贝叶斯统计学中先验分布的选择方法新探*

李 勇,孙 荣

(重庆工商大学 数学与统计学院,重庆 400067)

摘 要:在一个参数的可选先验分布类中选择一个合理先验的问题,类似于从参数空间中估计一个恰当参数的问题。因此,可利用贝叶斯分析的后验分布理论,先求出参数的后验分布,再根据后验分布中各个先验的相对似然选取似然最大的先验为合理先验,从而建立一个基于参数的后验分布的先验选择方法,它也是 ML - Ⅱ先验的一个拓广。

关键词:先验选择;后验分布;贝叶斯似然合理先验

中图分类号:O212.8 文献标识码:A 文章编号:1008 - 6439(2007)05-0067-03

Research into method selection of Bayesian Prior distribution

LI Yong, SUN Rong

(*School of Mathematics and Statistics, Chongqing Technology and Business University, Chongqing* 400067, *China*)

Abstract: Selecting a reasonable prior density in a series of selectable prior distributions of a parameter is similar to estimating a suitable parameter from parameter space. From this point, posterior distribution of Bayesian analysis can be used to solve posterior distribution of a parameter, then reasonable prior of maximum likelihood can be selected by depending on the relative likelihood of each prior of posterior distribution, a prior selection method of posterior distribution based on parameters is established. This method also extends ML - Ⅱ prior.

Key words: prior selection; posterior distribution; Bayesian Likelihood Reasonable Prior

贝叶斯统计学与经典统计学的主要分歧就在于先验分布的问题。其中先验分布的确定与选择是贝叶斯统计学首要的基本问题。对于先验分布的确定方法,已经有很多学者的研究成果。但所有这些确定先验分布的方法都只是提供了求先验分布的手段,而对于先验分布的合理性和准确性缺少合理选择评价体系。换句话说,对于同一个问题,不同的人根据不同的经验和规则,可以得出完全不同的先验分布,那么,对同一个问题,面对具有多个先验分布可供选择时,应该选择哪一个更加合理?这就是先验分布的选择问题。

对于这一问题,笔者在已发表文章(参考文献[1]、[2]、[3])中阐述了一个观点:在先验类中选择一个合理先验的问题,与在参数空间中估计一个恰当参数的问题类似,并提出了一个基于先验的后验分布的选择方法。本文在此基础上,进一步提出了一个基于参数的贝叶斯先验选择方法。其基本思路是:设 $X \sim f(x|\theta)$, $\theta \in \Theta$,参数 θ 的可选先验族 $\Gamma = \{\pi_i : \pi_i(\theta)$ 为 θ 的先验$\}$。先由可选先验 $\pi_i(\theta)$ 的先验分布 $\mu(\pi_i)$ 求出 θ 的先验 $\overline{\pi}(\theta) = \sum_i \mu(\pi_i)\pi_i$,再结合数据 x 求出参数 θ 的后验分布 $\overline{\pi}(\theta|x)$,最后根据参数 θ 的后验分布 $\overline{\pi}(\theta|x)$ 中各个先验 $\pi_i(\theta)$ 的相对似然选取似然最大的先验为合理先验。这样建立的先验选择方法称为基于参数的贝叶斯先验选择方法。

* 收稿日期:2007-07-25

作者简介:李 勇(1970—),男,重庆南川人,重庆工商大学数学与统计学院,讲师,从事统计学理论及应用研究。
孙 荣(1972—),男,重庆工商大学数学与统计学院,讲师,从事风险统计研究。

一、参数的后验分布计算方法

定理1：设 $X\sim f(x|\theta)$，$\theta\in\Theta$，参数 θ 的可选先验族 $\Gamma=\{\pi_i:\pi_i(\theta)$ 为 θ 的先验$\}$，对应的后验分布为 $\pi_i(\theta|x)$，相应的边缘分布为 $m(x|\pi_i)$，$\pi_i(\theta)$ 的先验分布为 $\mu(\pi_i)$。令 $\overline{\pi}(\theta)=\sum_i\mu(\pi_i)\pi_i$ 作为参数 θ 的先验密度，相应边缘分布为 $m(x|\overline{\pi})$，则参数 θ 的后验分布为：

$$\overline{\pi}(\theta|x)=\frac{1}{m(x|\overline{\pi})}\sum_i\mu(\pi_i)m(x|\pi_i)\pi_i(\theta|x)$$

证明：$\overline{\pi}(\theta|x)=\dfrac{f(x|\theta)\overline{\pi}(\theta)}{\int_\Theta f(x|\theta)\overline{\pi}(\theta)d\theta}$

$$=\sum_i\frac{\mu(\pi_i)m(x|\pi_i)}{m(x|\overline{\pi})}\cdot\frac{f(x|\theta)\pi_i}{\int_\Theta f(x|\theta)\pi_i(\theta)d\theta}$$

$$=\frac{1}{m(x|\overline{\pi})}\sum_i\mu(\pi_i)m(x|\pi_i)\pi_i(\theta|x)$$

推论：θ 的可选先验 $\Gamma=\{\pi_1(\theta),\pi_2(\theta)\}$，对应的后验分布为 $\pi_1(\theta|x)$，$\pi_2(\theta|x)$，其边缘分布分别为 $m(x|\pi_1)$、$m(x|\pi_2)$。又选择 π_1、π_2 的先验概率为 μ_1、μ_2，且 $\mu_1+\mu_2=1$，取 θ 的先验为 $\overline{\pi}(\theta)=\mu_1\pi_1(\theta)+\mu_2\pi_2(\theta)$，其边缘分布为 $m(x|\overline{\pi})$，则 θ 的后验分布为[4]：$\overline{\pi}(\theta|x)=\lambda(x)\pi_1(\theta|x)+[1-\lambda(x)]\pi_2(\theta|x)$，其中 $\lambda(x)=\dfrac{\mu_1m(x|\pi_1)}{m(x|\overline{\pi})}=\left[1+\dfrac{\mu_2m(x|\pi_2)}{\mu_1m(x|\pi_1)}\right]^{-1}$，$\mu_1+\mu_2=1$。

证明：由定理1得：

$$\overline{\pi}(\theta|x)=\frac{\mu_1m(x|\pi_1)}{m(x|\overline{\pi})}\pi_1(\theta|x)+\frac{\mu_2m(x|\pi_2)}{\mu_1m(x|\pi_1)+\mu_2m(x|\pi_2)}\pi_2(\theta|x)$$

$$=\frac{\mu_1m(x|\pi_1)}{m(x|\overline{\pi})}\pi_1(\theta|x)+\left[1-\frac{\mu_1m(x|\pi_1)}{m(x|\overline{\pi})}\right]\pi_2(\theta|x)$$

$$=\lambda(x)\pi_1(\theta|x)+[1-\lambda(x)]\pi_2(\theta|x)$$

$\lambda(x)=\dfrac{\mu_1m(x|\pi_1)}{m(x|\overline{\pi})}=\left[1+\dfrac{\mu_2m(x|\pi_2)}{\mu_1m(x|\pi_1)}\right]^{-1}$，$\mu_1+\mu_2=1$，故得证。

二、先验的选择方法

由定理1：利用似然原理，可以建立基于参数的后验分布的先验选择方法。首先给出（贝叶斯）极大似然的先验选择原理：设 $X\sim f(x|\theta)$，$\theta\in\Theta$，参数 θ 的可选先验族 $\Gamma=\{\pi_i:\pi_i(\theta)$ 为 θ 的先验$\}$，$\pi_i(\theta)$ 的先验分布为 $\mu(\pi_i)$，$m(x|\pi_i)$ 是 π_i 的似然函数。则在数据 x 下的一个合理先验就是选取似然程度与先验乘积 $\mu(\pi_i)m(x|\pi_i)$ 最大的那个先验。

定义1：设 $X\sim f(x|\theta)$，$\theta\in\Theta$，参数 θ 的可选先验族 $\Gamma=\{\pi_i:\pi_i(\theta)$ 为 θ 的先验$\}$，对应的后验分布为 $\pi_i(\theta|x)$，相应的边缘分布为 $m(x|\pi_i)$，$\pi_i(\theta)$ 的先验分布为 $\mu(\pi_i)$。令 $\overline{\pi}(\theta)=\sum_i\mu(\pi_i)\pi_i$ 作为参数 θ 的先验分布，相应边缘分布为 $m(x|\overline{\pi})$，且参数 θ 的后验分布为 $\overline{\pi}(\theta|x)=\dfrac{1}{m(x|\overline{\pi})}\sum_i\mu(\pi_i)m(x|\pi_i)\pi_i(\theta|x)$。若存在 $\pi_0(\theta)\in\Gamma$，使得 $\pi_0(\theta)=\max_{\pi_i\in\Gamma}\{\mu(\pi_i)m(x|\pi_i)\}$，则称先验 $\pi_0(\theta)$ 为 Γ 中的贝叶斯似然合理先验。

定理2：设参数 θ 的可选先验族 $\Gamma=\{\pi_i:\pi_i(\theta)$ 为 θ 的先验$\}$，$\pi_i(\theta)$ 的先验分布为 $\mu(\pi_i)$ 为均匀分布。则 Γ 中的贝叶斯似然合理先验就是 $ML-\mathrm{II}$ 先验。

证明：由于 $\mu(\pi_i)$ 为均匀分布，则贝叶斯似然合理先验 $\pi_0(\theta)=\max_{\pi_i\in\Gamma}\{\mu(\pi_i)m(x|\pi_i)\}=\mu(\pi_i)\max_{\pi\in\Gamma}\{m(x|\pi_i)\}$。

显然，$\pi_0(\theta)$ 的最值与 $\max_{\pi\in\Gamma}\{m(x|\pi_i)\}$ 最值一致，而后者正是 $ML-\mathrm{II}$ 先验。故得证。

从上面定理可以看出，用 $ML-II$ 先验方法来选择先验，只限于对可选先验本身就是合理的，且对其选择的主观概率大体相近时，才能得出较好的选择。并不能反映选择者的主观信息。而从本文建立的选择方法来看，选择合理先验还与对可选先验正确性的主观概率有关。同时也说明 $ML-II$ 先验选择方法是本文所建立的方法的特殊情形。

三、实例应用

例1：设 $X\sim N(\theta,1)$，$\theta\in\Theta=(-\infty,+\infty)$。参数 θ 的先验分布的中位数为0，4分位数（即1/4分位点与3/4分位点）为 -1 和1，且参数的先验分布仅限于正态和柯西分布。查表可得参数 θ 的可选先验 $\Gamma=\{\pi_c,\pi_N\}$，其中 $\pi_N=N(0,2.19)$，$\pi_c=C(0,1)$（柯西分布）。下面结合样本数据 x，根据不同情

况对参数的两个可选先验$\{\pi_c, \pi_N\}$作出合理选择。

(1) 根据 $ML-\mathrm{II}$ 先验法选择合理先验。

(2) 设选择 π_c, π_N 的先验分布为$\{\mu(\pi_c), \mu(\pi_N)\}$，其中$\mu(\pi_c)=p_c, \mu(\pi_N)=p_N$。求贝叶斯似然合理先验。

解：(1) 设样本数据为 x，其边缘分布分别为 $m(x|\pi_N)=N(0,3.19)$，

$$m(x|\pi_c)=\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}(x-\theta)^2\right\}\frac{1}{\pi(1+\theta^2)}d\theta$$

下表给出各种 x 的 $m(x|\pi_N)$、$m(x|\pi_c)$值：

x	0	4.5	6.0	10
$m(x\|\pi_N)$	0.22	0.0093	0.00079	3.5×10^{-6}
$m(x\|\pi_c)$	0.21	0.0180	0.00940	0.0032

显然，当 x 较小时，数据对二者的支持度差不多；当 x 较大时，π_c 为 $ML-\mathrm{II}$ 先验。

(2) 可得 θ 的先验分布为 $\pi(\theta)=\mu(\pi_N)\pi_N(\theta)+\mu(\pi_c)\pi_c(\theta)$，由推论1得其后验分布为：$\overline{\pi}(\theta|x)=\lambda(x)\pi_N(\theta|x)+[1-\lambda(x)]\pi_C(\theta|x)$，其中 $\lambda(x)=\left[1+\frac{p_c m(x|\pi_c)}{p_N m(x|\pi_N)}\right]^{-1}$。若$\frac{p_c m(x|\pi_c)}{p_N m(x|\pi_N)}>1$，有 $\lambda(x)<\frac{1}{2}$，则贝叶斯似然合理先验应为 π_c。

若 $p_N=0.3, p_c=0.7$，得：

x	0	4.5	6.0	10
$p_N m(x\|\pi_N)$	0.066	0.00279	0.000237	1.05×10^{-6}
$p_c m(x\|\pi_c)$	0.147	0.01260	0.006580	0.00224

若 $p_N=0.7, p_c=0.3$，得：

x	0	4.5	6.0	10
$p_N m(x\|\pi_N)$	0.154	0.00651	0.000553	2.45×10^{-6}
$p_c m(x\|\pi_c)$	0.063	0.00540	0.002820	0.00096

若选择 π_N 与 π_C 的主观概率一样，则贝叶斯似然合理先验正是 $ML-\mathrm{II}$ 先验。

从例1可知，在选择合理先验的过程中，恰当的先验信息与数据信息的有机结合[3][4]，可以得出有效的先验选择。这比只利用样本信息作出选择(即利用 $ML-\mathrm{II}$ 先验选择)要合理些，因为按贝叶斯似然合理先验的选择反映了选择者对被选择先验的主观信息。但是，也可以得出，即是对 $\mu(\pi_N)$ 赋予较大的先验，当样本数据 x 增大后($x\geq6$)，后验值对它仍不利。也就是在选择似然合理先验中，样本数据的信息占主导地位。而且当样本信息足够多时，两种似然选择标准的结果完全一致；在缺乏样本信息的情况下，先验信息的地位就显露出来，这时此处的方法就明显的优于 $ML-\mathrm{II}$ 先验选择方法。

贝叶斯理论中先验分布的确定，是利用贝叶斯方法的关键。能够选出合理的先验分布，对于问题解决至关重要。本文在参考文献[1]、[2]、[3]的基础上，建立了基于参数的后验分布的先验选择方法，它对 $ML-\mathrm{II}$ 先验选择方法有一定的拓展，它更好地体现了选择者的主观信息，更具有先验选择的合理性。对应用贝叶斯方法解决实际问题，提供了一定的有效手段。

参考文献：

[1]李勇. 基于参数的贝叶斯先验选择方法[J]. 西南师范大学学报(自然科学版)，2007，32(2)：1-2.

[2]李勇. 基于先验的贝叶斯先验选择方法[J]. 重庆工商大学学报(自然科学版)，2006，23(6)：548-550.

[3]李勇，易文德. 贝叶斯分析中先验分布优选的方法[J]. 渝西学院学报(自然科学版)，2005，4(4)：5-7.

[4] Berger J O. Statistical Decision Theory and Bayesian Analysis[M]. New York: Springer-Verlag, 1985.

[5]Basu S. Posterior sensitivity to the sampling distribution and the prior: more than one observation[J]. Ann Inst Statist Math, 1999, 51(3): 499-513.

(责任编校：夏　冬)