第四章 核估计

王 成 *
http://math.sjtu.edu.cn/faculty/chengwang/

上海交通大学数学系

1 介绍

给定一组iid的样本$X_1,\ldots,X_n$，如何刻画其分布函数$F(x)$?

Definition 1.1 (Empirical distribution function) 我们可以定义经验分布函数(Empirical distribution function)，

$$F_n(x)=\frac{1}{n}\sum_{k=1}^{n}I(X_k\le x). \tag{1}$$

作为$F(x)$的无偏估计，EDF有如下的性质：

1. For any fixed x,

$$E(\hat{F}_n(x))=F(x),\ \text{and}\ Var(\hat{F}_n(x))=\frac{F(x)(1-F(x))}{n}. \tag{2}$$

   Further, if $0<F(x)<1$, we have

$$\sqrt{n}(\hat{F}_n(x)-F(x))\rightsquigarrow N(0,F(x)(1-F(x)). \tag{3}$$

2. Gilvenko-Cantelli Theorem:

$$\sup_x|F_n(x)-F(x)|\xrightarrow{a.s.}0. \tag{4}$$

3. Dvoretzky - Kiefer - Wolfowitz (DKW) inequality: For any $\epsilon>0$,

$$P(\sup_x|\hat{F}_n(x)-F(x)|>\epsilon)\le 2e^{-2n\epsilon^2}. \tag{5}$$

*关于讲义中的任何错误或者建议，请联系chengwang@sjtu.edu.cn

我们知道对于很多随机变量，我们更多时候是以其密度函数来判断或者刻画分布的。那么，如何通过iid样本来估计或者刻画这组数据的密度函数$f(x)$?

□ 我们可以仿照之前的V统计量，现在要估计$F'(x)$,那么我们用经验分布函数的导数$F_n'(x)$来估计，我们有：

$$F_n'(x) = \frac{1}{n}\sum_{k=1}^{n}\delta_{X_k}(x)$$

这里的δ是 Dirac delta 函数：

$$\delta_x(t) = \begin{cases} \infty & \text{if } t = x; \\ 0 & \text{if } t \neq x. \end{cases}$$

并且满足$\int \delta_x(t)dt = 1$.

□ 直方图.如果确定直方图的结点$t_0 < t_1 < \cdots < t_{m-1} < t_m$，定义一个新的函数

$$\gamma_x(t) = \sum_{k=1}^{m} I(t_{k-1} < x, t \leq t_k)$$

那么计数直方图可以表示为：

$$H_n(x) = \sum_{i=1}^{n}\gamma_{X_i}(x). \tag{6}$$

如果我们想展示密度函数，可以定义

$$\gamma_x(t) = \sum_{k=1}^{m} I(t_{k-1} < x, t \leq t_k)/(t_k - t_{k-1})$$

然后直方图表示为：

$$H_n(x) = \frac{1}{n}\sum_{i=1}^{n}\gamma_{X_i}(x). \tag{7}$$

实际上，这里的

$$\gamma_x(t) = \begin{cases} \frac{1}{t_k - t_{k-1}} & t_{k-1} < t \leq t_k; \\ 0 & \text{others}. \end{cases}$$

这里对于某个k，$t_{k-1} < x \leq t_k$，且满足$\int \gamma_x(t)dt = 1$. 对于一般直方图，我们会选取一个起点h_0及固定的宽度δ,而实际问题中这两个参数的选取对于最后的结果影响也是很大的。

2 核密度估计 Kernel Density Estimation

2.1 Motivations

仔细检查直方图，我们会发现直方图对于所有发生在$(t_{k-1}, t_k]$中的样本都同等的对待。简单起见我们考虑$[0,1]$区间，如果有两个样本取值$0^+, 1^-$,那么我们应该认为前者表达的信息是分布在0周围有密度，而后者应该是在1周围有密度，而不能把两个完全同等的看待。

因此，对于直方图，给定宽度δ，我们认为样本X_i应该反应区间$[X_i - \delta/2, X_i + \delta/2]$的信息，例如我们定义

$$K_\Delta(x) = I(-\delta/2 \le x \le \delta/2)/\delta.$$

那么对应的估计为：

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_\delta(X_i - x). \tag{8}$$

再进一步我们把宽度δ看成参数，直接定义

$$K(x) = I(|x| \le 1)/2.$$

然后估计为：

$$f_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{X_i - x}{h}) \tag{9}$$

当然这里我们取的是同等权重，如果考虑一般的权，那么我们也就得到了一般的核密度估计形式：

Definition 2.1 (Kernel density Estimate)

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\frac{X_t - x}{h}) = \int K_h(\mu - x)d\hat{F}(\mu), \tag{10}$$

这里$K_h(\cdot) = K(\cdot/h)/h$称为核函数，$h$ 是 **窗宽**(Bandwidth) 参数。

2.2 核函数 Kernel function

What is the condition that K should satisfy?

1. $K(x) \ge 0$ for all x;
2. $\int K(x)dx = 1$.

Therefore, $K(\cdot)$ is a density function and any density function can act as a kernel function in theory.

Actually, a kernel function is usually a nonnegative symmetric, unimodal probability density function.

Commonly used kernel functions:

☐ Uniform

$$K(\mu) = \frac{1}{2} I(|\mu| \leq 1); \tag{11}$$

☐ Triangular

$$K(\mu) = (1 - |\mu|) I(|\mu| \leq 1); \tag{12}$$

☐ Epanechnikov:

$$K(\mu) = \frac{3}{4}(1 - \mu^2) I(|\mu| \leq 1); \tag{13}$$

☐ Quartic (biweight):

$$K(\mu) = \frac{15}{16}(1 - \mu^2)^2 I(|\mu| \leq 1); \tag{14}$$

☐ Triweight:

$$K(\mu) = \frac{35}{32}(1 - \mu^2)^2 I(|\mu| \leq 1); \tag{15}$$

☐ Tricube:

$$K(\mu) = \frac{70}{81}(1 - |\mu|^3)^3 I(|\mu| \leq 1); \tag{16}$$

☐ Gaussian:

$$K(\mu) = \frac{1}{\sqrt{2\pi}} exp(-\mu^2/2); \tag{17}$$

☐ Cosine:

$$K(\mu) = \frac{\pi}{4} cos(\frac{\pi}{2}\mu) I(|\mu| \leq 1); \tag{18}$$

☐ Logistic:

$$K(\mu) = \frac{1}{e^{\mu} + 2 + e^{-\mu}} \tag{19}$$

Q: how to calculate the constant part such as $\frac{35}{32}$?

2.3 Bandwidth

Q: Why do we use the thresholding 1 not 2 or other parameters?

Remark 2.1 If $K(\mu)$ is a kernel function, for any $h > 0$,

$$\frac{1}{h}K(\frac{\mu}{h}) \quad \text{or} \quad hK(h\mu) \tag{20}$$

can both serve as the kernel functions.

Notes on the bandwidth

- A large bandwidth h-bias, over-smooth.
- A small bandwidth h-large variance.
- Optimal bandwidth should achieve a trade-off between bias and variance.

Conclusion: "It is well-known both empirically and theoretically that the choice of kernel functions is not very important to the kernel density estimator. As long as they are symmetric and unimodal, the resulting kernel density estimator performs nearly the same when the bandwidth h is optimally chosen. "

3 窗宽选择Bandwidth Selection

为了估计样本的密度函数$f(x)$, 我们构造了核估计

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\frac{X_i - x}{h}) = \int K_h(\mu - x)d\hat{F}(\mu), \tag{21}$$

作为一个估计, 我们可以考虑估计相关的性质。

3.1 期望 Expectation

我们直接计算核估计的期望:

$$\begin{aligned} E\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n}E\frac{1}{h}K(\frac{X_i - x}{h}) &= E\frac{1}{h}K(\frac{X_1 - x}{h}) \\ &= \int \frac{1}{h}K(\frac{t-x}{h})f(t)dt \\ &= \int K(t)f(x+ht)dt. \end{aligned}$$

所以, 一般来说, 核估计是有偏的, 因为

$$\int K(\mu)f(x+h\mu)d\mu \neq f(x).$$

我们可以看下不同核下的期望:

1. Uniform $K(\mu) = \frac{1}{2}I(|\mu| \leq 1)$，我们有

$$E\hat{f}_h(x) = \int K(t)f(x+ht)dt = \frac{1}{2}\int I(|t| \leq 1)f(x+ht)dt = \frac{1}{2}\int_{-1}^{1} f(x+ht)dt.$$

即从$[x-h, x+h]$上的平均密度。窗宽度h决定选取的区间宽度，直观上窗宽越小误差越小，但是方差越大。反之选取的区间越大，估计的偏差越大，但是波动性越小。

2. Gaussian: $K(\mu) = \frac{1}{\sqrt{2\pi}}exp(-\mu^2/2)$,

$$E\hat{f}_h(x) = \int K(t)f(x+ht)dt = \frac{1}{2\pi}\int exp(-t^2/2)f(x+ht)dt = Ef(x+hY).$$

这里的$Y \sim N(0,1)$. 即整个实数区间上的加权平均，窗宽也有类似现象。

3.2 方差 Variance

注意到核估计是独立同分布随机变量的样本平均，所以

$$\begin{aligned} nVar(\hat{f}_h(x)) &= Var(\frac{1}{h}K(\frac{X_1 - x}{h})) \\ &= \int \frac{1}{h^2}K^2(\frac{t-x}{h})f(t)dt - (\int K(t)f(x+ht)dt)^2 \\ &= \frac{1}{h}\int K^2(t)f(x+ht)dt - (\int K(t)f(x+ht)dt)^2 \end{aligned}$$

所以估计的方差为：

$$Var(\hat{f}_h(x)) = \frac{1}{nh}\int K^2(t)f(x+ht)dt - \frac{1}{n}(\int K(t)f(x+ht)dt)^2$$

类似的我们可以看下Uniform核的情况：当$K(\mu) = \frac{1}{2}I(|\mu| \leq 1)$，我们有

$$Var(\hat{f}_h(x)) = \frac{1}{2nh}\int_{-1}^{1} f(x+ht)dt - \frac{1}{n}(\int_{-1}^{1} f(x+ht)dt)^2$$

3.3 均方损失 Mean Square Error

因为核估计是有偏差的，所以我们考虑均方损失来衡量估计的好坏：

$$\begin{aligned} MSE(x) &= E(\hat{f}_h(x) - f(x))^2 \\ &= Var(\hat{f}_h(x)) + (f(x) - E(\hat{f}_h(x)))^2 \\ &= \frac{1}{nh}\int K^2(t)f(x+ht)dt - \frac{1}{n}(\int K(t)f(x+ht)dt)^2 \\ &\quad + (f(x) - \int K(t)f(x+ht)dt)^2 \end{aligned}$$

3.4 窗宽选取

当样本$n \to \infty$时候，注意到核估计是iid的平均，所以经典CLT可以给出其渐近分布，这里我们具体来看均方损失的极限状况。

一般窗宽$h \to 0$，把$f(x+ht)$在x处做Taylor展开：

$$f(x+ht) = f(x) + f'(x)ht + \frac{1}{2}f''(x)h^2t^2 + O(h^3t^3), \tag{22}$$

我们有：

$$\begin{aligned}\frac{1}{h}\int K^2(t)f(x+ht)dt &= \frac{1}{h}\int K^2(t)(f(x)+f'(x)ht+\frac{1}{2}f''(x)h^2t^2+O(h^3t^3))dt \\ &= \frac{f(x)}{h}\int K^2(t)dt + \frac{h}{2}f''(x)\int K^2(t)t^2dt + \cdot\end{aligned}$$

类似的

$$\begin{aligned}\int K(t)f(x+ht)dt &= \int K(t)(f(x)+f'(x)ht+\frac{1}{2}f''(x)h^2t^2+O(h^3t^3))dt \\ &\approx f(x) + \frac{h^2}{2}f''(x)\int K(t)t^2dt\end{aligned}$$

所以

$$MSE(x) \approx \frac{f(x)}{nh}\int K^2(t)dt + \frac{1}{4}(f''(x))^2h^4(\int t^2K(t)dt)^2,$$

这里$h \to 0$，所以只有主项。如果我们为了让在x处核估计的MSE最小，优化上面的MSE即可。但是一般的这里我们考虑的是一个密度函数的估计，所以不能仅仅从某一点x来判断h的选取。为此，我们引入 Mean Integrated Square Error (MISE)：

$$\begin{aligned}\mathsf{MISE} &= E\int(\hat{f}_h(x)-f(x))^2dx \\ &\approx \int[\frac{f(x)}{nh}\int K^2(t)dt + \frac{1}{4}(f''(x))^2h^4(\int t^2K(t)dt)^2]dx \\ &\approx \underline{\frac{1}{nh}\int K^2(t)dt} + \frac{h^4}{4}\int(f''(x))^2dx(\int t^2K(t)dt)^2.\end{aligned}$$

最小化MISE，我们得到最优窗宽(实际上是渐近最优窗宽)

$$h_{opt} = \mathop{\mathsf{argmin}}_h \mathsf{MISE} = n^{-1/5}(\int(f''(x))^2dx)^{-1/5}(\int K^2(t)dt)^{1/5}(\int t^2K(t)dt)^{-2/5}. \tag{23}$$

记

$$\|g\|_2^2 = \int g^2(x)dx, \quad \mu_2(K) = \int t^2K(t)dt, \tag{24}$$

最优窗宽为：

$$h_{opt} = (\frac{\|K\|_2^2}{n\mu_2^2(K)\|f''\|_2^2})^{1/5} \tag{25}$$

此时最小的MISE为：

$$\begin{aligned} MISE_{opt} &= \frac{5}{4}\frac{\|K\|_2^2}{n}(\frac{\|K\|_2^2}{n\mu_2^2(K)\|f''\|_2^2})^{-1/5} \\ &= \frac{5}{4}(\frac{\|K\|_2^2}{n})^{4/5}(\mu_2^2(K)\|f''\|_2^2)^{1/5}. \end{aligned} \tag{26}$$

4 最优核 Optimal Kernel function

这一节我们比较核函数对于核估计的影响。从最优MISE的结果，我们发现：

$$\begin{aligned} MISE_{opt} &= \frac{5}{4}\frac{\|K\|_2^2}{T}(\frac{\|K\|_2^2}{T\mu_2^2(K)\|f''\|_2^2})^{-1/5} \\ &= \frac{5}{4}(\frac{\|K\|_2^2}{T})^{4/5}(\mu_2^2(K)\|f''\|_2^2)^{1/5}. \end{aligned} \tag{27}$$

所以最好的核函数应该是：

$$K_{opt} \in \text{argmin}(\|K\|_2^2)^2(\mu_2(K)) = (\int K^2(\mu)d\mu)^2(\int \mu^2 K(\mu)d\mu) := \beta(K). \tag{28}$$

In addition, K_{opt} should satisfy

$$\int K(\mu)d\mu = 1, \quad \int \mu K(\mu)d\mu = 0. \tag{29}$$

Note the fact

$$\beta(K(x)) = \beta(\frac{1}{h}K(\frac{x}{h})). \tag{30}$$

Therefore, if $K_{opt}(\cdot)$ is an optimal kernel, $\frac{1}{h}K(\frac{\cdot}{h})$ is also an optimal kernel. 所以我们只需要最小化

$$\int K^2(t)dt,$$

$K(t)$需要满足条件

$$\int K(t)dt = 1, \quad \int tK(t)dt = 0, \quad \int t^2K(t)dt = 1. \tag{31}$$

记Epanechnikov核

$$K_0(t) = \frac{3}{4}(1-t^2)I(|t| \le 1); \tag{32}$$

我们考虑函数$\delta(t) = K(t) - K_0(t)$,我们有

$$\int t^m \delta(t) dt = 0, \ \ m = 0, 1, 2.$$

所以

$$\int (1 - t^2)\delta(t) dt = 0.$$

现在考虑:

$$\begin{aligned} \frac{4}{3} \int \delta(t) K_0(t) dt &= \int_{|t| \le 1} \delta(t)(1 - t^2) dt \\ &= -\int_{|t| > 1} \delta(t)(1 - t^2) dt \\ &= \int K(t)(t^2 - 1) dt \ge 0. \end{aligned}$$

所以

$$\int K(t)^2 dt = \int K_0^2(t) dt + 2 \int \delta(t) K_0(t) dt + \int \delta^2(t) dt \ge \int K_0^2(t) dt.$$

因此Epanechnikov核为最优核，或者更一般的，最优核函数为:

$$K_{opt}(t) = \frac{3}{4\alpha}(1 - t^2/\alpha^2) I(|t| \le \alpha); \tag{33}$$

5 窗宽的经验选取

对于正态核函数:

$$h_{opt} \approx 1.06 \hat{\sigma} n^{-1/5}.$$

对于Epanechnikov核:

$$h_{opt} \approx 2.34 \hat{\sigma} n^{-1/5}.$$

这里的样本协方差是从 $\int (f''(x))^2 dx$项得到的，对于正态分布，两次导数得到的是方差。