第 8 讲：贝叶斯模型选择

张伟平

目录

1.1 贝叶斯因子

- 模型选择是统计建模里常见的问题之一。变量选择是模型选择的一种形式，比如线性回归模型中研究协变量对响应变量的影响，亦即选择对预测响应变量比较重要的协变量。
- 一般的模型选择问题是要在给定的样本下，在一类候选模型中依照某个准则选择最优的模型。
- 对模型的选择和评价依赖于抽样分布的结构、模型参数的先验分布指定等。已经有一些方法被提出来处理这类问题，本讲我们介绍模型选择和评价中常用的 Bayes 因子、BIC 准则，PBIC 准则和 DIC 准则等。

假设检验回顾

首先我们从模型选择的角度来回顾假设检验问题的 Bayes 推断方法。假设总体 $X \sim f(x|\theta)$，其中 θ 为一未知参数且 $\theta \in \Theta$，而我们感兴趣的假设 $H_0 : \theta \in \Theta_0 \leftrightarrow H_1 : \theta \in \Theta_1$ 等价于比较两个模型

$$M_0 : X \text{ 有密度 } f(x|\theta), \text{ 其中 } \theta \in \Theta_0;$$

$$M_1 : X \text{ 有密度 } f(x|\theta), \text{ 其中 } \theta \in \Theta_1;$$

其中 $\Theta_0 = \Theta - \Theta_1$。令 $g_i(\theta)$ 分别表示给定真实模型为 M_i 下 θ 的先验密度 $(i = 0, 1)$。则当有了样本 $\mathbf{X} = (X_1, \ldots, X_n) = \mathbf{x}$ 后，我们可以使用 Bayes 因子来比较 M_0 和 M_1：

$$BF_{01} = \frac{P(\Theta_0|\mathbf{x})}{P(\Theta_1|\mathbf{x})} \Big/ \frac{\pi_0}{1-\pi_0} = \frac{m_0(\mathbf{x})}{m_1(\mathbf{x})},$$

其中 $\pi_0 = P^\pi(M_0) = P^\pi(\Theta_0), P^\pi(M_1) = P^\pi(\Theta_1) = 1 - \pi_0$，

$$m_i(\mathbf{x}) = \int_{\Theta_i} f(\mathbf{x}|\theta) g_i(\theta) d\theta, \ i = 0, 1.$$

因此

$$P(M_0|\mathbf{x}) = \left\{1 + \frac{1-\pi_0}{\pi_0} BF_{01}^{-1}(\mathbf{x})\right\}^{-1}.$$

从而，如果先验密度 g_0, g_1 可以被指定，则可以仅仅使用 Bayes 因子 BF_{01} 进行模型选择。

进一步，如果 π_0 可以被指定，则可以计算得到模型 M_0 和 M_1 的后验机会比，因而也可以使用后验机会比进行模型选择。

但是 Bayes 因子或者后验机会比未必总是可以容易计算的，即便是先验密度完全指定时候也可能是得不出积分值。此时，可以使用 BIC 来近似 Bayes 因子，我们将在后面详细讨论。当先验分布也难以指定时，Bayes 因子的计算就更加困难了。

1.1.1 正常先验下的贝叶斯因子

- 假设候选模型 $M_1, \ldots, M_r$, 在每个模型 M_k 下抽样密度为 $f_k(x|\theta_k)$, 其中 $\theta_k \in \Theta_k \subset \mathcal{R}^p$ 为未知的 p 维参数向量. 我们要从中选择一个最佳模型.

- 记 $\pi_k(\theta_k)$ 表示在模型 M_k 下参数 θ_k 的先验密度, 则在样本 $\mathbf{X}_n = (X_1, \ldots, X_n) = \mathbf{x}_n$ 下, 模型 M_k 的后验密度为

$$P(M_k|\mathbf{x}_n) = \frac{P(M_k)\int f_k(\mathbf{x}_n|\theta_k)\pi_k(\theta_k)d\theta_k}{\sum_{k=1}^{r} P(M_k)\int f_k(\mathbf{x}_n|\theta_k)\pi_k(\theta_k)d\theta_k},$$

其中 $f_k(\mathbf{x}_n|\theta_k)$ 为样本 $\mathbf{X}_n$ 在模型 M_k 下的密度 (似然函数), $P(M_k)$ 为模型 M_k 的先验概率。

后验概率最大原则

- 后验模型概率 $P(M_1|\mathbf{x}_n), P(M_2|\mathbf{x}_n), \ldots, P(M_r|\mathbf{x}_n)$ 即为模型选择中我们感兴趣的量。
- 这等价于最大化

$$P(M_k) \int f_k(\mathbf{x}_n|\theta_k)\pi_k(\theta_k)d\theta_k$$

其中积分

$$P(\mathbf{x}_n|M_k) = \int_{\Theta_k} f_k(\mathbf{x}_n|\theta_k)\pi_k(\theta_k)d\theta_k$$

为样本 $\mathbf{X}_n$ 在模型 M_k 下的边际概率密度（边际似然），表示了指定的先验分布对样本的拟合程度。

- 对模型的先验概率来说，常用的一种指定为均匀分布

$$P(M_k) = \frac{1}{r}, \quad k = 1, \ldots, r.$$

显然这个先验是无信息先验，表示我们对所有的候选模型一样偏好。在此先验下，上式与边际似然成正比，而且后验概率为

$$P(M_k|\mathbf{x}_n) = \frac{\int f_k(\mathbf{x}_n|\theta_k)\pi_k(\theta_k)d\theta_k}{\sum_{k=1}^{r}\int f_k(\mathbf{x}_n|\theta_k)\pi_k(\theta_k)d\theta_k}.$$

- 尽管均匀先验应用起来很方便，但有时候仍然偏好非均匀先验。比如对线性回归模型 $y = \sum_{i=1}^{p}\beta_i x_i + \epsilon$，我们或许希望对简单的模型赋予更多的先验概率，为此, Denison et al. (1998) 使用 Poisson 分布来作为先验分布

$$P(M_k) \propto \lambda^{p_k} e^{-\lambda},$$

其中 p_k 为模型 M_k 下的解释变量个数，而 λ 衡量了模型中期望的解释变量个数。类似的另外一种先验分布 (如 Smith and

Kohn 1996) 为

$$P(M_k) \propto \prod_{j=1}^{p} \pi_j^{r_j} (1-\pi_j)^{1-r_j},$$

其中 π_j 表示解释变量 x_j 被包含在模型中的概率，$r_j = 1$ 表示 x_j 被包含进模型，而 $r_j = 0$ 表示 x_j 没有被包含进模型。

- 因此模型 M_k 和 M_j 的 Bayes 因子为

$$\begin{aligned} BF_{kj} &= \frac{P(M_k|\mathbf{x}_n)}{P(M_j|\mathbf{x}_n)} \Big/ \frac{P(M_k)}{P(M_j)} = \frac{P(\mathbf{x}_n|M_k)}{P(\mathbf{x}_n|M_j)} \\ &= \frac{\int_{\Theta_k} f_k(\mathbf{x}_n|\theta_k)\pi_k(\theta_k)d\theta_k}{\int_{\Theta_j} f_j(\mathbf{x}_n|\theta_j)\pi_j(\theta_j)d\theta_j}. \end{aligned}$$

从而模型 M_k 的后验概率为

$$P(M_k|\mathbf{x}_n) = \left[\sum_{j=1}^{r} \frac{P(M_j)}{P(M_k)} BF_{jk}\right]^{-1}.$$

因此可以根据模型的后验概率或者所有候选模型进行两两比较 Bayes 因子，从中选出最优的模型。

Jeffreys (1961) 建议将 Bayes 因子解释为证据的程度。如下表表示:

Bayes 因子 B_{kj} 的值所表示的模型支持强度解释

Bayes 因子	解释
$B_{kj} < 1$	否定模型 M_k
$1 < B_{kj} < 3$	对模型 M_k 的支持证据微乎其微
$3 < B_{kj} < 10$	较强的证据支持 M_k
$10 < B_{kj} < 30$	强烈的证据支持 M_k
$30 < B_{kj} < 100$	非常强烈的证据支持 M_k
$100 < B_{kj}$	肯定支持 M_k

1.1.2 非正常先验下的贝叶斯因子

在使用 Bayes 因子中最常见的困难是其对先验选择的敏感性，而如果对参数假设不正常先验，则一般会导致 Bayes 因子没有很好的被唯一定义.

- 由非正常先验的定义

$$\pi(\theta) \propto h(\theta), \quad \int h(\theta)d\theta = \infty.$$

因此对任意正的常数 C，我们可以使用 $q(\theta) = C\pi(\theta)$ 来作为先验，从而后验密度为

$$\pi(\theta|\mathbf{x}_n) = \frac{f(\mathbf{x}_n|\theta)q(\theta)}{\int f(\mathbf{x}_n|\theta)q(\theta)d\theta} = \frac{f(\mathbf{x}_n|\theta)\pi(\theta)}{\int f(\mathbf{x}_n|\theta)\pi(\theta)d\theta}.$$

显然只要 $\int f(\mathbf{x}_n|\theta)\pi(\theta)d\theta$ 存在非零，则后验密度就是正常的密度。

- 但是使用先验 $q_k(\theta_k)$ 的 $q_j(\theta_j)$ 的 Bayes 因子为

$$BF_{kj} = \frac{\int f_k(\mathbf{x}_n|\theta_k)\pi_k(\theta_k)d\theta_k}{\int f_j(\mathbf{x}_n|\theta_j)\pi_j(\theta_j)d\theta_j} \times \frac{C_k}{C_j}.$$

  可以看出，此时 Bayes 因子没有被很好的唯一定义，因为其依赖于任意的常数 C_k/C_j。(而对正常先验来说，C_k, C_j 是唯一有限的，故 C_k/C_j 唯一，因此 Bayes 因子是唯一的)

对这种由先验选择所带来的 Bayes 因子不唯一性，已经有一些方法解决。

潜在贝叶斯因子

- 由于不正常先验的不确定性导致 Bayes 因子的不唯一性，而且注意到不正常先验下的后验分布可以是正常的分布，因此一种自然的想法就是使用部分样本作为“训练”样本，把先验分布“估计”出来 (得到了基于部分样本的后验分布)，再使用剩余的样本来进行模型比较。

- 记样本 $\mathbf{X}_n = (X_1, \ldots, X_n)$，$X_{n(l)}$ 为 $\mathbf{X}_n$ 的子集，记 $X_{-n(l)}$ 表示 $\mathbf{X}_n$ 除去 $X_{n(l)}$ 外剩余的样本。则在先验 π 下

$$\pi(\theta|X_{n(l)}) = \frac{f(X_{n(l)}|\theta)\pi(\theta)}{\int f(X_{n(l)}|\theta)\pi(\theta)d\theta},$$

  其中子集 $X_{n(l)}$ 的选择要使的此后验分布为正常分布。

- 然后使用 $\pi(\theta|X_{n(l)})$ 作为先验分布，结合剩余的样本 $X_{-n(l)}$，计算的 Bayes 因子即称为在给定样本 $X_{n(l)}$ 时的**潜在 Bayes 因**

子(Intrinsic Bayes factor, IBF) 或者部分 Bayes 因子 (因其仅使用了部分样本)，即模型 M_k 和 M_j 在给定样本 $X_{n(l)}$ 下的 Bayes 因子为

$$
\begin{aligned}
IBF_{kj}\ (n(l)) &= \frac{\int f_k(X_{-n(l)}|X_{n(l)},\theta_k)\pi_k(\theta_k|X_{n(l)})d\theta_k}{\int f_j(X_{-n(l)}|X_{n(l)},\theta_j)\pi_j(\theta_j|X_{n(l)})d\theta_j} \\
&= \frac{\int f_k(\mathbf{X}_n|\theta)\pi_k(\theta_k)d\theta_k}{\int f_j(\mathbf{X}_n|\theta_j)\pi_j(\theta_j)d\theta_j} \times \frac{\int f_j(X_{-n(l)}|\theta)\pi_j(\theta_j)d\theta_j}{\int f_k(X_{-n(l)}|\theta_k)\pi_k(\theta_k)d\theta_k} \\
&= BF_{kj} \times BF_{jk}(n(l)).
\end{aligned}
$$

显然此量是唯一的，不确定项 C_k/C_j 在上式中被消去。

- 为考虑样本选择问题, Berger & Pericchi 提出了算术潜在贝叶斯因子和几何潜在贝叶斯因子.

分数贝叶斯因子

O'Hagan (1995) 为了避免潜在贝叶斯因子的数据子集选择问题，从另一角度提出了分数 Bayes 因子 (Fractional Bayes factor, FBF)。

- 记观测的样本 $\mathbf{x}_n = (\mathbf{z}_{n-m}, \mathbf{y}_m)$，其中 $\mathbf{y}$ 作为“训练”样本，则由 IBF_{kj} 定义知部分 Bayes 因子为

$$PBF_{kj}(\mathbf{z}|\mathbf{y}) = \frac{\int \pi_k(\theta_k|\mathbf{y}) f_k(\mathbf{z}|\theta, \mathbf{y}) d\theta_k}{\int \pi_j(\theta_j|\mathbf{y}) f_j(\mathbf{z}|\theta, \mathbf{y}) d\theta_j}.$$

- 注意到

$$q_i(\mathbf{z}|\mathbf{y}) = \int \pi_i(\theta_i|\mathbf{y}) f_i(\mathbf{z}|\theta, \mathbf{y}) d\theta_i = \frac{\int \pi_i(\theta_i) f_i(\mathbf{x}|\theta) d\theta_i}{\int \pi_i(\theta_i) f_i(\mathbf{y}|\theta) d\theta_i}$$

而当 n 和 m 都很大时，在简单样本情况下似然 $f(\mathbf{y}|\theta)$ 趋于 $[f(\mathbf{x}|\theta)]^b$，其中 $b = m/n$，因此定义

$$FBF_{kj}^b = \frac{q_k(\mathbf{x}, b)}{q_j(\mathbf{x}, b)},$$

称为**分数 Bayes 因子**, 其中

$$q_i(\mathbf{x}, b) = \frac{\int \pi_i(\theta_i) f_i(\mathbf{x}|\theta_i) d\theta_i}{\int \pi_i(\theta_i) [f_i(\mathbf{x}|\theta_i)]^b d\theta_i}, i = k, j.$$

可以看出,

$$\begin{aligned} FBF_{kj}^b &= \frac{\int \pi_k(\theta_k) f_k(\mathbf{x}|\theta_k) d\theta_k}{\int \pi_j(\theta_j) f_j(\mathbf{x}|\theta_j) d\theta_j} \times \frac{\int \pi_j(\theta_j) [f_j(\mathbf{x}|\theta_j)]^b d\theta_i}{\int \pi_k(\theta_k) [f_i(\mathbf{x}|\theta_k)]^b d\theta_k} \\ &= BF_{kj} \times \frac{\int \pi_j(\theta_j) [f_j(\mathbf{x}|\theta_j)]^b d\theta_i}{\int \pi_k(\theta_k) [f_i(\mathbf{x}|\theta_k)]^b d\theta_k}, \end{aligned}$$

因此不受不确定量 C_k/C_j 的影响。

后验贝叶斯因子

- Aitkin (1991) 提出后验 Bayes 因子 (Posterior Bayes factor) 来克服不正常先验情形下 Bayes 因子的定义缺点，他称

$$PBF_{kj} = \frac{\bar{L}_k}{\bar{L}_j} = \frac{\int f_k(\mathbf{x}_n|\theta_k)\pi_k(\theta_k|\mathbf{x}_n)d\theta_k}{\int f_j(\mathbf{x}_n|\theta_j)\pi_j(\theta_j|\mathbf{x}_n)d\theta_j}, \tag{1.1}$$

其中 $\pi_i(\theta_i|\mathbf{x}_n) = f_i(\mathbf{x}_n|\theta_i)\pi_i(\theta_i)/\int f_i(\mathbf{x}_n|\theta_i)\pi_i(\theta_i)d\theta_i$ 为 θ_i 在样本 $\mathbf{x}_n$ 下的后验密度。显然，后验 Bayes 因子 PBF 为不同模型下似然函数的后验均值之比。Aitkin (1991) 指出 PBF 和 Bayes 因子的使用类似，PBF_{12} 的值小于 1/20、1/100 和 1/1000 分别表示有强、非常强和极强的证据来否定模型 M_1 而支持模型 M_2。

基于 CV 的拟贝叶斯因子

- 在预测问题中，使用交叉验证方法来检验模型的预测能力是自然的。Gelfand et al. (1992) 提出使用交叉验证预测密度

$$CVPD = \prod_{i=1}^{n} \int f(x_i|\theta)\pi(\theta|x_{-i})d\theta,$$

其中 X_{-i} 表示除去 X_i 后剩余的所有样本。从而

$$PSBF_{kj} = \frac{CVPD_k}{CVPD_j}, \tag{1.2}$$

称为拟 Bayes 因子 (pseudo-Bayes factor, PSBF)。显然，$f(x_i|x_{-i}) = \int f(x_i|\theta)\pi(\theta|x_{-i})d\theta$ 为预测密度，Geisser and Eddy (1979) 提出使用 $CVPD$ 来替代 $f(\mathbf{x})$。交叉验证方法的优点是适用于各种实际情形，当然，当样本量比较大时计算时间也很可观。

1.2 RJMCMC

在一些情况下，特别是模型选择问题，MCMC 过程需要在两个不同维数的参数空间之间移动。前述标准的 MCMC 算法就不适用了。Green（1995）提出的可逆跳算法是标准 MCMC 的推广，称为可逆跳转马尔可夫链蒙特卡洛（Reversible-Jump MCMC）方法。

- 假定我们有模型集合 $k, k \in \mathcal{K}$，$\mathcal{K}$ 为一可数集，且模型 k 有连续的参数空间 $\Theta_k \in \mathcal{R}^{n_k}$，不同模型的参数维数可能是不相同的。假定模型的先验分布为

$$P(k) = p_k, \quad k \in \mathcal{K}, \qquad \sum_{k \in \mathcal{K}} p_k = 1.$$

而对每个模型 k，参数 θ_k 的先验分布为 $\pi^{prior}(\theta|k)$. 记样本为 y，假设 $f(y|k, \theta_k)$ 为在模型 k 下的似然函数，则 (k, θ_k) 的后

验分布为

$$\pi(k,\theta_k|y)=\frac{p_k\pi^{prior}(\theta_k|k)f(y|k,\theta_k)}{\sum_{j\in\mathcal{K}}p_j\int_{\Theta_j}\pi^{prior}(\theta_j|j)f(y|j,\theta_j)d\theta_j} \tag{1.3}$$

- 在模型选择中经常需要计算模型 k 相对于模型 l 的 Bayes 因子，即

$$\frac{P^\pi(k|y)}{P^\pi(l|y)}\frac{\pi_l}{\pi_k},$$

其中

$$P^\pi(k|y)=\frac{p_k\int_{\Theta_k}\pi^{prior}(\theta_k|k)f(y|k,\theta_k)d\theta_k}{\sum_j p_j\int_{\Theta_j}\pi^{prior}(\theta_j|j)f(y|j,\theta_j)d\theta_j}$$

为模型 k 的后验概率。

一种计算方法就是单独计算每个 $P^\pi(k|y)$，然后选择模型后验概率最大的模型，或者使用后验概率 $P^\pi(k|y)$ 为权进行模型平均。这种

方法对存在大量的可选模型时是不可行的和效率低下的。另外一种方法就是使用所谓的可逆跳转马尔科夫链蒙特卡洛方法。

- 记 $x = (k, \theta)$，则 x 的取值空间为

$$\Theta = \prod_{k \in \mathcal{K}} \{\{k\} \times \Theta_k\},$$

其有分布 (1.3)。因此按照 M-H 算法，我们提出构造一个以 (1.3) 为平稳分布的马氏链。但是和前面的单模型下的算法不同，这里状态空间是由一些不同维数的子空间构成。因此需要运行算法在不同维数的子空间 (不同模型) 之间跳转。

- 按照 M-H 算法的要求，我们需要使用一个提议分布 $q(\cdot|x)$ 在当前状态 x 下，产生一个候选的状态 x'，并以一定的接受概率 $\alpha(x, x') = \min\{1, \frac{\pi(x'|y)q(x|x')}{\pi(x|y)q(x'|x)}\}$ 接受其为下一步状态。此处，$x = (k, \theta_k), x' = (m, \theta_m)$，由于 $q(m, \theta_m|x) = q(m|x)q(\theta_m|x, m)$，

可以更特殊的取 $q(m|x) = q(m|k)$，即仅依赖于当前模型指示变量 k。类似地，$q(\theta_m|x,m) = q(\theta_m|\theta_k)$。因此接受概率

$$\alpha(x,x') = \min\{1, \frac{\pi(m,\theta_m|y)q(k|m)q(\theta_k|\theta_m)}{\pi(k,\theta_k|y)q(m|k)q(\theta_m|\theta_k)}\}.$$

但是由于 θ_m 与 θ_k 的维数差异，选择合适的密度 $q(\theta_m|\theta_k)$ 比较困难。在从 $q(m|k)$ 中产生了 m 后，建立 θ_k 到 θ_m 的转移的一个常用的办法就是选取辅助变量 u 和 v，使得 (θ_k,u) 和 (θ_m,v) 的维数是匹配的:

$$k + dim(u) = m + dim(v).$$

因此，假设我们从某个提议密度 $g(u|\theta_k,m)$ 中产生一个 u，然后通过一一映射 ϕ 建立

$$(\theta_m,v) = \phi(\theta_k,u).$$

由密度变换公式知道 (θ_m, v) 的密度为

$$g(v|\theta_m)q(\theta_m) = \frac{g(u|\theta_k)q(\theta_k)}{|det(J_\phi(\theta_k, u))|},$$

其中 $det(J_\phi(\theta_k, u))$ 为一一映射 ϕ 在 (θ_i, u) 处的 Jacobi 行列式值。于是

$$\frac{q(\theta_k|\theta_m)}{q(\theta_m|\theta_k)} = \frac{q(\theta_k)}{q(\theta_m)} = \frac{g(v|\theta_m)}{g(u|\theta_k)}\Big|det(J_\phi(\theta_k, u))\Big|.$$

所以接受概率

$$\alpha(x, x') = \min\left\{1, \frac{\pi(m, \theta_m)q(k|m)g(v|\theta_m)}{\pi(k, \theta_k)q(m|k)g(u|\theta_k)}\Big|det(J_\phi(\theta_k, u))\Big|\right\}.$$

综上所述，可逆跳转马尔科夫蒙特卡洛算法为

迭代 t: 若 $x_t = (k, \theta_k)$，则

1. 以概率 $q(m|k)$ 选择模型 m;

2. 从提议分布 $g(u|\theta_k, m)$ 中产生 u，令 $(\theta_m, v) = \phi(\theta_k, u)$;

3. 以概率

$$\alpha(x_t, x_{t+1}) = \min\left\{1, \frac{\pi(m, \theta_m|y)q(k|m)g(v|\theta_m)}{\pi(k, \theta_k|y)q(m|k)g(u|\theta_k)}\left|det(J_\phi(\theta_k, u))\right|\right\}$$

令 $x_{t+1} = (m, \theta_m)$，否则 $x_{t+1} = x_t$。

例 5.5.10:

↑Example
↓Example

给定容量为 n 的简单样本 y 时，在参数为 $\lambda > 0$ 的 Poisson 分布下似然函数为

$$L(y|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i}}{y_i!} e^{-\lambda_i y_i}.$$

而在参数 $\lambda > 0, \kappa > 0$ 的负二项分布下似然函数为

$$L(y|\lambda, \kappa) = \prod_{i=1}^{n} \frac{\lambda^{y_i}}{y_i!} \frac{\Gamma(1/\kappa + y_i)}{\Gamma(1/\kappa)(1/\kappa + \lambda)^{y_i}} (1 + \kappa\lambda)^{-1/\kappa}$$

Poisson 分布和负二项分布的均值都为 λ，而负二项分布的方差为 $\lambda(1 + \kappa\lambda)$，因而更适合过度分散的数据。

假设两个模型示性变量的先验分布为 $P(k=1)=P(k=2)=0.5$。参数 $\theta_1=\lambda, \theta_2=(\theta_{21},\theta_{22})=(\lambda,\kappa)$ 的先验分布取为 Gamma 分布，即 $\theta_1|k=1 \sim Gamma(\alpha_\lambda,\beta_\lambda)$ 以及 $\theta_2|k=2=(\theta_{21}|k=2)(\theta_{22}|k=2)$, 其中 $\theta_{21} \sim Gamma(\alpha_\lambda,\beta_\lambda), \theta_{22} \sim Gamma(\alpha_\kappa,\beta_\kappa)$. 于是得到后验分布

$$\pi(k,\theta_k|y) \propto \begin{cases} \frac{1}{2}p(\theta_1|k=1)L(y|\theta_1), & k=1 \\ \frac{1}{2}p(\theta_{21},\theta_{22})|k=2)L(y|\theta_2), & k=2 \end{cases},$$

其中 $p(\theta_1|k=1)$, $p(\theta_{21},\theta_{22}|k=2)=p(\theta_{21}|k=2)p(\theta_{22}|k=2)$ 为相应 Gamma 分布的密度函数。

下面我们构造从模型 1 到模型 2 的合适转移。

- 记 $x=(1,\lambda)$ 为当前链的状态，$x'=(2,\theta)$ 为下一个候选值，其中 $\theta=(\lambda,\kappa)$, 则接受概率

$$\alpha(x,x')=\min\left\{1, \frac{\pi(2,\theta|y)q(1|2)q(\lambda|\theta)}{\pi(1,\lambda|y)q(2|1)q(\theta|\lambda)}\right\}$$

其中 $q(1|2)/q(2|1) = P(k=1)/P(k=2) = 1$. 由于模型 1 中没有和参数 κ 等价的分量，我们这里采用一个独立的方法。具体地，

$$(\lambda, \kappa) = \phi(\lambda, u) = (\lambda, \mu e^u)$$

其中 $u \sim N(0, \sigma^2)$，μ, σ^2 是固定的. 换句话说，就是 λ 在变换中不变，而 κ 是一个对数正态随机变量。因而变换的 Jacobi 行列式值为

$$|J| = \begin{vmatrix} 1 & 0 \\ 0 & \mu e^u \end{vmatrix} = \mu exp(u).$$

所以

$$q(\lambda, \kappa) = q(\lambda, u)/|J|$$

从而从模型 1 到模型 2 的接受概率为 $\min\{1, A_{12}\}$，其中

$$A_{12} = \frac{\pi(2, \theta|y)}{\pi(1, \lambda|y)} \left\{ \frac{1}{\sqrt{2\pi}\sigma} exp[-u^2/(2\sigma^2)] \right\}^{-1} \mu exp(u)$$

- 从模型 2 到模型 1 的转移，令 $(\lambda, u) = \phi'(\lambda, \kappa) = (\lambda, log(\kappa)/\mu)$. 而从模型 2 转移到模型 1 的接受概率为 $\min\{1, A_{21}\}$，其中

$$A_{21} = \frac{\pi(1, \lambda|y)}{\pi(2, \theta|y)} \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{(\log(\kappa/\mu))^2}{2\sigma^2}\right] \frac{1}{\kappa}$$

1.3 贝叶斯模型评价

1.3.1 Bayes 预测信息准则

记 $M_j = \{f(x;\theta_j) : \theta_j \in \Theta_j\}, j = 1, \ldots, r$. $\widehat{\theta}_j$ 为模型 M_j 下参数的最大似然估计. $\widehat{f}_j(x) = f(x;\widehat{\theta}_j)$. 估计 $\hat{f}_j(x)$ 与真实分布 $g(x)$ 之间的 Kullback-Leibler 距离为

$$\begin{aligned} KL\left(g||\widehat{f}_j\right) &= \int g(x) \log \frac{g(x)}{\hat{f}_j(x)} dx \\ &= \int g(x) \log g(x) dx - \int g(x) \log \widehat{f}_j(x) dx \end{aligned}$$

首项与 j 无关, 因此对 $j = 1, \ldots, r$ 最小化 $KL(g||\widehat{f}_j)$ 等价于最大化

$$K_j = \int g(x) \log \widehat{f}_j(x) dx$$

- 上式的一个” 直观” 估计为

$$\bar{K}_j = \frac{1}{n}\sum_{i=1}^{n} \log f\left(x_i; \widehat{\theta}_j\right) = \frac{\ell_j\left(\widehat{\theta}_j\right)}{n}$$

但是这个估计偏倚很大, 因为样本被使用了两次: 一次估计参数 θ_j; 第二次估计积分.

- AIC 准则 Akaike (1974) 注意到 (省略掉 j)

$$\begin{aligned} K &\approx \int g(x)\left(\log f(x;\theta_0) + (\widehat{\theta}-\theta_0)^T s(x,\theta_0)\right. \\ &\quad \left. + \frac{1}{2}(\widehat{\theta}-\theta_0)^T H(x,\theta_0)(\widehat{\theta}-\theta_0)\right) dx \\ &= K_0 - \frac{1}{2n} Z_n^T J Z_n \end{aligned}$$

其中 $s(x,\theta) = \partial \log f(x;\theta)/\partial\theta$, $H(x,\theta) = \partial^2 \log f(x;\theta)/\partial\theta\partial\theta'$, $Z_n = \sqrt{n}(\hat{\theta}-\theta_0)$, $J = -EH(X,\theta_0)$. $K_0 = \int g(x)\log f\left(x;\theta_0\right) dx$.

另一方面,

$$\begin{aligned}\bar{K} &\approx \frac{1}{n}\sum_{i=1}^{n}\Big(\ell(x_i,\theta_0)+(\widehat{\theta}-\theta_0)^T s(x_i,\theta_0)\\ &\quad+\frac{1}{2}(\widehat{\theta}-\theta_0)^T H(x_i,\theta_0)(\widehat{\theta}-\theta_0)\Big)\\ &= K_0+\Lambda_n+\frac{(\widehat{\theta}-\theta_0)^T S_n}{\sqrt{n}}-\frac{1}{2n}Z_n^T J_n Z_n\\ &\approx K_0+\Lambda_n+\frac{Z_n^T S_n}{n}-\frac{1}{2n}Z_n^T J Z_n\end{aligned}$$

其中 $J_n=-\frac{1}{n}\sum_{i=1}^{n}H(x_i,\theta_0)\xrightarrow{P}J$, $\Lambda_n=\frac{1}{n}\sum_{i=1}^{n}(\ell(x_i,\theta_0)-K_0)$, $S_n=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}s(x_i,\theta_0)\to N(0,V)$ 因此利用 $Z_n\approx J^{-1}S_n$ 有

$$E(\bar{K}-K)\approx\mathbb{E}(\Lambda_n)+E\left(\frac{Z_n^T S_n}{n}\right)=0+\frac{trace\left(J^{-1}V\right)}{n}$$

即有

$$K \approx \bar{K} - \frac{\text{trace}\left(J^{-1}V\right)}{n}$$

如果模型是正确的, 则 $J = V$, 故 $trace(J^{-1}V) = d = dim\theta$, 从而

$$K \approx \bar{K} - \frac{d}{n}$$

现在, 定义

$$AIC(j) = -2n\widehat{K}_j = -2\ell_j(\widehat{\theta}_j) + 2d_j$$

对 $AIC(j)$ 进行最小化即可.

- BIC **准则** 由于

$$\begin{aligned} f\left(x_1, \ldots, x_n | M_j\right) &= \int f\left(x_1, \ldots, x_n | M_j, \theta_j\right) \pi_j\left(\theta_j\right) d\theta_j \\ &= \int L\left(\theta_j\right) \pi_j\left(\theta_j\right) d\theta_j \end{aligned}$$

而

$$P\left(M_j|x_1,\ldots,x_n\right) \propto f\left(x_1,\ldots,x_n|M_j\right) p_j$$

其中 $p_j = P(M_j)$. 因此, 我们可以选择 j 最大化

$$\log \int L\left(\theta_j\right) \pi_j\left(\theta_j\right) d\theta_j + \log p_j$$

使用 Laplace 近似有

$$\int L\left(\theta_j\right) \pi_j\left(\theta_j\right) d\theta_j \approx L(\hat{\theta})(2\pi)^d n^{-d/2} det[\Delta(\hat{\theta})]^{-1/2} \pi(\hat{\theta})$$

忽略掉 $n \to \infty$ 时的有界项后有

$$\log \int L\left(\theta_j\right) \pi_j\left(\theta_j\right) d\theta_j + \log p_j \approx \ell_j(\widehat{\theta}_j) - \frac{d_j}{2} \log n$$

因此定义

$$BIC(j) = -2\ell_j(\widehat{\theta}_j) + d_j \log n$$

- BPIC **准则** Ando (2007) 提出了最大化对数似然的后验均值

$$\eta(G) = \int \left\{ \int log f(z|\theta)\pi(\theta|\mathbf{x}_n)d\theta \right\} dG(z)$$

其中 $\mathbf{x}_n$ 为观测到的样本值, G 为真实模型.

但是真实的模型 G 往往是未知的，因此建立 $\eta(G)$ 的一个估计是有必要的。利用经验分布函数，一个自然的估计为

$$\eta(\hat{G}) = \frac{1}{n} \int log f(\mathbf{x}_n|\theta)\pi(\theta|\mathbf{x}_n)d\theta.$$

$\eta(\hat{G})$ 相对于 $\eta(G)$ 来说一般是有偏的，这是因为估计模型参数和计算期望对数似然的后验均值使用了相同的样本数据。因而

需要考虑修正这个偏差。偏差定义为

$$\begin{aligned} b(G) &= \int [\eta(\hat{G}) - \eta(G)] dG(\mathbf{x}_n) \\ &= \int \left[\frac{1}{n} \int log f(\mathbf{x}_n|\theta) \pi(\theta|\mathbf{x}_n) d\theta \right. \\ &\quad \left. - \int \int log f(z|\theta) \pi(\theta|\mathbf{x}_n) d\theta dG(z) \right] dG(\mathbf{x}_n), \end{aligned}$$

其中 $G(\mathbf{x}_n)$ 为样本 $\mathbf{X}_n$ 的密度。

记偏差 $b(G)$ 的估计为 $\hat{b}(G)$，则 $\eta(G)$ 的一个偏差修正的估计为

$$\eta(G) \leftarrow \frac{1}{n} \int log f(\mathbf{x}_n|\theta) \pi(\theta|\mathbf{x}_n) d\theta - \hat{b}(G).$$

此估计量常常表示为

$$IC = -2 \int log f(\mathbf{x}_n|\theta) \pi(\theta|\mathbf{x}_n) d\theta + 2n\hat{b}(G).$$

上式右边第一项度量了模型的拟合程度，第二项则是对模型复杂程度的一个惩罚。

假设参数模型 $f(x|\theta)$ 包含了真实的模型 $g(x) = f(x;\theta_0)$，以及 $log\pi(\theta) = O_p(1)$，Anho (2007) 证明了渐近偏差为 $\hat{b} = p/n$，因此提出 Bayes 预测信息准则 (the Bayesian predictive information criterion, BPIC)：

$$BPIC = -2\int logf(\mathbf{x}_n|\theta)\pi(\theta|\mathbf{x}_n)d\theta + 2p, \qquad (1.4)$$

其中 p 为模型中的参数个数。最优模型可以通过最小化 BPIC 得到。在实际应用中，对数似然的后验均值往往没有解析表达，此时可以使用蒙特卡洛逼近：

$$\int logf(\mathbf{x}_n|\theta)\pi(\theta|\mathbf{x}_n)d\theta \approx \frac{1}{L}\sum_{j=1}^{L} logf(\mathbf{x}_n|\theta^{(j)}),$$

其中 $\theta^{(1)}, \ldots, \theta^{(L)}$ 为从后验分布 $\pi(\theta|\mathbf{x}_n)$ 中抽取的后验样本。

- DIC **偏差信息准则** 偏差 (Deviance) 的经典定义

$$D(\theta) = -2log f(\mathbf{x}_n|\theta) + 2log f_S(\mathbf{x}_n),$$

其中 $f_S(\mathbf{x}_n)$ 一个仅依赖于样本的标准化项，可以视为饱和模型的似然函数最大值，在模型比较中此项不起作用，因此常取 $f_S(\mathbf{x}_n) = 1$。Spiegelhalter et al. (2002) 指出对数似然的后验均值，$\bar{D} = E[D(\theta)|\mathbf{x}_n]$，可以作为模型拟合程度的一个 Bayes 度量。一个模型拟合数据的程度越高，则似然应越大，相应的 DIC 就越小，$\bar{D}$ 越大，表明模型拟合数据的程度越差。

通过定义有效参数个数来刻画模型的复杂程度:

$$p_D = \bar{D} - D(\bar{\theta}_n) = 2log f(\mathbf{x}_n|\bar{\theta}_n) - 2\int log f(\mathbf{x}_n|\theta)\pi(\theta|\mathbf{x}_n)d\theta,$$

其中 $\bar{\theta}_n$ 为后验均值。p_D 越大，则模型拟合数据越容易。Spiegelhalter et al. (2002) 定义偏差信息准则 (Deviance information

criterion, DIC) 为

$$DIC = p_D + \bar{D} = D(\bar{\theta}_n) + 2p_D. \tag{1.5}$$

DIC 可以视为是 $AIC = D(\hat{\theta}) + 2p$ 的推广，其中 $\hat{\theta}$ 为极大似然估计。对非分层模型而言，$p \approx p_D, \hat{\theta} \approx \bar{\theta}_n$，从而 $DIC \approx AIC$。

DIC 准则和 Bayes 因子、BIC 准则在形式上和目的上均有所不同。BIC 试图来识别真实的模型，而 DIC 并没有假设“真实模型”，其目的在于考察短期的预测能力。BIC 要求指定一些参数，而 DIC 估计有效参数个数。BIC 提供了一种方法来进行模型平均，而 DIC 则没有。DIC 值默认包含在 WinBUGS 软件分析结果中。

1.4 模型平均

假设我们要基于数据 $D = \{Y_1, \ldots, Y_n\}$ 预测一个新观测 Y, 有不同的候选模型 $\mathcal{M}_j$, 各模型先验概率相同, 则贝叶斯模型平均方法使用下式进行预测

$$p(y|D) = \sum_j p\left(y|D, \mathcal{M}_j\right) P\left(\mathcal{M}_j|D\right)$$

其中

$$P\left(\mathcal{M}_j|D\right) = \frac{\int L\left(\theta_j\right) \pi_j\left(\theta_j\right) d\theta_j}{\sum_s \int L\left(\theta_s\right) \pi_s\left(\theta_s\right) d\theta_s} \approx \frac{e^{\mathrm{BIC}_j}}{\sum_s e^{\mathrm{BIC}_s}}$$

记 $\hat{Y}_j = E[Y|D, M_j]$, 则后验均值和方差为

$$E[Y \mid D] = \sum_j \hat{Y}_j \operatorname{pr}\left(M_j \mid D\right)$$

$$\operatorname{Var}[Y \mid D] = \sum_j \left(\operatorname{Var}\left[Y \mid D, M_j\right] + \hat{Y}_j^2\right) \operatorname{pr}\left(M_j \mid D\right) - E[Y \mid D]^2$$

↑Example

考虑线性模型

$$y = \alpha_\gamma + X_\gamma \beta_\gamma + \varepsilon \quad \varepsilon \sim N\left(0, \sigma^2 I\right)$$

其中 $X_\gamma \in \{X\}$, 假设 X 包含了 K 个可能有用的协变量, 贝叶斯模型平均估计所有 2^K 个变量组合, 然后使用后验模型概率进行平均.

↓Example

第 γ 个模型的后验概率为

$$p\left(M_\gamma | y, X\right) = \frac{p\left(y | M_\gamma, X\right) p\left(M_\gamma\right)}{\sum_{s=1}^{2^K} P\left(y | M_s, X\right) p\left(M_s\right)}$$

从而, 对任何量 θ(e.g. 回归系数 β) 有

$$p(\theta | y, X) = \sum_{\gamma=1}^{2^K} p\left(\theta | M_\gamma, y, X\right) p\left(M_\gamma | X, y\right)$$

在正态线性模型框架下, β_γ 的先验分布常取为 Zellner's g:

$$\beta_\gamma | g \sim N\left(0, \psi^{-1}\left(\frac{1}{g} X_\gamma' X_\gamma\right)^{-1}\right)$$

在模型 M_γ 下的后验分布,

$$\beta_\gamma \mid \psi, X_\gamma, y \sim \mathrm{N}\left[q\hat{\beta}_\gamma + (1-q)\beta_0, \frac{q}{\psi}\left(X_\gamma^\top X_\gamma\right)^{-1}\right]$$

其中 $q = g/(1+g)$, $\beta_0 = 0$, $\hat{\beta}_\gamma = \left(X_\gamma^\top X_\gamma\right)^{-1} X_\gamma^\top y$. 此时

$$p\left(y | M_\gamma, X, g\right) \propto (y-\bar{y})'(y-\bar{y})^{-\frac{N-1}{2}} (1+g)^{-\frac{k\gamma}{2}} \left(1 - \frac{g}{1+g}\right)^{-\frac{N-1}{2}}$$

对超参数 g 的选择常默认选 unit information prior (UIP), 即 $g = N$. (详细讨论见阅读材料.)