

第 31 章 贝叶斯估计简介

31.1 贝叶斯估计的思想

在统计学中有两派, 主流是“频率学派”(frequentist school), 也称“古典学派”(Classical school), 即数理统计课的常规内容;

另一派是“贝叶斯学派”(Bayesian school), 由 18 世纪英国统计学家贝叶斯(Thomas Bayes)创立。

二者的主要区别在于，频率学派假设总体服从某个分布，比如 $f(x; \theta)$ ，其中 θ 为待估计、未知、给定的参数(或参数向量)。

贝叶斯学派认为，既然 θ 有不确定性，应将 θ 本身也视为随机变量(或随机向量)，并用概率分布来描述。

由于 θ 的分布在研究者看到数据之前就有，故称为“先验分布”(prior distribution)。

得到样本数据后，可根据贝叶斯定理(Bayes' Theorem)将先验分布更新(update)为后验分布(posterior distribution)，并以后验分布作为统计推断的依据。

由于先验分布的主观性(subjectivity), 有些学者对贝叶斯学派持有保留意见。但在大样本中, 先验分布的作用将变得很小。

还可使用“相对不含信息”(relatively uninformative)的先验分布, 并对后验分布对于先验分布的依赖性进行“敏感度分析”(sensitivity analysis)。

贝叶斯估计的主要优点:

(1) 古典学派一般通过最优化(比如, MLE, OLS)进行参数估计, 但有时不易求得最优解。贝叶斯学派只要反复使用贝叶斯定理即可, 不需要进行最优化。虽然贝叶斯分析常没有解析解, 随着计算方法的发展, 这已基本不成问题。

(2) 古典学派需要用不同的统计量来估计期望、方差、中位数、分位数等，而贝叶斯学派可以直接得到参数的整个后验分布。从后验分布，可容易地算出其各阶矩。

(3) 对于古典学派的统计量，常常不易找出其“精确的有限样本分布”(exact finite-sample distribution)，故只能退而求其次，推导大样本渐近分布。贝叶斯学派一般可直接计算精确的有限样本分布，不需要渐近理论。

31.2 贝叶斯定理

贝叶斯估计的实质在于，反复使用贝叶斯定理，将先验分布与样本数据综合为后验分布。

对于随机事件 A 与 B ，有如下贝叶斯公式：

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B)}$$

其中，第一个等号为条件概率的定义，而第二个等号使用了概率的乘法公式。

如果将 $P(A)$ 视为先验概率，将 B 视为样本数据，则贝叶斯公式给出了在看到样本数据 B 后，如何将先验概率 $P(A)$ 更新为后验概率 $P(A|B)$ 的规则。

一般地，对于随机向量 θ (视为参数)与随机向量 y (视为样本数据)，根据贝叶斯定理(Bayes' Theorem)可知，

$$f(\theta|y) = \frac{f(\theta, y)}{f(y)} = \frac{f(y|\theta)\pi(\theta)}{f(y)}$$

其中， $f(\theta|y)$ 为看到数据 y 之后 θ 的条件分布密度(即后验分布)， $\pi(\theta)$ 为参数 θ 的先验分布密度， $f(\theta, y)$ 为 θ 与 y 的联合分布， $f(y|\theta)$ 为给定参数 θ 时 y 的密度函数，而 $f(y)$ 为 y 的边缘分布密度。

在联合分布 $f(\boldsymbol{\theta}, \mathbf{y})$ 中将随机参数 $\boldsymbol{\theta}$ 积分掉，可得 \mathbf{y} 的边缘密度：

$$f(\mathbf{y}) = \int f(\boldsymbol{\theta}, \mathbf{y}) d\boldsymbol{\theta} = \int f(\mathbf{y} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

在贝叶斯分析中，把后验分布 $f(\boldsymbol{\theta} | \mathbf{y})$ 记为 $p(\boldsymbol{\theta} | \mathbf{y})$ (p 表示 posterior)，而把 \mathbf{y} 的密度函数 $f(\mathbf{y} | \boldsymbol{\theta})$ 记为似然函数 $L(\boldsymbol{\theta}; \mathbf{y})$ 。

在后验分布公式中，分母为边缘分布 $f(\mathbf{y})$ ，不包含 $\boldsymbol{\theta}$ ，故可以将其视为常数。故后验分布与该公式的分子成正比：

$$f(\boldsymbol{\theta} | \mathbf{y}) \propto f(\mathbf{y} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta})$$

其中，“ \propto ” 表示 “成正比”。

省略常数可以简化后验分布的推导，被省的常数可在以后加上。

省去常数的密度函数被称为“密度核”(density kernel)。

$L(\boldsymbol{\theta}; \mathbf{y})\pi(\boldsymbol{\theta})$ 就是后验分布 $p(\boldsymbol{\theta} | \mathbf{y})$ 的密度核。

31.3 贝叶斯估计的一个例子

记随机样本为 $\mathbf{y} = (y_1 \ y_2 \ \cdots \ y_n)'$ ，其中 $y_i \sim N(\theta, \sigma^2)$ ，方差 σ^2 已知，而均值 θ 未知。

古典学派选择 θ 使似然函数最大化，得到 $\hat{\theta}_{\text{MLE}} = \bar{y} \equiv \frac{1}{n} \sum_{i=1}^n y_i$ 。

贝叶斯学派则要额外地设定 θ 的先验分布。

为了计算方便，选择先验正态分布，即 $\theta \sim N(\mu, \tau^2)$ ，其中先验均值 μ 与先验方差 τ^2 为已知常数。

如果 τ^2 较大，就表示先验分布的不确定性较大。

目标是求出后验分布 $p(\theta | \mathbf{y})$ 。把此计算过程分为以下三步。

第一步 写出先验分布密度 $\pi(\theta)$ 。

在贝叶斯分析中，使用方差的倒数有时更为方便。

定义 θ 的“精确度”(precision)为

$$h \equiv 1/\tau^2$$

精确度 h 越大, 方差 τ^2 就越小, 表明对随机变量 θ 知道得越精确。

由于样本均值 \bar{y} 的方差为 (σ^2/n) , 记 \bar{y} 的精确度为 $h^* \equiv n/\sigma^2$ 。

θ 的先验密度为

$$\begin{aligned}\pi(\theta) &= (2\pi\tau^2)^{-1/2} \exp\{-(\theta - \mu)^2 / 2\tau^2\} && \text{(一元正态密度)} \\ &\propto \exp\{-h(\theta - \mu)^2 / 2\} && \text{(去掉不含}\theta\text{的常数项,} \\ &&& \text{代入} h \equiv 1/\tau^2\text{)}\end{aligned}$$

第二步 写出样本数据 \mathbf{y} 的联合密度，即似然函数：

$$\begin{aligned} L(\theta; \mathbf{y}) &= \prod_{i=1}^n (2\pi\sigma^2)^{-1/2} \exp\left\{-(y_i - \theta)^2 / 2\sigma^2\right\} \quad (n \text{ 个一元正态密度相乘}) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\sum_{i=1}^n (y_i - \theta)^2 / 2\sigma^2\right\} \quad (\text{合并同类项}) \\ &\propto \exp\left[-\sum_{i=1}^n (y_i - \bar{y} + \bar{y} - \theta)^2 / 2\sigma^2\right] \quad (\text{去掉常数项, 加减 } \bar{y}) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2} \left[\sum_{i=1}^n (y_i - \bar{y})^2 + 2\sum_{i=1}^n (y_i - \bar{y})(\bar{y} - \theta) + \sum_{i=1}^n (\bar{y} - \theta)^2 \right]\right\} \\ &\quad (\text{展开平方项}) \\ &\propto \exp\left\{-\sum_{i=1}^n (\bar{y} - \theta)^2 / 2\sigma^2\right\} \end{aligned}$$

(去掉常数 $\sum_{i=1}^n (y_i - \bar{y})^2$, 而 $\sum_{i=1}^n (y_i - \bar{y})(\bar{y} - \theta) = (\bar{y} - \theta) \sum_{i=1}^n (y_i - \bar{y}) = 0$)

$$\propto \exp\left\{-n(\bar{y} - \theta)^2 / 2\sigma^2\right\} \quad (n \text{ 个相同的数相加变为乘法})$$

$$\propto \exp\left\{-h^*(\bar{y} - \theta)^2 / 2\right\} \quad (\text{定义 } h^* \equiv n/\sigma^2)$$

似然函数的形式与先验密度相似, 这为后面的计算提供了方便。

第三步 根据贝叶斯定理，写出后验分布的密度核。

$$\begin{aligned} L(\theta; \mathbf{y})\pi(\theta) &\propto \exp\left\{-h^*(\bar{y}-\theta)^2/2\right\} \cdot \exp\left\{-h(\theta-\mu)^2/2\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[h^*(\bar{y}-\theta)^2+h(\theta-\mu)^2\right]\right\} \quad (\text{指数相加}) \end{aligned}$$

其中，指数项中的方括弧项可以简化为

$$\begin{aligned} &h^*(\bar{y}-\theta)^2+h(\theta-\mu)^2 \\ &=h^*(\bar{y}^2-2\bar{y}\theta+\theta^2)+h(\theta^2-2\mu\theta+\mu^2) \quad (\text{展开平方项}) \\ &=(h+h^*)\theta^2-2(h\mu+h^*\bar{y})\theta+h^*\bar{y}^2+h\mu^2 \quad (\text{根据}\theta\text{合并同类项}) \\ &=\bar{h}\theta^2-2\bar{h}\bar{\mu}\theta+h^*\bar{y}^2+h\mu^2 \quad (\text{定义}\bar{h}\equiv h+h^*, \bar{\mu}\equiv\frac{h\mu+h^*\bar{y}}{\bar{h}}) \\ &=\bar{h}(\theta^2-2\bar{\mu}\theta)+h^*\bar{y}^2+h\mu^2 \quad (\text{提取公因子}\bar{h}) \\ &=\bar{h}(\theta^2-2\bar{\mu}\theta+\bar{\mu}^2)-\bar{h}\bar{\mu}^2+h^*\bar{y}^2+h\mu^2 \quad (\text{配方}) \\ &\propto \bar{h}(\theta-\bar{\mu})^2 \quad (\text{去掉不含}\theta\text{的常数项}) \end{aligned}$$

根据以上推导可知，

$$L(\theta; \mathbf{y})\pi(\theta) \propto \exp\left\{-\frac{1}{2}\left[\bar{h}(\theta - \bar{\mu})^2\right]\right\}$$

这还是一个正态分布的密度核，故后验分布仍是正态分布。

后验分布的精确度提高为

$$\bar{h} \equiv h + h^*$$

即先验精确度 h 与样本精确度 h^* 之和。后验分布的期望值调整为

$$\bar{\mu} \equiv (h\mu + h^*\bar{y})/\bar{h}$$

即先验均值 μ 与样本均值 \bar{y} 之加权平均，权重为各自的精确度。

如果先验信息不精确，即 h 较小，则先验均值对于后验均值的影响就较小。

上述方程给出了如何将“先验信息”与“样本信息”综合成“后验信息”的规则，即“贝叶斯更新规则”(Bayesian updating rule)。

由于 $\bar{h} = h + h^* = h + \frac{n}{\sigma^2}$ ，故当 $n \rightarrow \infty$ 时，后验精确度 $\bar{h} \rightarrow \infty$ (后验方差趋于 0)；而样本精确度对后验精确度的贡献变大，即 $h^*/\bar{h} \rightarrow 1$ 。

当 $n \rightarrow \infty$ 时，后验均值 $\bar{\mu} \rightarrow \bar{y}$ ，即完全由样本数据决定，不受先验分布的影响。

直观来看，当 $n \rightarrow \infty$ 时，后验分布 $\theta | \mathbf{y} \xrightarrow{d} N(\bar{y}, \sigma^2/n)$ 。对比古典学派的结果则为 $\bar{y} \xrightarrow{d} N(\theta, \sigma^2/n)$ 。

当样本容量越来越大时，先验分布所起的作用越来越小。

此结论在一般情况下也成立。回到后验分布密度核的一般表达式：

$$p(\boldsymbol{\theta} | \mathbf{y}) \propto L(\boldsymbol{\theta}; \mathbf{y})\pi(\boldsymbol{\theta})$$

$$\propto L(\boldsymbol{\theta}; y_1, \dots, y_n)\pi(\boldsymbol{\theta})$$

$$\propto L(\boldsymbol{\theta}; y_1) \cdots L(\boldsymbol{\theta}; y_n)\pi(\boldsymbol{\theta}) \quad (\text{假设样本为 iid})$$

后验分布密度核的对数为

$$\ln p(\boldsymbol{\theta} | \mathbf{y}) \propto \ln \pi(\boldsymbol{\theta}) + \sum_{i=1}^n \ln L(\boldsymbol{\theta}; y_i)$$

当样本容量增大时，对数先验密度 $\ln \pi(\boldsymbol{\theta})$ 始终不变，而对数似然函数之和 $\sum_{i=1}^n \ln L(\boldsymbol{\theta}; y_i)$ 包含的项数越来越多，故越来越以后者为主。

此结论有助于冲淡人们对于贝叶斯估计结果依赖于主观先验分布的顾虑。

得到后验分布之后，可以把它视为“修正的先验分布”(revised prior)，并作为未来的先验分布，如此反复，不断更新我们对世界的认识。

31.4 基于后验分布的统计推断

有了后验分布，就可进行一系列的统计推断。

1. 边缘后验分布

假设参数 $\boldsymbol{\theta}$ 为向量, $\boldsymbol{\theta} = (\theta_1 \cdots \theta_q)'$ 。

知道 $\boldsymbol{\theta}$ 的后验分布 $p(\boldsymbol{\theta} | \mathbf{y})$ 之后, 可求得单个参数 θ_k ($1 \leq k \leq q$)的“边缘后验分布”(marginal posterior):

$$p(\theta_k | \mathbf{y}) = \int p(\boldsymbol{\theta} | \mathbf{y}) d\theta_1 \cdots d\theta_{k-1} d\theta_{k+1} \cdots d\theta_q$$

边缘后验分布常常既不对称也非“单峰”(unimodal), 这与古典学派的统计量不同。

2. 后验分布的各阶矩

根据后验分布，可以计算其各阶矩，比如均值、中位数、方差等。

3. 点估计

在贝叶斯分析中，未知参数 θ 被视为随机变量，而非一个固定的点，故点估计(point estimation)在贝叶斯分析中不那么重要，关注的重点是 θ 的整个后验分布。

后验均值(posterior mean)或后验中位数(posterior median)常常被作为点估计来汇报。

4. 区间估计

知道参数 θ 的整个后验分布后，很容易找到置信区间或置信区域 (confidence interval or region)。

但置信度为 $(1-\alpha)$ (比如， $\alpha = 5\%$)的置信区间并不唯一。一个简单的做法是选择分位数 $\alpha/2$ 与分位数 $(1-\alpha/2)$ 之间的区间。

对贝叶斯置信区间的解释与频率派的解释完全不同。假设 θ 的 95% 后验置信区间为 $(1, 3)$ ，可直接说 θ 落在区间 $(1, 3)$ 的概率为 95%。

对于频率派而言，如果 θ 的 95% 置信区间为 $(1, 3)$ ，则只能说，如果进行 100 次同样的随机抽样，大约有 95 次，这些随机的置信区间能覆盖真实的参数值 θ 。

5. 假设检验

由于贝叶斯分析认为参数不是固定的，故像“ $H_0: \theta = \theta_0$ ”这类的假设检验通常也就没有意义(假设 θ 为连续变量)。

但可检验参数 θ 是否属于参数空间 Θ 的某个子集 Θ_1 ，比如检验原假设“ $H_0: \theta \in \Theta_1 \subset \Theta$ ”，而替代假设为“ $H_1: \theta \in \Theta_2 \subset \Theta$ ”，其中 Θ_1 与 Θ_2 可以有交集，即允许“非嵌套式”(non-nested)检验。

在古典学派中，“ $H_0: \theta \in \Theta_1$ ”成立的概率要么为 0，要么为 1。贝叶斯学派直接计算“ $H_0: \theta \in \Theta_1$ ”成立的后验概率：

$$P(H_0 | \mathbf{y}) = \frac{P(\mathbf{y} | H_0)P(H_0)}{P(\mathbf{y})}$$

其中， \mathbf{y} 为样本数据。

替代假设 “ $H_1: \theta \in \Theta_2$ ” 成立的后验概率为

$$P(H_1 | \mathbf{y}) = \frac{P(\mathbf{y} | H_1)P(H_1)}{P(\mathbf{y})}$$

两个假设的后验概率之比称为“后验几率比” (posterior odds ratio):

$$\underbrace{\frac{P(H_0 | \mathbf{y})}{P(H_1 | \mathbf{y})}}_{\text{后验几率比}} = \underbrace{\frac{P(\mathbf{y} | H_0)}{P(\mathbf{y} | H_1)}}_{\text{贝叶斯因子}} \cdot \underbrace{\frac{P(H_0)}{P(H_1)}}_{\text{先验几率比}}$$

其中, $B_{01} \equiv \frac{P(y | H_0)}{P(y | H_1)}$ 被称为 “贝叶斯因子” (Bayes factor),

$\frac{P(H_0)}{P(H_1)}$ 称为 “先验几率比” (prior odds ratio)。

后验几率比等于贝叶斯因子与先验几率比之乘积。

如果后验几率比大于 1, 则数据显示我们更应相信 H_0 而不是 H_1 。

贝叶斯假设检验的实质是, 计算不同假设的后验概率, 然后进行比较。这与古典假设检验的小概率事件的 “反证法” 思想很不相同。

31.5 先验分布的选择

1. 无信息先验分布

贝叶斯分析面临的一个主要挑战是如何确定先验分布(specification of the prior)。

由于先验分布是贝叶斯分析中主观性的来源，为了减少主观性，一种方法是选择基本不带信息的先验分布，即“无信息先验分布”(uninformative prior)，以减少先验分布对后验分布的影响。

如果参数空间 Θ 有界，可使用均匀分布作为无信息先验分布，即 $\pi(\boldsymbol{\theta}) = c > 0, \forall \boldsymbol{\theta} \in \Theta$ ，称为“均匀先验分布”(uniform prior)。

如果参数空间 Θ 是无界的，则均匀先验分布是“非正常密度函数”(improper density)，因为 $\int \pi(\theta) d\theta = \infty$ 而不是1。这可能导致后验分布也是非正常密度函数。

均匀先验分布的另一缺点是，在作参数变换(reparameterization)时，不满足不变性(invariance)。

【例】 假设一维参数 $\theta > 0$ 服从均匀先验分布，即 $\pi(\theta) = c > 0$ 。考虑参数变换 $\gamma = \ln \theta$ ，则 $-\infty < \gamma < \infty$ 。

根据随机变量函数的概率分布公式， γ 的先验密度为 $\pi^*(\gamma) = \pi(\theta) \left| \frac{d\theta}{d\gamma} \right| = ce^\gamma$ ，故不再是均匀分布。似乎无信息的参数 θ 经过参数转换变成 γ 后，却带有明显的信息了。

为避免均匀先验分布的非正常密度函数，同时维持很少的信息，可假设先验分布的方差很大。比如， $\theta \sim N(\mu, \tau^2)$ ，其中 τ^2 很大。

这种先验分布称为“模糊先验分布”或“弥漫先验分布”(vague or diffuse prior)，但在参数变换下也不满足不变性。

实践中较常用的无信息先验分布是 Jeffreys(1946)提出的如下“杰佛里先验分布”(Jeffreys' prior)：

$$\pi(\boldsymbol{\theta}) \propto |\mathbf{I}(\boldsymbol{\theta})|^{1/2} = \left| -\mathbf{E} \left(\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right) \right|^{1/2}$$

杰佛里先验分布的密度核是信息矩阵 $\mathbf{I}(\boldsymbol{\theta})$ 的行列式的平方根，其中 $L \equiv L(\boldsymbol{\theta}; \mathbf{y})$ 为似然函数。

杰佛里先验分布的优点是具有不变性，无论作什么参数变换，先验分布的形式不变。

下面以一维情形为例来证明其不变性。

证明：对于参数 θ ，其杰佛里先验分布为 $\pi(\theta) \propto |\mathbf{I}(\theta)|^{1/2}$ 。

假设参数变换为 $\gamma = h(\theta)$ ，则需要证明， γ 的杰佛里先验分布为 $\pi^*(\gamma) \propto |\mathbf{I}(\gamma)|^{1/2}$ 。

根据随机变量函数的分布公式，

$$\pi^*(\gamma) = \pi(\theta) \left| \frac{\partial \theta}{\partial \gamma} \right| \propto |\mathbf{I}(\theta)|^{1/2} \left| \frac{\partial \theta}{\partial \gamma} \right|$$

因此，只要证明 $|\mathbf{I}(\gamma)|^{1/2} = |\mathbf{I}(\theta)|^{1/2} \left| \frac{\partial \theta}{\partial \gamma} \right|$ 即可。

为此，计算 $\mathbf{I}(\gamma) = -\mathbf{E} \left[\frac{\partial^2 \ln L}{\partial \gamma^2} \right]$ 。

$$\frac{\partial \ln L}{\partial \gamma} = \frac{\partial \ln L}{\partial \theta} \cdot \frac{\partial \theta}{\partial \gamma} \quad (\text{使用微分连锁法则求一阶导})$$

$$\frac{\partial^2 \ln L}{\partial \gamma^2} = \frac{\partial^2 \ln L}{\partial \theta^2} \cdot \left(\frac{\partial \theta}{\partial \gamma} \right)^2 + \frac{\partial \ln L}{\partial \theta} \cdot \frac{\partial^2 \theta}{\partial \gamma^2} \quad (\text{使用微分连锁法则求二阶导})$$

其中， $\frac{\partial \ln L}{\partial \theta}$ 为得分函数，其期望值为 $\mathbf{E} \left[\frac{\partial \ln L}{\partial \theta} \right] = 0$ 。因此，

$$\mathbf{I}(\gamma) = -\mathbf{E} \left[\frac{\partial^2 \ln L}{\partial \gamma^2} \right] = -\mathbf{E} \left[\frac{\partial^2 \ln L}{\partial \theta^2} \right] \left(\frac{\partial \theta}{\partial \gamma} \right)^2 = \mathbf{I}(\theta) \left(\frac{\partial \theta}{\partial \gamma} \right)^2$$

对方程两边取行列式，然后开平方可得

$$|\mathbf{I}(\gamma)|^{1/2} = |\mathbf{I}(\theta)|^{1/2} \left| \frac{\partial \theta}{\partial \gamma} \right|$$

因此， $\pi^*(\gamma) \propto |\mathbf{I}(\gamma)|^{1/2}$ ，故杰佛里先验分布具备不变性。

例 正态分布的杰佛里先验分布。假设 $y \sim N(\mu, \sigma^2)$ 。

情形一， μ 未知而 σ^2 已知，则 $I(\mu) = -E\left[\frac{\partial^2 \ln L}{\partial \mu^2}\right] = \frac{n}{\sigma^2}$ 。

由于 σ^2 已知，故杰佛里先验分布为 $|I(\mu)|^{1/2} \propto c$ ，其中 c 为常数。

此时，杰佛里先验分布为非正常密度函数。

情形二， σ^2 未知而 μ 已知，则 $I(\sigma^2) = -E\left[\frac{\partial^2 \ln L}{\partial(\sigma^2)^2}\right] = \frac{n}{2\sigma^4}$ ，故杰佛里先验分布为 $|I(\sigma^2)|^{1/2} \propto \sigma^{-2}$ 。

情形三， μ 与 σ^2 皆未知，则

$$I(\mu, \sigma^2) = -E \begin{bmatrix} \frac{\partial^2 \ln L}{\partial \mu^2} & \frac{\partial^2 \ln L}{\partial \mu \partial (\sigma^2)} \\ \frac{\partial^2 \ln L}{\partial \mu \partial (\sigma^2)} & \frac{\partial^2 \ln L}{\partial (\sigma^2)^2} \end{bmatrix} = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix} = \frac{n^2}{2\sigma^6},$$

故杰佛里先验分布为 $|I(\mu, \sigma^2)|^{1/2} \propto \sigma^{-3}$ 。

在此例中，在第一种情形下，杰佛里先验分布为均匀分布，但在后两种情形下都不是均匀分布。

究竟杰佛里先验分布在何种意义上是无信息先验分布并不清楚。

2. 自然共轭先验分布

对于不同的样本密度函数(sample density)，常选择合适的先验分布，使得样本密度函数、先验分布与后验分布都“属于同一族的密度函数”(in the same class of densities)，具有同样的函数形式。

这种先验分布称为“自然共轭先验分布”(natural conjugate prior)，而这样的先验分布与样本密度函数被称为“自然共轭对”(natural conjugate pair)。

比如，在前例，方差已知而均值未知的正态分布，其自然共轭先验分布就是正态分布。

由于样本密度函数与先验分布的密度核在形式上都是指数函数，故把它们相乘时，得到的后验分布仍然具有指数函数的形式。

使用共轭先验分布的另一好处是，由于后验分布与先验分布的形式相同，容易把后验分布作为下一轮估计的先验分布。

3. 敏感度分析

由于贝叶斯分析的结果依赖于先验分布，而我们对先验分布的具体形式可能并无把握，故常需要进行敏感度分析，即考虑使用不同的先验分布对后验分布的影响有多大；或考虑使用不同的样本密度函数的影响。

31.6 多元回归的贝叶斯分析

考虑以下的多元回归模型：

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

其中， \mathbf{X} 为 $n \times K$ 满列秩数据矩阵，扰动项 $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ 。

给定 $(\mathbf{X}, \boldsymbol{\beta}, \sigma^2)$ ，条件分布 $\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$ 。

以无信息的杰佛里先验分布为例。

根据前例，对于 $y \sim N(\mu, \sigma^2)$ ，当 μ 未知而 σ^2 已知时， μ 的杰佛里先

验分布为常数；当 σ^2 未知而 μ 已知时， σ^2 的杰佛里先验分布与 $(1/\sigma^2)$ 成正比。

推广到多元回归的情形，意味着 $\pi(\beta_j) \propto c, j=1, \dots, K$ ，而 $\pi(\sigma^2) \propto 1/\sigma^2$ 。

假设 $\boldsymbol{\beta}$ 与 σ^2 相互独立，则 $(\boldsymbol{\beta}, \sigma^2)$ 的联合先验分布(joint prior)为

$$\pi(\boldsymbol{\beta}, \sigma^2) \propto 1/\sigma^2$$

样本数据的似然函数可以写为

$$L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}, \mathbf{X}) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$$

定义最小二乘估计量为 $\hat{\beta} \equiv (X'X)^{-1} X'y$ ，残差向量 $e \equiv y - X\hat{\beta}$ 。

由于 $y = X\hat{\beta} + e$ ，故 $y - X\beta = X\hat{\beta} + e - X\beta = e - X(\beta - \hat{\beta})$ 。可得：

$$\begin{aligned} L(\beta, \sigma^2; y, X) &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} [e' - (\beta - \hat{\beta})' X'] [e - X(\beta - \hat{\beta})]\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} [e'e + (\beta - \hat{\beta})' X'X(\beta - \hat{\beta})]\right\} \\ &\quad \text{(因为 } X'e = 0\text{)} \\ &\propto (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} [(n-K)s^2 + (\beta - \hat{\beta})' X'X(\beta - \hat{\beta})]\right\} \end{aligned}$$

(其中，样本方差 $s^2 \equiv e'e/(n-K)$)

因此，联合后验分布满足

$$\begin{aligned}
 p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}, \mathbf{X}) &\propto \left(1/\sigma^2\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-K)s^2 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right]\right\} \underbrace{\left(1/\sigma^2\right)} \\
 &\propto \left(1/\sigma^2\right)^{\frac{n}{2}+1} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-K)s^2 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right]\right\} \\
 &\propto \left(1/\sigma^2\right)^{\frac{n}{2}+1} \exp\left\{-\frac{(n-K)s^2}{2\sigma^2} - \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \left(\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right\} \\
 &\propto \left(1/\sigma^2\right)^{K/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \left(\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right\} \\
 &\times \left(1/\sigma^2\right)^{\frac{n-K}{2}+1} \exp\left\{-(n-K)s^2/2\sigma^2\right\} \quad (\text{拆成两部分, 第二部分不含 } \boldsymbol{\beta})
 \end{aligned}$$

给定 σ^2 ，则 $\boldsymbol{\beta}$ 的条件后验分布(conditional posterior)为：

$$p(\boldsymbol{\beta} | \sigma^2, \mathbf{y}, \mathbf{X}) \propto (1/\sigma^2)^{K/2} \exp \left\{ -\frac{1}{2} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' (\sigma^2 (\mathbf{X}'\mathbf{X})^{-1})^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right\}$$

这是一个 K 维正态分布，其期望为 $\boldsymbol{\beta}$ ，方差为 $\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$ 。

要得到 $\boldsymbol{\beta}$ 的边缘后验分布(marginal posterior)，需将联合后验分布 $p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X})$ 中的 σ^2 积分积掉：

$$\begin{aligned} p(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) &\propto \int_0^\infty (1/\sigma^2)^{\frac{n}{2}+1} \exp \left\{ -\frac{1}{2\sigma^2} \left[(n-K)s^2 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \mathbf{X}'\mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right] \right\} d\sigma^2 \\ &\propto \int_\infty^0 z^{\frac{n}{2}+1} \exp \left\{ -\frac{z}{2} \left[(n-K)s^2 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \mathbf{X}'\mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right] \right\} (-z^{-2}) dz \end{aligned}$$

(作积分变换 $z \equiv 1/\sigma^2$, $\sigma^2 = 1/z$, $d\sigma^2 = -(1/z^2)dz$)

$$\propto \int_0^\infty z^{\frac{n}{2}-1} \exp\left\{-\frac{z}{2}\left[(n-K)s^2 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \mathbf{X}' \mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right]\right\} dz$$

此式在形式上非常接近于 Γ 函数的定义式, 即 $\Gamma(c) \equiv \int_0^\infty z^{c-1} e^{-z} dz$ 。

可以证明,

$$\int_0^\infty z^c e^{-az} dz = \Gamma(c+1)/a^{c+1}$$

令 $c = \frac{n}{2} - 1$, $a = -\frac{1}{2}\left[(n-K)s^2 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \mathbf{X}' \mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right]$, 可得

$$p(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) \propto \Gamma(n/2) \left\{ -\frac{1}{2}\left[(n-K)s^2 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \mathbf{X}' \mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right] \right\}^{-n/2}$$

$$\propto \left[(n-K)s^2 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \mathbf{X}'\mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right]^{-n/2} \quad (\text{去掉常数项})$$

$$\propto \left\{ (n-K)s^2 \left[1 + \frac{1}{(n-K)} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \left(s^2 (\mathbf{X}'\mathbf{X})^{-1} \right)^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right] \right\}^{-n/2}$$

(提取因子)

$$\propto \left[1 + \frac{1}{(n-K)} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \left(s^2 (\mathbf{X}'\mathbf{X})^{-1} \right)^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right]^{-\frac{n-K+K}{2}} \quad (\text{去掉常数项})$$

这正是 K 维 t 分布的密度核，其中心在 $\hat{\boldsymbol{\beta}}$ ，自由度为 $(n-K)$ ，而协方差矩阵为 $s^2 (\mathbf{X}'\mathbf{X})^{-1}$ 。这个结果与古典多元回归类似。单个参数 β_j 的后验分布则为一维 t 分布。

最后，要得到 σ^2 的边缘后验分布(marginal posterior)，只要将联合后验分布 $p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X})$ 中的 $\boldsymbol{\beta}$ 积分积掉即可。回到联合后验分布：

$$p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}) \propto \left(1/\sigma^2\right)^{K/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \left(\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right\} \\ \times \left(1/\sigma^2\right)^{\frac{n-K}{2}+1} \exp\left\{-(n-K)s^2/2\sigma^2\right\}$$

其中，第二部分不含 $\boldsymbol{\beta}$ ，而第一部分为 K 维正态分布的密度核，积分为1，故

$$p(\sigma^2 | \mathbf{y}, \mathbf{X}) \propto \left(1/\sigma^2\right)^{\frac{n-K}{2}+1} \exp\left\{-(n-K)s^2/2\sigma^2\right\}$$

这是“ Γ 分布的平方根之倒数的密度核”(kernel of an inverted

square-root gamma density)。

总之，使用无信息先验分布得到的结果与古典多元回归相类似。

31.7 马尔可夫链蒙特卡罗法

古典学派的核心为点估计。在贝叶斯估计中，则对应于后验均值的计算：

$$E(\boldsymbol{\theta} | \mathbf{y}) = \int \boldsymbol{\theta} p(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta}$$

其中， $p(\boldsymbol{\theta} | \mathbf{y})$ 为后验密度。

但 $p(\boldsymbol{\theta} | \mathbf{y})$ 通常并无解析表达式(可能包括复杂的积分)。

常使用蒙特卡罗积分方法(参见第 19 章)。此处的一个技术难题是，如果 $p(\boldsymbol{\theta} | \mathbf{y})$ 无解析式，如何从后验分布 $p(\boldsymbol{\theta} | \mathbf{y})$ 中获得随机样本？

“吉布斯抽样法” (Gibbs sampler) 的基本思想是，通过更简单的条件分布来抽样。

比如，联合后验分布 $p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X})$ 的表达式较复杂，但条件后验分布 $p(\boldsymbol{\beta} | \sigma^2, \mathbf{y}, \mathbf{X})$ 较简单(为 K 维正态)。

一般地，考虑从二维联合分布 $f(x, y)$ 进行随机抽样，但由于 $f(x, y)$ 的表达式太复杂而无法直接抽样。

另一方面，假设条件分布 $f(x | y)$ 与 $f(y | x)$ 相对简单，可从中进行

一维随机抽样。

可使用以下迭代法得到联合分布 $f(x, y)$ 的随机样本。

吉布斯抽样法的步骤：

- (1) 选择 x_0 (只要在条件分布 $f(x|y)$ 的取值范围即可);
- (2) 从条件分布 $f(y|x_0)$ 中随机抽取 y_0 ;
- (3) 从条件分布 $f(x|y_0)$ 中随机抽取 x_1 ;
- (4) 从条件分布 $f(y|x_1)$ 中随机抽取 y_1 ;
- (5) 从条件分布 $f(x|y_{t-1})$ 中随机抽取 x_t ;
- (6) 从条件分布 $f(y|x_t)$ 中随机抽取 y_t ;

重复以上步骤(5)与(6)几千次之后, 即可得到联合分布 $f(x, y)$ 的随机抽样(通常去掉前面几千个抽样值, 以避免初始值的影响, 称为“burn in”)。

吉布斯抽样的理论基础在于, 当 $t \rightarrow \infty$ 时, (x_t, y_t) 的极限分布(limiting distribution)就是 $f(x, y)$ 。

使用类似方法, 可以得到更高维分布(比如 $f(x, y, z)$)的随机样本。

吉布斯抽样并非真正的随机抽样, 因为每次抽样都是上次抽样的函数, 只不过上次抽样也是随机的。

吉布斯抽样所得到的序列其实是马尔可夫链。

定义：考虑时间为离散的随机过程 $\{x_t\}_0^\infty$ 。如果对于任何 t ，都有 $P(x_{t+1} \leq x \mid x_t, x_{t-1}, \dots, x_0) = P(x_{t+1} \leq x \mid x_t)$ ，则称 $\{x_t\}$ 为“马尔可夫链”(Markov chain)。

在给定整个历史的情况下， x_{t+1} 的条件分布仅与 x_t 有关。

马尔可夫链的未来状态(x_{t+1})仅与现在状态(x_t)有关，而与历史状态(x_{t-1}, \dots, x_0)无关。

由于吉布斯抽样法使用了蒙特卡罗抽样，而得到的序列又是马尔可夫链，故也称为“马尔可夫链蒙特卡罗法”(Markov Chain Monte Carlo, 简记 MCMC)，在贝叶斯分析中应用广泛。