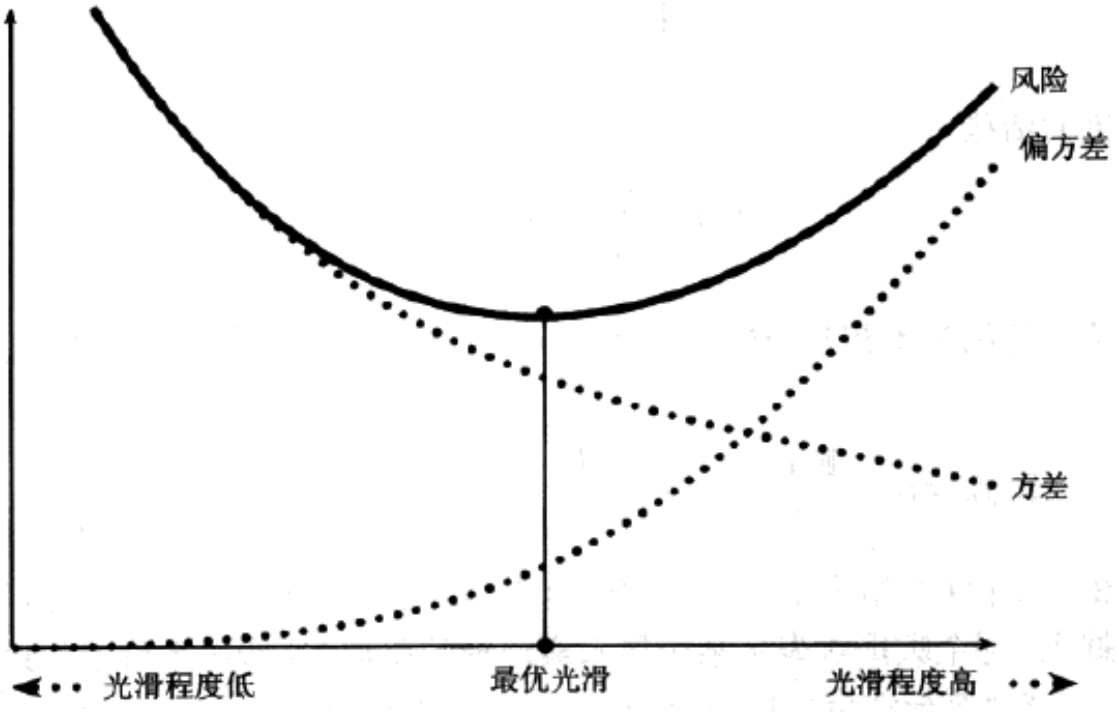

图 20.2 偏差 – 方差平衡

1 20.1 偏差-方差平衡

2 20.2 直方图

3 20.3 核密度估计

4 20.4 非参数回归

直方图

令 $X_1, X_2, \ldots, X_n$ 为区间 $[0,1]$ 上的简单随机样本且具有密度函数 f。令 m 为一整数且定义窗格

$$B_1 = [0, \frac{1}{m}),\ B_2 = [\frac{1}{m}, \frac{2}{m}),\ \ldots, B_m = [\frac{m-1}{m}, 1].$$

定义窗宽为 $h = 1/m$，ν_j 表示 B_j 中的观测数，令

$$\widehat{p}_j = \frac{\nu_j}{n},\ p_j = \int_{B_j} f(u)du.$$

直方图估计可以定义为

$$\widehat{f}_n(x) = \sum_{j=1}^{m} \frac{\widehat{p}_j}{h} I(x \in B_j).$$

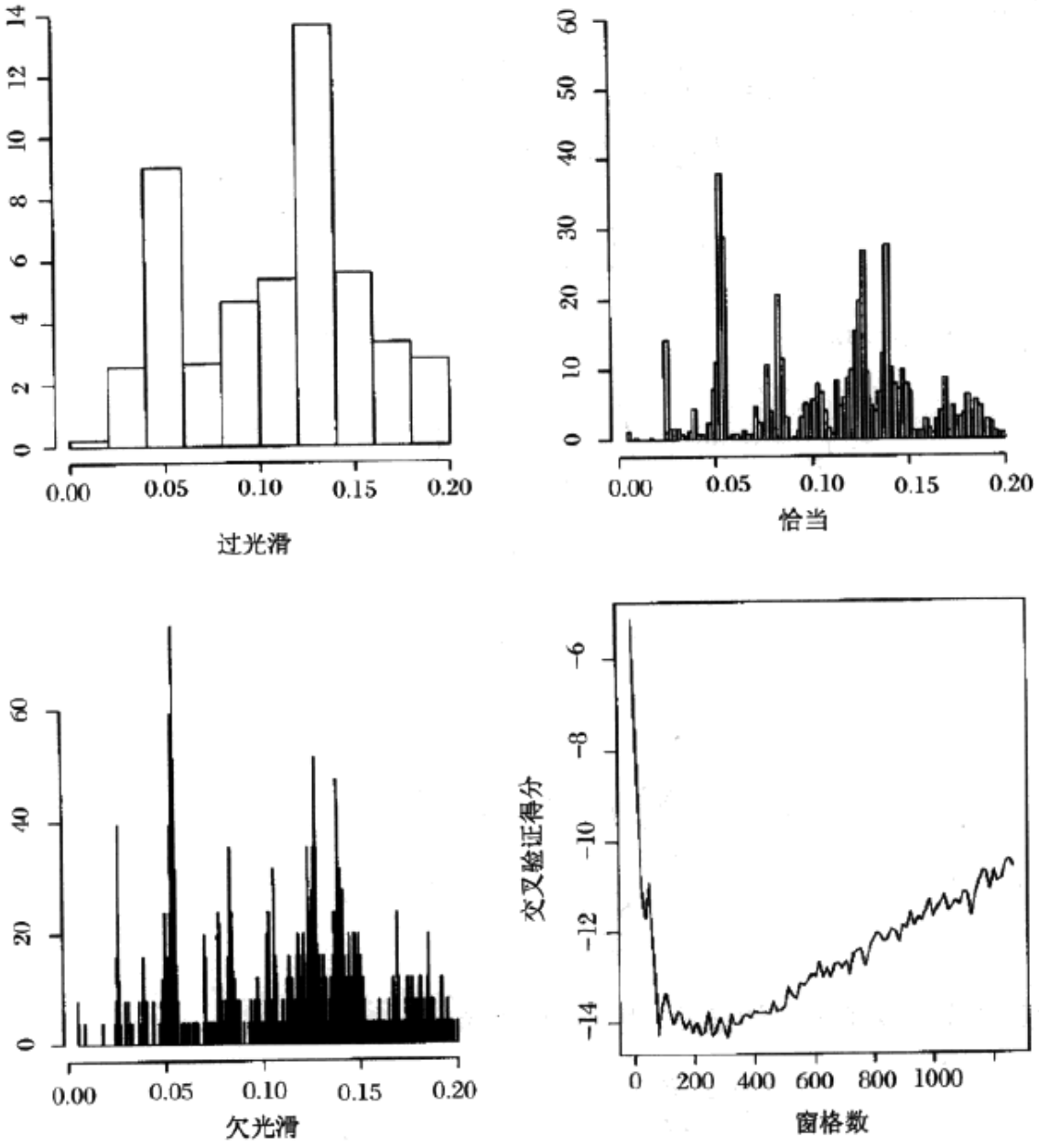

图 20.3 天文学数据的三个直方图

定理 1

考虑固定的 x 和固定的 m，且令 B_j 为含有 x 的窗格，则

$$E(\widehat{f}_n(x)) = \frac{p_j}{h}, \quad V(\widehat{f}_n(x)) = \frac{p_j(1-p_j)}{nh^2}.$$

注意到

$$\begin{aligned} p_j &= \int_{B_j} f(u)du \approx \int_{B_j} (f(x) + (u-x)f'(x))du \\ &= f(x)h + hf'(x)\left(h(j-\frac{1}{2}) - x\right). \\ b(x) &= \frac{p_j}{h} - f(x) = f'(x)\left(h(j-\frac{1}{2}) - x\right). \end{aligned}$$

若 $\tilde{x}_j$ 是窗格的中心，则

$$\begin{aligned}
\int_{B_j} b^2(x)dx &\approx \int_{B_j} \left(f'(x)\left(h(j-\frac{1}{2}) - x \right) \right)^2 dx \\
&\approx \left(f'(\tilde{x}_j)\right)^2 \int_{B_j} \left(h(j-\frac{1}{2}) - x \right)^2 dx \\
&= \left(f'(\tilde{x}_j)\right)^2 \frac{h^3}{12},
\end{aligned}$$

$$\begin{aligned}
\int_0^1 b^2(x)dx &= \sum_{j=1}^m \int_{B_j} b^2(x)dx \approx \sum_{j=1}^m \left(f'(\tilde{x}_j)\right)^2 \frac{h^3}{12} \\
&= \frac{h^2}{12} \sum_{j=1}^m h\left(f'(\tilde{x}_j)\right)^2 \\
&\approx \frac{h^2}{12} \int_0^1 \left(f'(x)\right)^2 dx.
\end{aligned}$$

方差

由

$$v(x) \approx \frac{p_j}{nh^2} = \frac{f(x)h + hf'(x)(h(j-1/2)-x)}{nh^2} \approx \frac{f(x)}{nh}.$$

得到 $\int_0^1 v(x)dx \approx 1/(nh)$。

定理 2

假设 $\int (f'(u))^2 du < \infty$，则

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

极小化右边的值 $h^\star$ 为

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

在这个窗宽选择下，

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

其中

$$C = (3/4)^{2/3} \left(\int (f'(u))^2 du \right)^{1/3}.$$

窗宽选择

另一个选择窗宽的方法是估计风险函数然后关于 h 极小化。

$$\begin{aligned} L(h) &= \int (\widehat{f}_n - f(x))^2 dx \\ &= \int (\widehat{f}_n(x))^2 dx - 2\int (\widehat{f}_n(x))(f(x))dx + \int (f(x))^2 dx \\ &= J(h) + \int (f(x))^2 dx. \end{aligned}$$

定义 1 (风险的交叉验证估计)

为

$$\widehat{J}(h) = \int (\widehat{f}_n(x))^2 dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{(-i)}(X_i).$$

其中 $\widehat{f}_{(-i)}$ 是去掉第 i 个观测后得到的直方图估计，称 $\widehat{J}(h)$ 为交叉验证得分或估计风险。

定理 3

交叉验证估计几乎是无偏的，

$$E(\widehat{J}(h)) \approx E(J(h)).$$

定理 4

$$\widehat{J}(h) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)h}\sum_{j=1}^{m} \widehat{p}_j^2.$$

证明.

根据定义，我们有

$$\int (\widehat{f}_n(x))^2 dx = \sum_{j=1}^{m} \frac{\widehat{p}_j^2}{h}, \ \widehat{f}_{(-i)}(X_i) = \sum_{j=1}^{m} \frac{n\widehat{p}_j - 1}{h(n-1)} I(X_i \in B_j).$$

代入即得

$$\begin{aligned}
\widehat{J}(h) &= \frac{1}{h}\sum_{j=1}^{m}[\widehat{p}_j^2 - \frac{2(n\widehat{p}_j - 1)}{n(n-1)}\sum_{i=1}^{n} I(X_i \in B_j)] \\
&= \frac{1}{h}\sum_{j=1}^{m}[\widehat{p}_j^2 - \frac{2(n\widehat{p}_j - 1)\widehat{p}_j}{(n-1)}] \\
&= \frac{2}{(n-1)h} - \frac{n+1}{(n-1)h}\sum_{j=1}^{m}\widehat{p}_j^2.
\end{aligned}$$

□

定义

$$\bar{f}_n(x) = \mathsf{E}(\widehat{f}_n(x)) = \frac{p_j}{h},\ x \in B_j,$$

其中 $p_j = \int_{B_j} f(u)du$.

定义 2

函数对 $(\ell_n(x), u_n(x))$ 是一个 $1-\alpha$ 置信带 *(*或称置信包络*)* 若

$$P(\ell_n(x) \le \bar{f}_n(x) \le u_n(x),\ \forall x) \ge 1-\alpha.$$

定理 5

令 $m = m(n)$ 为直方图 $\widehat{f}_n(x)$ 中的窗格数，假设当 $n \to \infty$ 时且 $m(n) \to \infty$ 时有 $m(n) \log n/n \to 0$. 定义

$$\ell_n(x) = \left(\max\left\{ \sqrt{\widehat{f}_n(x)} - c, 0 \right\} \right)^2, \ u_n(x) = \left(\sqrt{\widehat{f}_n(x)} + c \right)^2,$$

其中 $c = \frac{z_{\alpha/(2m)}}{2} \sqrt{\frac{m}{n}}$，则 $(\ell_n(x), u_n(x))$ 是一个近似的 $1 - \alpha$ 置信带。

证明.

$$
\begin{aligned}
& P(\ell_n(x) \le \bar{f}_n(x) \le u_n(x), \forall x) \\
= \quad & 1 - P(\max_x |\sqrt{\widehat{f}_n(x)} - \sqrt{\bar{f}_n(x)}| > c) \\
\ge \quad & 1 - \sum_{j=1}^{m} P(|2\sqrt{n}(\sqrt{\widehat{p}_j} - \sqrt{p_j})| > z_{\alpha/(2m)}).
\end{aligned}
$$

中心极限定理有 $\sqrt{n}(\widehat{p}_j - p_j) \approx N(0, p_j(1-p_j))$，结合 Delta 方法，有 $2\sqrt{n}(\sqrt{\widehat{p}_j} - \sqrt{p_j}) \approx N(0, 1-p_j)$，代入上式即得。 □

20.3 核密度估计

1 20.1 偏差-方差平衡
2 20.2 直方图
3 20.3 核密度估计
4 20.4 非参数回归

定义 3 (核)

核定义为一个光滑函数 K 使得

$$K(x) \geq 0, \int K(x)dx = 1, \int xK(x)dx = 0,$$

并且 $\sigma_K^2 = \int x^2 K(x)dx > 0.$

例如:

(1) Epanechnikov 核

$$K(x) = \begin{cases} \frac{3}{4}\left(\frac{1-x^2}{5}\right)/\sqrt{5} & |x| < \sqrt{5} \\ 0 & \text{其他} \end{cases}$$

(2) 高斯 (正态) 核 $K(x) = (2\pi)^{-1/2}\exp(-x^2/2).$

其他常用核函数

- Uniform: $K(\mu) = \frac{1}{2} I(|\mu| \leq 1)$;
- Triangular: $K(\mu) = (1 - |\mu|) I(|\mu| \leq 1)$;
- Quartic (biweight): $K(\mu) = \frac{15}{16}(1 - \mu^2)^2 I(|\mu| \leq 1)$;
- Triweight: $K(\mu) = \frac{35}{32}(1 - \mu^2)^2 I(|\mu| \leq 1)$;
- Tricube: $K(\mu) = \frac{70}{81}(1 - |\mu|^3)^3 I(|\mu| \leq 1)$;
- Cosine: $K(\mu) = \frac{\pi}{4} cos(\frac{\pi}{2}\mu) I(|\mu| \leq 1)$;
- Logistic: $K(\mu) = \frac{1}{e^{\mu} + 2 + e^{-\mu}}$.

核密度估计

定义 4 (核密度估计)

给定一组 $i.i.d$ 样本 $X_1, \ldots, X_n$，给定一个核 K 与一个正数 h，称为带宽。核密度估计为

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\frac{x - X_i}{h}).$$

说明：K 的选择不是关键的，带宽 h 的选择是非常重要的。

窗宽

对于给定的核函数，窗宽 h 反映了样本点的影响范围:

$$\int \frac{1}{h}K(\frac{x}{h})dx = \int K(y)dy = 1, \text{ 变量代换 } y = x/h.$$

例如

- 对于均匀核 $K(\mu) = \frac{1}{2}I(|\mu| \leq 1)$ 对应于 $U[-1,1]$, $\frac{1}{h}K(\frac{x}{h})$ 对应的是 $U[-h,h]$.
- Gaussian 核是标准正态分布 $N(0,1)$, $\frac{1}{h}K(\frac{x}{h})$ 对应于 $N(0,h^2)$.

窗宽选择

在核密度估计中，窗宽 h 的选择对最终结果影响非常大:

- 窗宽 h 选取的太小，核密度估计曲线波动非常厉害，方差很大 (极端的 $h \to 0$ 对应的就是经验分布函数导数).

- 窗宽 h 选取的太大，曲线过于光滑，偏差非常大。

所以核密度估计需要找到合适的窗宽，方差和偏差之间的平衡 (The Bias-Variance Tradeoff)。理论分析可以给出窗宽的数量级如 $h \sim n^{-1/5}$. 经验上，

- 对于正态核函数：$h_{opt} \approx 1.06\hat{\sigma} n^{-1/5}$.
- 对于 Epanechnikov 核：$h_{opt} \approx 2.34\hat{\sigma} n^{-1/5}$.

定理 6

在 f 和 K 的弱假设下，

$$R(f,\widehat{f}_n) \approx \frac{1}{4}\sigma_K^4 h^4 \int (f''(x))^2 dx + \frac{\int K^2(x)dx}{nh},$$

其中 $\sigma_K^2 = \int x^2 K(x)dx > 0$. 最优的带宽为

$$h^\star = \frac{c_1^{-2/5} c_2^{1/5} c_3^{-1/5}}{n^{1/5}},$$

其中

$$c_1 = \int x^2 K(x)dx,\ c_2 = \int (K(x))^2 dx,\ c_3 = \int (f''(x))^2 dx.$$

在这个带宽选择下，$R(f,\widehat{f}_n) \approx \frac{c_4}{n^{4/5}}$ 对于某个常数 $c_4 > 0$.

定理 7

对于任意 $h > 0$, $E(\widehat{J}(h)) = E(J(h))$ 且

$$\widehat{J}(h) \approx \frac{1}{hn^2} \sum_i \sum_j K^\star \left(\frac{X_i - X_j}{h} \right) + \frac{2}{nh} K(0),$$

其中

$$K^\star(x) = K^{(2)}(x) - 2K(x), \quad K^{(2)}(x) = \int K(x-y)K(y)dy.$$

可以选择能够最小化 $\widehat{J}(h)$ 的带宽，其合理性见下面的定理：

定理 8 (Stone 定理)

假设 f 有界，令 $\widehat{f}_h$ 表示带宽为 h 的核估计且令 h_n 表示由交叉验证得到的带宽，则

$$\frac{\int (f(x) - \widehat{f}_{h_n}(x))^2 dx}{\inf_h \int (f(x) - \widehat{f}_h(x))^2 dx} \xrightarrow{P} 1.$$

最优核函数-Epanechnikov 核

通过对核密度估计的理论分析，Epanechnikov 核为最优核，或者更一般的，最优核函数为：

$$K_{opt}(t) = \frac{3}{4\alpha}(1 - t^2/\alpha^2)I(|t| \leq \alpha);$$

核估计可以推广到 d 维的情况，令 $\boldsymbol{h} = (h_1, \ldots, h_d)^\top$ 为一个带宽向量且定义

$$\widehat{f}_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\boldsymbol{h}}(\boldsymbol{x} - \boldsymbol{X}_i),$$

其中

$$K_{\boldsymbol{h}}(\boldsymbol{x} - \boldsymbol{X}_i) = \frac{1}{nh_1 h_2 \ldots h_d}\left\{\prod_{j=1}^{d} K\left(\frac{x_j - X_{ij}}{h_j}\right)\right\}.$$

作业：1

20.4 非参数回归

1 20.1 偏差-方差平衡
2 20.2 直方图
3 20.3 核密度估计
4 20.4 非参数回归

回归函数

在**均方损失**下，回归函数是最优的预测函数：

$$\arg\min_g E(Y - g(X))^2 = E(Y|X = x) := r(x).$$

在监督学习中，统计模型的目标都是来估计出回归函数 $r(x)$.

Regressongram

类似于直方图 Histogram, 在回归函数估计问题中也可以把解释变量 $X_1, \cdots, X_n$ 所在的全部空间分成 m 组 $B_1, \cdots, B_m$, 用每一组数据的平均值作为对应组的条件期望。

Regressongram

对于任意的 x, 一定存在一个组使得 $x \in B_j$, 定义回归函数为

$$\hat{r}(x) = \frac{\sum_{i:\ X_i \in B_j} Y_i}{\#\{i :\ X_i \in B_j\}}, \tag{1}$$

即把每一组内的 Y_i 的平均作为这一组内回归函数的取值。

Local Average

在 Regressongram 中，组数和如何分组对结果影响很大，每组的数据都是同等对待，对于处于两个组边缘的样本也不太公平。

换一种方式来估计回归函数, 对于任意的 x, 我们选择一个邻域 $B_x = \{y : \|y - x\| \leq h\}$，定义回归函数

$$\hat{r}(x) = \frac{\sum_{i:\ X_i \in B_x} Y_i}{\#\{i :\ X_i \in B_x\}}. \tag{2}$$

如果选择一般的距离函数，理论上我们可以处理多维的情形，这里的 h 可以视为窗宽，其大小决定我们用 x 周围多少的邻居来估计回归函数 $r(x)$.

Kernel Regression-核回归

Local Average 思想赋予了领域中每个点同样的权重，进一步的让每个样本点贡献值和距离相关即是核回归估计：

定义 5 (Nadaraya–Watson 核估计)

对于样本 $(X_1, Y_1), \cdots, (X_n, Y_n)$, 定义回归函数为：

$$\widehat{r}(x) = \sum_{i=1}^{n} \omega_i(x) Y_i,$$

这里的 $K(x), h$ 是核函数和窗宽且

$$\omega_i(x) = \frac{K(\frac{x - X_i}{h})}{\sum_{i=1}^{n} K(\frac{x - X_i}{h})}$$

- 和密度函数的核估计一样，相对于窗宽这里的核函数选取没有那么的重要。
- 理论上，我们也可以在模型假设下去计算理论损失，然后优化损失函数得到最优窗宽表达式。
- 对于回归函数来说，更为简单实用的方式是通过重抽样等交叉验证的方式选择窗宽和核函数。

定理 9

假设 $V(\epsilon_i) = \sigma^2$. *Nadaraya–Watson* 核估计的风险为

$$R(\widehat{r}_n, r) \approx \frac{h^4}{4}\left(\int x^2 K^2(x)dx\right)^4 \int \left(r''(x) + 2r'(x)\frac{f'(x)}{f(x)}\right)^2 dx + \int \frac{\sigma^2 \int K^2(x)dx}{nhf(x)}dx.$$

最优宽带以 $n^{-1/5}$ 的速率递减且在该选择下其风险以 $n^{-4/5}$ 的速率递减。

定理 10

$\widehat{J}$ 可以写为

$$\widehat{J}(h) = \sum_{i=1}^{n} (Y_i - \widehat{r}(X_i))^2 \frac{1}{\left(1 - K(0)/\sum_{j=1}^{n} K\left(\frac{x_i - x_j}{h}\right)\right)^2}.$$

令 $\bar{r}_n(x) = \mathsf{E}(\widehat{r}_n(x))$ 且

$$\omega_i(x) = \frac{K\left(\frac{x - x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right)}, \quad \widehat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2.$$

定理 11 (核回归的置信带)

$\bar{r}_n(x)$ 的一个近似 $1-\alpha$ 置信带为

$$\ell_n(x) = \widehat{r}_n(x) - q\widehat{se}(x),\ u_n(x) = \widehat{r}_n(x) + q\widehat{se}(x),$$

其中 $m = (b-a)/\omega$,

$$\widehat{se}(x) = \widehat{\sigma}\sqrt{\sum_{i=1}^{n}\omega_i^2(x)},\ q = \Phi^{-1}\left(\frac{1+(1-\alpha)^{1/m}}{2}\right).$$

这里 ω 为核的宽度。

作业：8