关于Johns的经验Bayes估计的几点注记*

陈希孺

（中国科技大学）

设$\mathscr{X}$为一离散样本空间．$\{P_\theta,\ \theta\in I\}$为其上之一概率分布族，$I$为$R^1$上的一有限或无限的区间．$g(\theta)$为一可估函数，即存在其一无偏估计．我们可取$g(\theta)$的一定的无偏估计$h(x)$，例如，其UMVU估计．

设在I上给定了θ的先验分布G，且损失为平方函数$(a-\theta)^2$．假定

$$E[h^2(X)]=\int_I\Big[\sum_{x\in\mathscr{X}}P_\theta(x)h^2(x)\Big]dG(\theta)<\infty \qquad (1)$$

现在对X进行了r次独立随机观察，得样本$X^*=(X_1,\cdots,X_r)$．要由X^*作$g(\theta)$的Bayes估计（在平方损失及先验分布G之下）．不难证明：这个Bayes估计就是

$$\delta_G(x^*)=E(g(\theta)\,|\,X^*=x^*) \qquad (2)$$

其Bayes风险为

$$R(G)=E[g^2(\theta)]-[E(\delta_G(X^*))]^2 \qquad (3)$$

在条件(1)之下不难证明$R(G)<\infty$．

一般，先验分布G未知，这时在一定情况下可使用H. Robbins在[1]中所引进的经验Bayes估计(以下简记为EB估计)．对此处讨论的问题，Johns在[2]中引进了一个有些奇特（因而不大自然）的结构：在每个历史参数值θ_i处各进行$r+1$次独立随机观察，得样本

$$X_i=(X_{i1},\cdots,X_{ir},X_{i,r+1})=(X_i^*,X_{i,r+1}),\ i=1,\cdots,n.$$

定义

$$M_i=M_i(X_i^*,X^*)=\begin{cases}1, & 若\{X_i^*\}=\{X^*\}\\ 0, & 其他\end{cases} \qquad (4)$$

此处$\{a\}=\{b\}$表示由a的元构成之集（重复要计入）与由b的元构成之集相同．例如，$a=(1,2,1)$，$b=(2,1,1)$，则$\{a\}=\{b\}$．反之，若$a=(1,2,1)$而$b=(2,1,2)$，则$\{a\}\neq\{b\}$

记$M_{(n)}=\sum_{i=1}^n M_i$．Johns用

* 1981年2月16日收到．

$$\delta_n(x^*)=\delta_n(x_1,\cdots,x_n,x^*)=\begin{cases}0, & \text{若} M_{(n)}=0\\ \dfrac{1}{M_{(n)}}\sum_{i=1}^{n}M_i h(X_{i,r+1}), & \text{若 } M_{(n)}>0\end{cases}\quad (5)$$

作为$g(\theta)$的EB估计。并证明了：若先验分布族$\mathscr{F}$中任一先验分布G都满足条件(1)，则由(5)定出的δ_n是$g(\theta)$的渐近最优（简记为$(a.o.)$EB估计。

1970年，Maritz注意到，由于在给定θ时，$X_{i1},X_{i2},\cdots X_{i,r+1}$独立同分布，由(5)定义的EB估计$\delta_n$可作如下的修改：为此定义

$$M_i'=M_i(X_i^*,X^*)=1,\text{ 或 } 0$$

且$M_i'=1$，当且仅当存在k_i,使在$X_{i1},\cdots,X_{i,r+1}$中去掉X_{ik_i}后，剩下来的r个元所构成的X_i'满足条件$\{X_i'\}=\{X^*\}$，然后令$M_{(n)}'=\sum_{i=1}^{n}M_i'$，及

$$\delta_n'(x)=\delta_n'(x_1,\cdots,x_n,x^*)=\begin{cases}\dfrac{1}{r}\sum_{j=1}^{r}h(X_j), & \text{若 } M_{(n)}'=0\\ \dfrac{1}{M_{(n)}'}\sum_{i=1}^{n}M_i'h(X_{ik_i}), & \text{若 } M_{(n)}'>0\end{cases}$$

以之作为$g(\theta)$的EB估计。

Maritz在[3]中只提出了这个修改而未对其性质作任何讨论。本文的目的是对于这个EB估计的性质提出两点注记。

首先是关于这个EB估计的$a.o.$性。仔细检查Johns[2]的定理1的证明过程，不难发现，它对Maritz的修改δ_n'仍然适用。唯一不同之处在于Johns在证明中用了关系式

$$E[h(X_{i,r+1})\,|\,X_i^*=X^*]=\delta_G(X^*),\quad (6)$$

在此要修改为

$$E[h(X_{i,k_i})\,|\,\{X_i'\}=\{X^*\}]=\delta_G(X^*).\quad (7)$$

注意到X_i的（边缘）联合分布，即

$$P(X_i\in B)=\int_I P_\theta(X_i\in B)dG(\theta)\quad (8)$$

为对称分布，(7)不难由下述引理推出。我们把这引理写为更一般的形状以适合下文的需要。

引理1. 设$Y_1,\cdots,Y_m$的联合分布对称，而$1\leqslant q<m$. 从自然数$1,2,\cdots m$中随机地无放回地逐一取出q个，设结果依次为$j_1,\cdots,j_q$,则对任何$u,u=1,2,\cdots,m$,但$u\neq j_i,i=1,\cdots,q$,有

$$E[f(Y_u)\,|\,Y_{j_i}=y_i,i=1,\cdots,q]=E[f(Y_{q+1})\,|\,Y_i=y_i,i=1,\cdots,q]$$

此处$(y_1,\cdots,y_q)$任意(可有一个零测度例外集)，f为任意Borel函数，致$E[|f(Y_1)|]<\infty$.

此引理的证明在概念上很简单，而在说明上稍嫌烦琐，故在此从略了。

比较一般的情况是：每个历史参数值θ_i处抽取的样本个数不一样，设为$X_{i1}\cdots,X_{im_i}$，$m_i>r,i=1,\cdots,n$. 这时可以将Maritz的EB估计δ_n'作如下的自然推广。为此定义

$$M_i''=m_i-r,\text{ 或 } 0$$

且$M_i'' = m_i - r$，当且仅当存在 r 个足标

$$1 \leqslant i_1 < i_2 < \cdots < i_r \leqslant m_i$$

致$\{(X_{ii_1}, X_{ii_2}, \cdots, X_{ii_r})\} = \{X^*\}$. 将$1, 2, \cdots, m_i$中不为$i_1, \cdots, i_r$的那些足标记为$i_1' \cdots i'_{m_i - r}$，及

$$Z_i = \sum_{j=1}^{m_i - r} h(X_{ii_j'}),\quad i = 1, \cdots, n$$

（当 $M_i'' = 0$时，定义 $Z_i = 0$），又令$M_{(n)}'' = \sum_{i=1}^{n} M_i''$. 然后定义

$$\delta_n''(x) = \begin{cases} \dfrac{1}{r} \sum_{j=1}^{r} h(X_j), \text{若 } M_{(n)}'' = 0, \\ \dfrac{1}{M_{(n)}''} \sum_{i=1}^{n} Z_i,\ \text{若} M_{(n)}'' > 0. \end{cases}$$

用 Johns [2] 中的证法，结合引理 1，可以证明在满足条件 (1) 的先验分布族之下，δ_n'' 是$g(\theta)$的$a.o.$ EB 估计。细节从略。

一个自然会提出的问题是：用 Maritz 修改所得的 δ_n' 代替 Johns 的δ_n，能产生多大的好处。由于 δ_n 和δ_n' 都是 $a.o.$ 的 EB 估计。在这一点上二者无优劣可言。我们必须进一步去比较 δ_n 和δ_n' 的"全面 Bayes 风险" $R(G, \delta_n)$和$R(G, \delta_n')$与 Bayes 估计 δ_G 的 Bayes 风险$R(G)$的差距。按周知的公式(例如，参见[4])，有

$$R(G, \delta_n) = \sum_{x^* \in \mathscr{X}_r} p(x^*) E[\delta_n(X_1, \cdots X_n, x^*) - \delta_G(x^*)]^2 \tag{9}$$

这里$\mathscr{X}_r = \{x_1, \cdots x_r): x_i \in \mathscr{X}, i = 1, \cdots, r\}$，$p(\cdot)$是 X^*在先验分布G之下的边缘分布，即

$$p(x^*) = P(X^* = x^*) = \int_I P_\theta(X^* = x^*) dG(\theta)$$

又$X_1, \cdots, X_n$为独立同分布，每一个的分布由(8)决定。$R(G, \delta_n')$当然有类似的公式。

由(9)式可知，要 $R(G, \delta_n) - R(G)$小，需要对每个x^*，$E[\delta_n(X_1, \cdots X_n, x^*) - \delta_G(x^*)]^2$尽可能小，根据 δ_n 的定义以及(6)式，这个期望值（在给定 x^* 的条件下）大体上等于

$$\sum_{k=1}^{n} \frac{\sigma_{x^*}^2}{k} P(M_{(n)} = k \mid x^*) \triangleq V(x^*, n) \tag{10}$$

这里 σ_x^2 表示在给定$\{X_i^*\} = \{x^*\}$的条件下，$X_{i, r+1}$ 的条件分布的方差。前文之所谓"大体上"，意思是我们假定 n 足够大，因而概率 $P(M_{(n)} = 0 \mid x^*)$小到可以忽略不计。不难证明

引理 2. 设$0 < p < 1, q = 1 - p$，而

$$a_n = \sum_{k=1}^{n} \frac{1}{k} \binom{n}{k} p^k q^{n-k}$$

则

$$\lim_{n \to \infty} nPa_n = 1 \tag{11}$$

证 任给 $\varepsilon>0$ 充分小，根据Hoeffding在[5]中证明的一个结果，有

$$\sum_{\{k:\,|k-np|\geqslant n\varepsilon\}}\binom{n}{k}p^k q^{n-k}\leqslant 2\exp\left(-\frac{n\varepsilon^2}{1+2\varepsilon}\right)$$

由此可知

$$a_n\leqslant 2\exp\left(-\frac{n\varepsilon^2}{1+2\varepsilon}\right)+\frac{1}{n(p-\varepsilon)}$$

$$a_n\geqslant\left[1-2\exp\left(-\frac{n\varepsilon^2}{1+2\varepsilon)}\right)\right]\frac{1}{n(p+\varepsilon)}$$

因此

$$\frac{p}{p+\varepsilon}\leqslant\lim_{n\to\infty} inf\ npa_n\leqslant\lim_{n\to\infty} supnqa_n\leqslant\frac{p}{p-\varepsilon}$$

由 $\varepsilon>0$ 的任意性即得(11)。

由(10)及此引理可知

$$E[\delta_n(X_1,\cdots X_n,x^*)-\delta_G(x^*)]^2\approx V(x^*,n)\approx[n\cdot w(x^*)]^{-1}\tag{12}$$

此处

$$w(x^*)=\mathrm{P}(X_i^*=x^*)$$

类似地有

$$E[\delta_n'(X_1,\cdots X_n,x^*)-\delta_G(x^*)]^2\approx[n\cdot w'(x^*)]^{-1}\tag{13}$$

由(12)、(13)可知，当 n 充分大时，问题取决于 $w(x^*)$ 和 $w'(x^*)$ 的比值，我们先对样本空间 $\mathscr{X}$ 有限（设有 b 个点）的情况来考虑。

引理 3. 设 $X_1,\cdots X_{r+1}$ 独立同分布，X_1 的分布为

$$\mathrm{P}(X_1=a_i)=p_i>0,\quad i=1,\cdots,b,\ \ (a_i\text{ 两两不同})$$

设 $n_1,\cdots,n_b$ 都是自然数，$\sum_{i=1}^{b}n_i=r$。以 A 计事件

$$A=\{X_1,\cdots,X_{r+1}\text{ 中，等于 }a_i\text{ 的个数}\geqslant n_i,i=1,\cdots,b\}$$

则

$$P(A)=(r+1)\left(1-\sum_{i=1}^{b}\frac{n_i}{n_i+1}p_i\right)Q\tag{14}$$

其中

$$Q=\frac{r!}{n_1!\ \cdots n_r!}p_1^{n_1}\ \cdots p_r^{n_r}$$

证 定义以下一些事件：

$A_i=\{$在 $X_1,\cdots,X_{i-1},X_{i+1},\cdots,X_{r+1}$ 中，恰有 n_j 个为 $a_j,j=1,\cdots,b\}$，$i=1\cdots,r+1$，

$B_j=\{$在 $X_1,\cdots,X_{r+1}$ 中，有 n_j+1 个为 a_j,n_u 个为 $a_u,u=1,\cdots,b,u\neq j\}$，$j=1,\cdots,b$。

不难验证以下几点

1°. $P(A_i)=Q$，因而 $\sum_{i=1}^{r+1}P(A_i)=(r+1)Q$。

2°. $A\subset\bigcup_{i=1}^{r+1}A_i$，且 $Bj_1\cap Bj_2=\phi$，当 $j_1\neq j_2$。

3°. $A_i-\bigcup_{j=1}^{b}B_j$，$i=1,\cdots,r+1$，两两互斥。

4°。属于 B_j 的每个“基本事件”$(x_1,\cdots,x_{r+1})$，在 $A_1,\cdots,A_{r+1}$ 中的 n_j+1 个出现。事实上，设 $x_1,\cdots,x_{r+1}$ 中等于 a_j 的 n_j+1 个为 x_{u_k}，$k=1,\cdots,n_j+1$，则“基本事件”$(x_1,\cdots,x_{r+1})$ 在 A_{u_k} 中出现，$k=1,\cdots,n_j+1$。

综合以上各点，可得

$$P(A)=\sum_{i=1}^{r+1}P(A_i)-\sum_{j=1}^{b}n_jP(B_j)$$

$$=(r+1)Q-\sum_{j=1}^{b}n_j\frac{(r+1)!}{n_1!\ \cdots n_{j-1}!\ (n_j+1)!\ n_{j+1}!\ \cdots n_b!}p_1^{n_1}\cdots p_{j-1}^{n_{j-1}}\,p_j^{n_j}p_{j+1}^{n_{j+1}}\cdots p_b^{n_b}$$

$$=(r+1)Q-(r+1)Q\sum_{j=1}^{b}\frac{n_j}{n_j+1}p_j$$

即(14)，这证明了本引理。

现在回到原来的问题。样本空间$\mathscr{X}$中包含 b 个点 $a_1,\cdots a_b$。设当前样本 x^* 中，a_i 有 n_i 个，$i=1,\cdots,b$。又设

$$p_i=\int_I P_\theta(a_i)dG(\theta),\quad i=1,\cdots,b$$

这时有

$$w(x^*)=Q。$$

$$w'(x^*)=P(A)=(14)\text{的右边。}$$

因而

$$w'(x^*)/w(x^*)=(r+1)\left(1-\sum_{i=1}^{b}\frac{n_i}{n_i+1}p_i\right)$$

$$=(r+1)\sum_{j=1}^{b}\frac{1}{n_j+1}p_j$$

这个比值当然依赖于 p_i，即依赖于分布族$\{P_\theta,\ \theta\in I\}$，先验分布 G，以及所给的 x^*。这比值总大于1，这在直观上当然很明显。有两个情况可以讨论一下：

a。$r=1$，b 相当大，而分布比较均匀，即 $p_i\approx 0$。这时，$w'(x^*)/w(x^*)=2-p_i$，当$x^*=a_i$。因为 $p_i\approx 0$，有 $w'(x^*)/w(x^*)\approx 2$。由(9)、(12)、(13)可知，在这种情况下，使用Maritz的修正，可以把Johns原来的EB估计 δ_n 所算出的差距 $R(G,\delta_n)-R(G)$大约缩小一半。

b。r 很大，而 b 很小，这时，$n_j/r\approx p_j$，而

$$w'(x^*)/w(x)\approx(r+1)\sum_{j=1}^{b}\frac{1}{rp_j+1}p_j\approx\frac{r+1}{r}b\approx b。$$

就是说，使用Maritz的修正，可以把Johns原来的EB估计 δ_n 所算出的差距 $R(G,\delta_n)-R(G)$缩小到其 b 分之一。

当样本空间$\mathscr{X}$包含无限个点时，可以通过用有限样本空间逼近的方法去讨论。这时可以证明：只要取r充分大，比值$w'(x^*)/w(x)$可以任意接近于1的概率超出指定的任意常数。细节在此从略。

在结束本文前，我们顺便提一下Johns模型的一个有趣的应用。

设$\mathscr{X}=\{0,1,\cdots,m\}$，$P_\theta$为二项分布$B(m,\theta),0\leqslant\theta\leqslant 1$。在[6]中证明了：若不对先验分布族加上性质很特殊的限制，则（在平方损失下）θ的$a.o.$EB将不存在。用Johns的方法（配合Maritz的修正）可以提出θ的一个较好的EB估计。这基于以下容易证明的周知的事实。

引理 4. 设$X\sim B(m,\theta)$，取X个1，$m-X$个0，随机地排成一列（即：这m个数看作m个相异的物件，每个排列法的概率为$1/m!$），得$X_1,\cdots,X_m$，则$X_1,\cdots,X_m$独立同分布，且

$$P_\theta(X_1=1)=1-P_\theta(X_1=0)=\theta.$$

现设历史样本为$x_1,\cdots,x_n$，当前样本为x。取x个1，$m-x$个0，从其中随机地抽出一个，设结果记为y。分两种情况：

1. $y=1$。分别以n_1和n_2记$x_1,\cdots,x_n$中等于$x-1$和x的个数。用$\dfrac{n_2}{n_1+n_2}$估计θ。

2. $y=0$。分别以n_1和n_2记$x_1,\cdots,x_n$中等于x和$x+1$的个数，用$\dfrac{n_2}{n_1+n_2}$估计θ。

容易看出，这正是在$\mathscr{X}=\{0,1\}$而$r=m$的情况下，用Maritz修正所得的EB估计。

这个EB估计（记为δ_n）在$P_\theta\sim B(m-1,\theta)$的体系内是$a.o.$EB估计。当$m$较大时，它（从全面Bayes风险的角度去考察）与在$P_\theta\sim B(m,\theta)$（即原问题）的体系下可能得到的$a.o.$EB估计不会相差很多。这个结论（只能作为启发性的而非严格证明）可以用下面的考察来论证：考虑β先验分布族$\{\beta(a,b):\ a>0,b>0\}$。不难证明，在$P_\theta\sim B(n,\theta)$（及平方损失）之下，Bayes解的Bayes风险为

$$r(n,a,b)=ab/[(n+a+b)(1+a+b)(a+b)]$$

故对任何$a>0,b>$，有

$$r(m,a,b)/r(m-1,a,b)=\frac{m-1+a+b}{m+a+b}>\frac{m-1}{m}$$

就是说，虽然为了迁就Johns模型而将m改为$m-1$，Bayes风险最多也只增为原来的$\dfrac{m}{m-1}$倍，当m较大时（在二项分布中估计θ，m不能太小），这个增加很微小，所以，在$m-1$的情况下作出的$a.o.$ EB估计，其全面Bayes风险，较之在m的情况下作出的$a.o.$EB估计，即使有增加，增量也很小。由于当a,b在$(0,\infty)$独立变化时，$\beta(a,b)$概括了各种在实用上有意义的先验概率分布的情况（这里指的是：分布为单峰且包括了各种不同范围内集中的情况），因此，大体上可以认为上述结论对一切先验分布G都适用。

参　考　文　献

[1] Robbins, H., An empirical Bayes approach to statistics. *Proc. Third Berkeley Symp Math. Statist. Prob.* 1, (1955) 157—163, Univ. Calfornia Press.

[2] Johns. Jr. M.V., Nonparametric empirical Bayes procedures. *Ann. Math. Statist.* 28, (1957). 649—669.

[3] Maritz. J. S., Empirical Bayes Methods, *Methuen Monograph*, Methuen, London. (1970).

[4] Singh. R.S., Empirical Bayes estimation in Lebesgue-exponential families with rates near the best possible rate. Ann. Statist., 7, (1979). 890—901.

[5] Hoeffding, W. Probability inequalities for sums of bounded random variables, *J. Amer. Statist Assoc.* 58. (1963). 13-30.

[6] 陈希孺，二项分布参数的经验Bayes估计（将在《中国科技大学学报》发表）.

Some Notes on Johns' Empirical Bayes Estimation

BY Chen Xiru (陈希孺)

Abstract

Some problems related to the empirical Bayers estimate proposed by Johns[1](δ_n), and its modification(δ_n') by Maritz[3], are considered in this artical. The main result concerns the comparison of the remainders $R(G,\delta_n)-R(G)$ and $R(G,\delta_n')-R(G)$, where $R(G,\delta_n)$ and $R(G,\delta_n')$ denote the "over-all" Bayes risk(under quadratic loss) of δ_n and δ_n' respectively, and $R(G)$ denotes the Bayes risk of Bayes estimate, and G is the prior distribution.