非参数估计理论

赵丰

2018 年 6 月 27 日

1　符号检测

考虑 n 次观测，$X_1, X_2, \ldots, X_n$，独立同分布但概率密度函数 $p(x)$ 未知。概率密度函数的中位数 t 满足 $P(X < t) = \frac{1}{2}$。针对假设检验问题

$$H_0 : \texttt{median} = 0$$
$$H_1 : \texttt{median} > 0$$

设在 H_1 假设下 $p = P(X \geq 0)$。针对每个观测 X_i 我们取它的符号 $u(X_i)$，在 H_0 假设下 $u(X) \sim Bern(\frac{1}{2})$，在 H_1 假设下 $u(X) \sim Bern(p)$ 因此 $u(X_1), u(X_2), \ldots, u(X_n)$ 为二项分布，根据似然比的方法构造最佳统计量 T。

$$\lambda(X) = \prod_{i=1}^{N} \frac{f(x_i|H_1)}{f(x_i|H_0)}$$

当 $x_i > 0$ 时，

$$\frac{f(x_i|H_1)}{f(x_i|H_0)} = 2p^{u(x_i)}(1-p)^{u(x_i)}$$

从而

$$\lambda(X) = 2^n p^{\sum_{i=1}^n u(x_i)}(1-p)^{n-\sum_{i=1}^n u(x_i)}$$

由 $\lambda(X) \gtrless \lambda$ 以及 $p > \frac{1}{2}$ 的性质可以得到最佳统计量 T 为

$$T = \sum_{i=1}^{n} u(X_i) \underset{H_0}{\overset{H_1}{\gtrless}} c$$

T 服从二项分布，$\begin{cases} T|H_0 & \sim B(n, \frac{1}{2}) \\ T|H_1 & \sim B(n, p) \end{cases}$，虚警概率为

$$P_F = \sum_{k=c+1}^{n} \binom{n}{k} \left(\frac{1}{2}\right)^n$$

令 $P_F \leq \alpha$ 求得最大的 c，由此求出检测概率为：

$$P_D = \sum_{k=c+1}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

如果做关于中位数是否是 M_0 的假设检验，只需将符号检测应用到 $Z_i = X_i - M$ 上即可。

2 Wilcoxon 检测

对于观测 $X_1, X_2, \ldots, X_n (i.i.d.) \sim X$，

$$\begin{aligned} &H_0 : \texttt{median}(X) = M_0 \\ &H_1 : \texttt{median}(X) > M_0 \end{aligned} \tag{2.1}$$

利用观测的绝对值大小信息 $|Z_i|$，将 $|Z_i|$ 从小到大排序，记 $r(|Z_i|)$ 为序号（$r(|Z_i|) = 1$ 表示 $|Z_i|$ 是最小的。）有 $\sum_{i=1}^{n} r(|Z_i|) = n(n+1)/2$ 记统计量 $T^+ = \sum_{i=1}^{n} u(Z_i) r(|Z_i|)$ 设 T_i 为 Bernoulli 随机变量，取值为 $\{0, 1\}$，$P(T_i = 1)$ 是 $\{|Z_i|\}$ 中第 i 小的数取值为 1 的概率。则

$$T^+ = \sum_{i=1}^{n} i T_i \tag{2.2}$$

可以证明在 H_0 的假设下，$T_1, T_2, \ldots, T_n$ 是相互独立的，且服从 $Bern(\frac{1}{2})$。因此在 H_0 的假设下

$$\mathbb{E}[T^+] = \sum_{i=1}^{n} i \mathbb{E}[T_i] = \frac{1}{2} \sum_{i=1}^{n} i = \frac{n(n+1)}{4} \tag{2.3}$$

$$\text{Var}[T^+] = \sum_{i=1}^{n} i^2 \text{Var}[T_i] = \frac{1}{4} \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{24} \tag{2.4}$$

于是关于 (2.1) 提出的假设检验问题可以用如下的判决：

$$T^+ - \frac{n(n+1)}{4} \underset{H_0}{\overset{H_1}{\gtrless}} k$$

其中 k 的取值是满足 $P_F \le \alpha$ 且尽量小。即 $\Pr(T^+ - \frac{n(n+1)}{4} \ge k|H_0) \le \alpha$

3 渐近相对效率 (ARE)

下面考虑弱信号检测问题，即随样本量 n 的增大，待估的参数 $\theta_n \to \theta_0$。

对下面的假设检验问题：

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta > \theta_0 \end{aligned} \tag{3.1}$$

若在 H_1 假设下求出 $\mathbb{E}_\theta[T]$，其对 θ 的导数在 $\theta = \theta_0$ 处取值, 而 $\mathrm{Var}[T]$ 在 H_0 的假设下求出，那么称 $e(T)$ 为统计量关于假设检验问题的效验 (efficacy). 即

$$e(T) = \frac{\left[\frac{\mathrm{d}\mathbb{E}_\theta[T]}{\mathrm{d}\theta}\Big|_{\theta_0}\right]^2}{\mathrm{Var}[T]} \tag{3.2}$$

若有两个检验统计量 T_n 和 T_n^* 其中 n 表示样本个数，在满足一定的正则化条件的情况下有：

$$\mathrm{ARE} = \lim_{n\to\infty} \frac{e(T_n)}{e(T_n^*)} \tag{3.3}$$

下面求解符号检测对线性检测的 ARE: 考虑 $f_0(x)$ 是关于 y 轴对称的概率密度函数，$F_0(x)$ 是 $f_0(x)$ 的累积分布函数。

$$\begin{aligned} H_0 &: X_i \sim F_0(x), \text{with } F_0(0) = \frac{1}{2} \\ H_1 &: X_i \sim F_0(x-\theta), \text{with } \theta > 0 \end{aligned} \tag{3.4}$$

H_0 表示中位数等于零，H_1 假设表示中位数大于零。$u(X_i)$ 为 Bernoulli 型随机变量，等于 1 的概率为 $1 - F_0(-\theta)$。由 X_i 是独立同分布，

$$\begin{aligned} \mathbb{E}_\theta[T_n] &= n(1 - F_0(-\theta)) \\ \mathrm{Var}[T^+] &= nF_0(-\theta)(1 - F_0(-\theta)) \end{aligned}$$

则 $e(T_n) = 4nf_0^2(0)$。对于线性检测的统计量 $T_n^* = \sum_{i=1}^n X_i$，在 H_1 假设下均值为 θ, 设在 H_0 假设下方差为 σ_x^2，则 $e(T_n^*) = \frac{n}{\sigma_x^2}$。从而得到符号检测对线性检测的 ARE:

$$\mathrm{ARE}_{\text{sign,linear}} = 4f_0^2(0)\sigma_x^2$$

当 X_i 是高斯分布时 $f_0 = \frac{1}{\sqrt{2\pi}\sigma_x} \Rightarrow \mathrm{ARE} = \frac{2}{\pi} \approx 0.64$，线性检测效果更好；当 X_i 是 Laplace 分布时，$f_0(x) = (\lambda/2)e^{-\lambda|x|} \Rightarrow f_0(0) = \lambda/2, \sigma_x^2 = 2/\lambda^2 \Rightarrow \mathrm{ARE} = 2$，符号检测效果更好。

当高斯分布的方差未知时，用 Student's t-test 检验统计量代替线性检测，即 $S_n^* = \frac{\sqrt{n}\bar{x}_n}{S_n}$ 服从 $n-1$ 个参数的 t 分布。其中 $\bar{x}_n$ 和 S_n 分别为样本均值和样本方差。

下面我们计算 Wilcoxon 检测的效验，首先可以将 T^+ 写成下面的形式（设 $M_0 = 0$ ）：

$$T^+ = \sum_{i=1}^{n}\sum_{j=1}^{i} u(X_i + X_j) \tag{3.5}$$

仍假设 $f_0(x)$ 的对称性，$\mathbb{E}[u(X_i)] = 1 - F_0(-\theta) = F_0(\theta)$ 且

$$\begin{aligned}\mathbb{E}[u(X_i + X_j)] &= P(X_i + X_j > 0)\\ &= 1 - \int_{-\infty}^{+\infty} f_0(\sigma - \theta)F_0(-\sigma - \theta)d\sigma\end{aligned}$$

于是有：

$$\left.\frac{\partial \mathbb{E}[T^+]}{\partial \theta}\right|_{\theta=\theta_0} = nf_0(\theta) + n(n-1)I \tag{3.6}$$

其中 $I = \int_{-\infty}^{\infty} f_0^2(\sigma)d\sigma$ 利用 (2.4)式的结果，得到 $e(T^+) = 12nI^2$。Wilcoxon 检验相对于 t 检验（等价于线性检测器）

$$\mathrm{ARE}_{\mathrm{Wilcoxon,t}} = 12\sigma_x^2 I^2$$

4 双输入系统

考虑将信号输入两个独立信道，得到的输出分别为 X_i 和 Y_i，现根据 $W = ((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))$ 检测是否有信号通过信道。

- H_0 假设为 X 和 Y 彼此独立，零均值，方差固定，$\mathbb{E}[X_1^2] = \sigma_1^2, \mathbb{E}[X_2^2] = \sigma_2^2$。
- H_1 假设为有信号 $S = (S_1, S_2, \ldots, S_n)$ 通过，信号零均值，方差为 σ_3^2，两信道的噪声分量与 H_0 中相同，与 S 独立。

假设是高斯加性噪声，且 $\sigma_1 = \sigma_2 = \sigma$，那么假设检验问题变成

- H_0 ：(X_i, Y_i) 是联合高斯的，$\rho = 0, \sigma_x = \sigma_y = \sigma$
- H_1 ：(X_i, Y_i) 是联合高斯的，$\rho > 0, \sigma_x^2 = \sigma_y^2 = \frac{\sigma^2}{1-\rho}$

通过似然比检验可以得到：

$$\sum_{i=1}^{n}(x_i + y_i)^2 \underset{H_0}{\overset{H_1}{\gtrless}} t \tag{4.1}$$

若通过非参数估计的形式，记

$$u(x_i, y_i) = \begin{cases} 1 & \text{当 } x_i, y_i \text{ 同号时} \\ 0 & \text{当 } x_i, y_i \text{ 异号时} \end{cases}$$

在 H_0 假设下 $u(x_i, y_i) \sim Bern(\frac{1}{2})$，在 H_1 假设下 $u(x_i, y_i) \sim Bern(p)$，其中 $p = \Pr(X > 0, Y > 0) + \Pr(X < 0, Y < 0)$ 且 $p > \frac{1}{2}$。符号检测的统计量为：

$$\sum_{i=1}^{n} u(x_i, y_i) \underset{H_0}{\overset{H_1}{\gtrless}} t$$

5 局部最优检测

在弱信号检测中，

$$H_0 : \theta = \theta_0 \tag{5.1}$$

$$H_1 : \theta > \theta_0, \theta \text{与 } \theta_0 \text{ 很接近} \tag{5.2}$$

关于上述假设检验问题，对于任意的检测 T_n，其虚警概率为 α_n，漏警概率为 β_n，n 表示样本数。若 T_n^* 与 T_n 有相同的虚警概率 α_n，其漏警概率 β_n^* 满足

$$\frac{\partial \beta_n^*(\theta)}{\partial \theta}\Big|_{\theta=\theta_0} \le \frac{\partial \beta_n(\theta)}{\partial \theta}\Big|_{\theta=\theta_0} \tag{5.3}$$

漏警概率 β_n 与检测概率 ϕ_n （或功效函数）有如下的关系：

$$\beta_n(\theta) = 1 - \phi_n(\theta) \tag{5.4}$$

因此 (5.3) 等价于

$$\frac{\partial \phi_n^*(\theta)}{\partial \theta}\Big|_{\theta=\theta_0} \ge \frac{\partial \phi_n(\theta)}{\partial \theta}\Big|_{\theta=\theta_0}$$

设 I 表示拒绝域，I^* 是 I 在观测空间的补集。则

$$\alpha_n = \int_I \cdots \int \prod_{i=1}^{n} f(x_i;\theta_0)dx \tag{5.5}$$

$$\phi_n(\theta) = \int_I \cdots \int \prod_{i=1}^{n} f(x_i;\theta)dx \tag{5.6}$$

在一定的正则化条件下，(5.6) 化为

$$\frac{\partial \phi_n(\theta)}{\partial \theta} = \int_I \cdots \int \frac{\partial}{\partial \theta} \prod_{i=1}^{n} f(x_i;\theta)dx \tag{5.7}$$

问题化为在 α_n 一定的情况下，极大化 (5.7) 式。考虑到

$$\frac{1}{\prod_{i=1}^{n} f(x_i;\theta_0)} \frac{\partial \phi_n(\theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ln f(x_i;\theta)}{\partial \theta}\Big|_{\theta=\theta_0}$$

由 NP 引理可以进一步得到我们的判别准则为

$$\sum_{i=1}^{n} \frac{\partial \ln f(x_i;\theta)}{\partial \theta}\Big|_{\theta=\theta_0} \underset{H_0}{\overset{H_1}{\gtrless}} c \tag{5.8}$$

c 的选取使得虚警概率为 α。

6 鲁棒检测

设 $(X_1, X_2, \ldots, X_n)$ 是独立同分布的随机变量，有相同的分布 p。问题是检验 $p = p_0$ 还是 $p = p_1$。

$$p_i = \{q | q = (1-\epsilon_i)f_i + \epsilon_i H_i, H_i \in H\}, i = 0, 1$$

f_i 被称作名义分布。记 $R(q_i', \phi(x))$ 为在检测 $\phi(x)$ 下的平均风险，其中 q_i' 是真实的分布。极大极小问题:

$$\min_{\phi} \sup_{q_1'} R(q_1', \phi) \text{ subject to } \sup_{q_0'} R(q_0', \phi) \leq \alpha \tag{6.1}$$

可以用最不利分布求解。设 q_i 为最不利分布，即对于任意的 $\phi(\cdot)$ 下式成立

$$R(q_i', \phi) \leq R(q_i, \phi)$$

于是 (6.1) 式可以化为

$$\min_{\phi} R(q_1, \phi) \text{ subject to } R(q_0, \phi) \leq \alpha$$

当取 R 为错误概率时上式即为 N-P 检验。

可以求出最不利分布具有如下形式：

$$q_0(x)=\begin{cases}(1-\epsilon_0)f_0(x) & \frac{f_1(x)}{f_0(x)}<c_0\\ \frac{1}{c_0}(1-\epsilon_0)f_1(x) & \frac{f_1(x)}{f_0(x)}\geq c_0\end{cases} q_1(x)=\begin{cases}(1-\epsilon_1)f_1(x) & \frac{f_1(x)}{f_0(x)}>c_1\\ c_1(1-\epsilon_1)f_0(x) & \frac{f_1(x)}{f_0(x)}\leq c_1\end{cases} \tag{6.2}$$

其中 $0\leq c_1<c_0<\infty$，根据 q_0,q_1 是概率密度函数可确定 c_0,c_1 的大小：

$$\begin{aligned}(1-\epsilon_0)\left\{\Pr(\frac{f_1}{f_0}<c_0|f_0)+\frac{1}{c_0}\Pr(\frac{f_1}{f_0}\geq c_0|f_1)\right\}=1\\ (1-\epsilon_1)\left\{\Pr(\frac{f_1}{f_0}>c_1|f_1)+c_1\Pr(\frac{f_1}{f_0}\leq c_1|f_0)\right\}=1\end{aligned} \tag{6.3}$$

根据 (6.2)式可以得到：

$$\frac{q_1(x)}{q_0(x)}=\begin{cases}bc_1 & \frac{f_1(x)}{f_0(x)}\leq c_1\\ b\frac{f_1(x)}{f_0(x)} & c_1<\frac{f_1(x)}{f_0(x)}<c_0\\ bc_0 & \frac{f_1(x)}{f_0(x)}\geq c_0\end{cases} \tag{6.4}$$

其中 $b=\frac{1-\epsilon_1}{1-\epsilon_0}$。

例 1. 考虑 *Huber* 模型的特殊情形，$\epsilon_0=\epsilon_1=\epsilon$，$f_0$ 是标准正态分布，$f_1(x)=f_0(x-\theta),\theta>0$。这对应着信道 $z=\theta+n$ 有常信号 θ 时观测的分布。这里 $b=1$，求该 *Huber* 模型的鲁棒检测。

解. 根据 (6.3) c_1,c_0 满足：

$$\begin{aligned}\Phi(\frac{\ln c_0}{\theta}+\frac{\theta}{2})+\frac{1}{c_0}\left[1-\Phi(\frac{\ln c_0}{\theta}-\frac{\theta}{2})\right]=\frac{1}{1-\epsilon}\\ 1-\Phi(\frac{\ln c_1}{\theta}-\frac{\theta}{2})+c_1\Phi(\frac{\ln c_1}{\theta}+\frac{\theta}{2})=\frac{1}{1-\epsilon}\end{aligned} \tag{6.5}$$

由 (6.4)式可得似然比为：

$$\frac{q_1(x)}{q_0(x)}=\begin{cases}c_1 & x\leq\frac{\ln c_1}{\theta}+\frac{\theta}{2}\\ \exp(\theta x-\frac{\theta^2}{2}) & \frac{\ln c_1}{\theta}+\frac{\theta}{2}<x<\frac{\ln c_0}{\theta}+\frac{\theta}{2}\\ c_0 & x\geq\frac{\ln c_0}{\theta}+\frac{\theta}{2}\end{cases}$$

为保证 p_1 和 p_0 两类的概率密度簇没有重叠，要求 $q_1\neq q_0$ 即 $c_1\neq 1$。在 (6.5) 中令 $c_1=1$ 得到 θ 的临界值 $\theta_\epsilon(\theta>\theta_\epsilon)$ 满足 $2\Phi(\frac{\theta_\epsilon}{2})=\frac{1}{1-\epsilon}$。

7 鲁棒估计

$X_1, X_2, \ldots, X_n$ 是独立同分布的随机变量，密度为 $f(x-\theta)$。假设 $f \in F$，其中 $F = \{f | f = (1-\epsilon)\phi + \epsilon h, h \in H\}$，其中 ϕ 是标准正态分布，H 是对称有界的概率密度函数族。对给定的损失函数 L，统计量 $\hat{\theta}(\boldsymbol{x})$ 是一致估计量，其中

$$\hat{\theta}(\boldsymbol{x}) = \arg\min_{\theta} \sum_{i=1}^{n} L(x_i - \theta) \tag{7.1}$$

且有 $\sqrt{n}(\hat{\theta}(\boldsymbol{x}) - \theta)$ 渐近分布是零均值，方差为 $V(l, f)$，其中

$$V(l, f) = \frac{\int l^2(x) f(x) dx}{\left[\int l'(x) f(x) dx\right]^2}, \text{ with } l = \frac{dL(x)}{dx} \tag{7.2}$$

对给定的 $f(x)$ 极小化 (7.2) 有:

$$\begin{aligned} V(l, f) &= \frac{\int l^2(x) f(x) dx}{\left[\int l(x) f'(x) dx\right]^2} \\ &\geq \frac{\int l^2(x) f(x) dx}{(\int (l\sqrt{f})^2 dx)(\int (\frac{f'}{\sqrt{f}})^2 dx)} \end{aligned}$$