学习笔记：贝叶斯因子（Bayes Factors）[1]

吴金龙（Jinlong Wu）
School of Mathematical Sciences
Peking University

March, 2010

Contents

1 Introduction

贝叶斯方法由 Jeffreys (1935, 1961) 引入到*假设检验（hypothesis testing）*中。

Hypothesis testing 与 model selection 的联系与区别：

- Hypothesis testing 一般用来验证是否该拒绝零*假设（null hypothesis）* H_0，所以其中的假设 H_0 和 H_1 是不对称的，而 Bayes Factors 中的各个模型之间是对称的。

相比于 non-Bayesian methods，贝叶斯方法可以解决以下困难：

- 当接受零假设时，零假设到底是以多大的优势打败对手。
- 可以很简单的包括其他信息。
- 非嵌套模型的比较和嵌套模型一样方便。
- 计入模型选择中的不确定性。

2 Applications

Application 1: Escherichia coli Mutagenesis
在两种条件下事件发生的概率 H_0：相同？ H_1：不同？

Application 2: The Hot Hand
每次比赛中投中球满足二项分布，那么不同的比赛中投中球的概率（二项分布的一个参数） H_0：相同？ H_1：不同？

Application 3: Ozone Exceedances
臭氧超标率随着时间是 H_0：不变？ H_1：逐渐变化？ H_2：突然变化？

Application 4: Educational Transitions

社会阶层背景、能力以及学校种类是如何影响一个人的成就的？连接函数是 H_1 ： log-log ？还是 H_2 ： logit ？这个应用中主要的问题是模型不确定性。我们可以获得感兴趣量在单个模型条件下的后验分布(30)，也可以获得它在组合模型条件下的后验分布(31)。

Application 5: Human Working Memory Failure in Computer-Based Tasks

3 Bayes Factors

3.1 Definition

- **D** ： 已知的数据
- H_k ： 第 k 个模型
- odds：$\text{odds} = \text{probability}/(1 - \text{probability})$
- 嵌套模型： H_0 : $(\psi = \psi_0, \beta)$ ； H_1 : (ψ, β) 。

假设有两个候选模型 H_1 和 H_2 ，则：

$$\frac{p(H_1|\mathbf{D})}{p(H_2|\mathbf{D})} = \frac{p(\mathbf{D}|H_1)}{p(\mathbf{D}|H_2)} \cdot \frac{p(H_1)}{p(H_2)} \quad , \tag{1}$$

也即

$$\text{posterior odds} = \text{Bayes factor} \times \text{prior odds} \quad 。 \tag{2}$$

定义

$$B_{12} \equiv \frac{p(\mathbf{D}|H_1)}{p(\mathbf{D}|H_2)} \tag{3}$$

为*贝叶斯因子（Bayes factor）*。所以

$$\text{Bayes factor} = \frac{\text{posterior odds}}{\text{prior odds}} \quad 。 \tag{4}$$

一般情况我们可以取 $H_1 = H_2 = 1/2$ ，此时模型的后验概率比由贝叶斯因子唯一决定。贝叶斯因子中所涉及的似然 $p(\mathbf{D}|H_k)$ 由下面的积分公式获得：

$$p(\mathbf{D}|H_k) = \int p(\mathbf{D}|\theta_k, H_k)\pi(\theta_k|H_k)d\theta_k \quad , \tag{5}$$

其中 θ_k 为模型 H_k 中的参数， $\pi(\theta_k|H_k)$ 为它的先验密度。记 θ_k 的维数为 d_k 。

先验分布 $\pi(\theta_k|H_k)$ 是很有必要的，但它的影响也是两方面的。一方面它可以包括模型参数的其他信息，另一方面是它在没有很多信息时又往往很难决定取何种形式。

3.2 Interpretation

Jeffreys (1961) 给了一个贝叶斯因子的粗略尺度，见表 1。

B_{10}	Evidence **AGAINST** H_0
$(1, 3.2]$	Not worth more than a bare mention
$(3.2, 10]$	Substantial
$(10, 100]$	Strong
$(100, \infty)$	Decisive

Table 1: Jeffreys' scale of evidence for Bayes factors

我们可以把数据在模型 H_k 下的对数边缘概率 $\log p(\mathbf{D}|H_k)$ 看成模型 H_k 获得的*预测分数（predictive score）*。那么 $\log B_{10}$ 就是模型 H_1 与 H_0 的预测分数之差。所以贝叶斯因子 B_{10} 可以视为测量模型 H_1 与 H_0 在预测数据 $\mathbf{D}$ 时的相对成功。很多应用中对数贝叶斯因子也被称为 *weight of evidence*。

4 Calculating Bayes Factors

为了计算贝叶斯因子，我们需要计算如下的积分：

$$I = \int p(\mathbf{D}|\theta, H)\pi(\theta|H)d\theta \quad 。 \tag{6}$$

当样本数比较大时，被积函数变得非常集中在它的最大值附近（highly peaked around its maximum）。

4.1 Asymptotic Approximation

4.1.1 Laplace's Method

当后验密度 $p(\theta|\mathbf{D}, H) \propto p(\mathbf{D}|\theta, H)\pi(\theta|H)$ 非常集中于它的最大值（也即 posterior mode）$\tilde{\theta}$ 附近时，我们可以使用 Laplace 方法获得(6)的很好近似。如果似然函数 $p(\mathbf{D}|\theta, H)$ 非常集中于它的最大值 $\hat{\theta}$ 附近，那么往往 $p(\theta|\mathbf{D}, H)$ 也非常集中于 $\tilde{\theta}$ 附近。这种条件在大数据量时可以获得满足。

记

$$\tilde{l}(\theta) \equiv \log(p(\mathbf{D}|\theta, H)\pi(\theta|H)) \quad , \tag{7}$$

$\tilde{l}(\theta)$ 在 $\tilde{\theta}$ 点处展开至二次项：

$$\begin{aligned} \tilde{l}(\theta) &\approx \tilde{l}(\tilde{\theta}) + (\theta - \tilde{\theta})^T \frac{d\tilde{l}(\tilde{\theta})}{d\theta} + \frac{1}{2}(\theta - \tilde{\theta})^T D^2\tilde{l}(\tilde{\theta})(\theta - \tilde{\theta}) \\ &= \tilde{l}(\tilde{\theta}) + \frac{1}{2}(\theta - \tilde{\theta})^T D^2\tilde{l}(\tilde{\theta})(\theta - \tilde{\theta}) \quad 。 \end{aligned} \tag{8}$$

把上面的展开结果代入(6)，得到 I 的近似值：

$$\begin{aligned}\hat{I} &= \int \exp\left\{\tilde{l}(\tilde{\theta}) + \frac{1}{2}(\theta-\tilde{\theta})^T D^2\tilde{l}(\tilde{\theta})(\theta-\tilde{\theta})\right\} d\theta \\ &= \int \exp\left\{\frac{1}{2}(\theta-\tilde{\theta})^T D^2\tilde{l}(\tilde{\theta})(\theta-\tilde{\theta})\right\} d\theta \cdot e^{\tilde{l}(\tilde{\theta})} \\ &= (2\pi)^{d/2}|\tilde{\Sigma}|^{1/2} \cdot p(\mathbf{D}|\tilde{\theta}, H)\pi(\tilde{\theta}|H) \quad ,\end{aligned} \tag{9}$$

其中 $\tilde{\Sigma} \equiv (-D^2\tilde{l}(\tilde{\theta}))^{-1}$ ，而 d 为 θ 的维数。 $\hat{I}$ 即为 I 的 *Laplace 近似*。在满足某些条件（Kass, Tierney, and Kadane (1990)）时可以得到：

$$I = \hat{I}(1 + O(n^{-1})) \ , \quad \text{当样本数 } n \to \infty \text{ 时} \quad 。 \tag{10}$$

Laplace 近似对于行为较好（well-behaved，也即似然函数与正态分布相差不是太远）的问题一般都可以获得足够好的近似。经验也表明当样本数小于 $5d$ 时它的近似精度令人担忧，而当样本数大于 $20d$ 时它一般都可以近似得比较好。

4.1.2 Variants on Laplace's Method

(9)的一个重要变形如下：

$$\hat{I}_{\text{MLE}} = (2\pi)^{d/2}|\hat{\Sigma}|^{1/2} \cdot p(\mathbf{D}|\hat{\theta}, H)\pi(\hat{\theta}|H) \quad , \tag{11}$$

其中 $\hat{\Sigma}$ 为*观测信息矩阵（observed information matrix）*，即

$$\hat{\Sigma} = -\left.\frac{\partial^2}{\partial\theta^2} l(\theta|\mathbf{D}, H)\right|_{\hat{\theta}} = -\left.\frac{\partial^2}{\partial\theta^2} \log p(\mathbf{D}|\theta, H)\right|_{\hat{\theta}} \quad ; \tag{12}$$

而 $\hat{\theta}$ 为*最大似然估计（MLE）*，即 $p(\mathbf{D}|\theta, H)$ 的最大值点。(11)的相对误差同样是 $O(n^{-1})$ 。

虽然在先验分布包含较多信息时(11)没有(9)近似准确，但它的计算在很多软件中都更容易得多。

一些软件包中会使用*期望信息矩阵（expected information matrix，或 Fisher information matrix）*

$$-\mathbb{E}\left[\left.\frac{\partial^2}{\partial\theta^2}\log p(\mathbf{D}|\theta, H)\right|\theta\right]\Big|_{\hat{\theta}} = -n \cdot \mathbb{E}\left[\left.\frac{\partial^2}{\partial\theta^2}\log p(D_i|\theta, H)\right|\theta\right]\Big|_{\hat{\theta}}$$

代替(11)中的观测信息矩阵 $\hat{\Sigma}$ ，其中的期望是关于 $p(\mathbf{D}|\theta)$ 的。这样产生的近似其渐进相对误差为 $O(n^{-1/2})$ ，但在很多问题中它的精度也够用了。

现在假设我们有嵌套模型假设： H_1 有参数 (β, ψ) ，其对应的先验分布为 $\pi(\beta, \psi|H_1)$ ；而 H_0 有 $\psi = \psi_0$ 以及先验分布 $\pi(\beta|H_0)$ 。使用(11)，我们获得了：

$$2\log B_{10} \approx \Lambda + \log|\hat{\Sigma}_1| - \log|\hat{\Sigma}_0| + \log\pi(\hat{\beta}, \hat{\psi}|H_1) - \log\pi(\hat{\beta}^*|H_0) + (d_1 - d_0)\log(2\pi) \quad , \tag{13}$$

其中 $\Lambda \equiv 2[\log p(\mathbf{D}|(\hat{\beta}, \hat{\psi}), H_1) - \log p(\mathbf{D}|\hat{\beta}, H_0)]$ 为*对数似然比（log-likelihood ratio）*， $\hat{\beta}^*$ 为模型 H_0 下的 MLE，而 $(\hat{\beta}, \hat{\psi})$ 为模型 H_1 下的 MLE 。依据在计算方差矩阵 $\hat{\Sigma}_k$ $(k = 0, 1)$ 时是使用观测信息矩阵还是期望信息矩阵，上面的近似误差分别为 $O(n^{-1})$ 或 $O(n^{-1/2})$ 。

Raftery (1993c) 也建议使用牛顿方法从 $\hat{\theta}$ 开始走一步获得 $\tilde{\theta}$ 的近似值，然后把它代入到(9)以获得更精确的 $2\log B_{10}$ 近似值。

4.1.3 The Schwarz Criterion

另一种避免在(3)中引入先验 $\pi_k(\theta_k|H_k)$ 的方法是使用 *Schwarz criterion*：

$$S = \log p(\mathbf{D}|\hat{\theta}_1, H_1) - \log p(\mathbf{D}|\hat{\theta}_2, H_2) - \frac{1}{2}(d_1 - d_2)\log n \quad , \tag{14}$$

其中 $\hat{\theta}_k$ 为模型 H_k 下的 MLE， d_k 为 $\hat{\theta}_k$ 的维数，而 n 为样本数。

当 $n \to \infty$ 时，有：

$$\frac{S - \log B_{12}}{\log B_{12}} \to 0 \quad 。 \tag{15}$$

所以 S 往往可以视为 $\log B_{12}$ 的近似值。通常我们称 $-2S$ 为 *Bayesian information criterion (BIC)* 。

相对于(9)和(11)，使用 e^S 近似 B_{12} 的相对误差为 $O(1)$，但在满足某些条件时（比较常见的条件）它的相对误差可以降为 $O(n^{-1/2})$ 。在大样本情况下，Schwarz criterion 应当能提供 evidence 的合理指示。

Schwarz criterion 的优点是即使先验分布 $\pi_k(\theta_k|H_k)$ 很难设定时仍可以使用。

4.2 Simple Monte Carlo, Importance Sampling, and Gaussian Quadrature

当不显示写出对模型 H 的依赖时，(6)变为：

$$p(\mathbf{D}) = \int p(\mathbf{D}|\theta)\pi(\theta)d\theta \quad 。 \tag{16}$$

对(16)的 *simple Monte Carlo* 积分估计如下：

$$\hat{p}_1(\mathbf{D}) = \frac{1}{m}\sum_{i=1}^{m} p(\mathbf{D}|\theta^{(i)}) \ , \quad \text{其中} \theta^{(i)} \sim \pi(\theta) \ 。 \tag{17}$$

$\hat{p}_1(\mathbf{D})$ 的一个主要困难是如果后验分布相对于先验分布更加集中，大多数的 $p(\mathbf{D}|\theta^{(i)})$ 都将有很小的值， $\hat{p}_1(\mathbf{D})$ 将由少数的大似然值所控制。整个模拟过程不高效，收敛到高斯分布的速度很慢，而且最终的 $\hat{p}_1(\mathbf{D})$ 方差会很大。

一种改进(17)近似的方法是使用 *importance sampling* ，也即：

$$\hat{I} = \frac{1}{\sum_{i=1}^{m} w_i}\sum_{i=1}^{m} w_i p(\mathbf{D}|\theta^{(i)}) \ , \quad \text{其中} \theta^{(i)} \sim \pi^*(\theta) \ , \ \text{且} w_i = \frac{\pi(\theta^{(i)})}{\pi^*(\theta^{(i)})} \ 。 \tag{18}$$

$\pi^*(\theta)$ 被称为 *importance sampling function* 。

另一个更加高效的估计方法是基于 *adaptive Gaussian quadrature* （见 Genz & Kass (1993)）。

4.3 Simulating from the Posterior

很多方法可以被用来近似地从后验分布 $p(\theta|\mathbf{D}) = p(\mathbf{D}|\theta)\pi(\theta)/p(\mathbf{D})$ 获得抽样，常见的如 *direct simulation* 、 *rejection sampling* 、 *Markov Chain Monte Carlo (MCMC)* 、*Metropolis-Hastingg* 、*Gibbs sampling* 和 *weighted likilihood bootstrap* （Newton & Raftery (1994)）等等。

取(18)中的 $\pi^*(\theta) = p(\theta|\mathbf{D})$ ，把它代入(18)，得到：

$$\begin{aligned}\hat{p}_2(\mathbf{D}) &= \frac{1}{\sum_{i=1}^m w_i}\sum_{i=1}^m w_i p(\mathbf{D}|\theta^{(i)}) \\ &= \frac{1}{\sum_{i=1}^m \pi(\theta^{(i)})\frac{p(\mathbf{D})}{p(\mathbf{D}|\theta^{(i)})\pi(\theta^{(i)})}}\sum_{i=1}^m \pi(\theta^{(i)})\frac{p(\mathbf{D})}{p(\mathbf{D}|\theta^{(i)})\pi(\theta^{(i)})}p(\mathbf{D}|\theta^{(i)}) \\ &= \left\{\frac{1}{m}\sum_{i=1}^m p(\mathbf{D}|\theta^{(i)})^{-1}\right\}^{-1} \quad ,\end{aligned} \tag{19}$$

也即似然值的*调和平均值（harmonic mean）*。而且当 $m \to \infty$ 时，

$$\hat{p}_2(\mathbf{D}) \to p(\mathbf{D}) \ , \quad \text{almost surely} \quad 。 \tag{20}$$

但一般这种收敛不满足中心极限定理。 $\hat{p}_2(\mathbf{D})$ 的近似是不稳定的，但它容易计算，且通常能够获得足够精确的近似来解释 $\log B_{10}$ 。

为了克服 $\hat{p}_2(\mathbf{D})$ 的不稳定性，Newton & Raftery (1994) 建议取(18)中的 $\pi^*(\theta) = \delta\pi(\theta)+(1-\delta)p(\theta|\mathbf{D})$ 。把它代入(18)，得到：

$$\begin{aligned}\hat{p}_3(\mathbf{D}) &= \frac{1}{\sum_{i=1}^m w_i}\sum_{i=1}^m w_i p(\mathbf{D}|\theta^{(i)}) \\ &= \frac{1}{\sum_{i=1}^m \pi(\theta^{(i)})/[\delta\pi(\theta^{(i)})+(1-\delta)p(\theta^{(i)}|\mathbf{D})]}\sum_{i=1}^m \frac{\pi(\theta^{(i)})p(\mathbf{D}|\theta^{(i)})}{\delta\pi(\theta^{(i)})+(1-\delta)p(\theta^{(i)}|\mathbf{D})} \\ &= \frac{1}{\sum_{i=1}^m 1/[\delta+(1-\delta)\frac{p(\mathbf{D}|\theta^{(i)})}{p(\mathbf{D})}]}\sum_{i=1}^m \frac{p(\mathbf{D}|\theta^{(i)})}{\delta+(1-\delta)\frac{p(\mathbf{D}|\theta^{(i)})}{p(\mathbf{D})}} \\ &= \frac{\sum_{i=1}^m p(\mathbf{D}|\theta^{(i)})/[\delta\hat{p}_3(\mathbf{D})+(1-\delta)p(\mathbf{D}|\theta^{(i)})]}{\sum_{i=1}^m 1/[\delta\hat{p}_3(\mathbf{D})+(1-\delta)p(\mathbf{D}|\theta^{(i)})]} \quad 。\end{aligned} \tag{21}$$

$\hat{p}_3(\mathbf{D})$ 具有 $\hat{p}_2(\mathbf{D})$ 的高效性，但同时避免了它的不稳定性。而且可以证明 $\hat{p}_3(\mathbf{D})$ 满足中心极限定理。但 $\hat{p}_3(\mathbf{D})$ 的麻烦之处是我们必须同时从先验和后验分布中获得模拟。

$\hat{p}_3(\mathbf{D})$ 需要先后验分布同时模拟的麻烦可以使用如下方法解决：(1)从后验分布中抽取所有的 m 个 θ 值；(2)“想象”从先验中另外抽取 $\delta m/(1-\delta)$ 个 θ 值，并且假设这 $\delta m/(1-\delta)$ 中的 $\theta^{(i)}$ 对应的 $p(\mathbf{D}|\theta^{(i)})$ 都等于 $p(\mathbf{D})$ ，也即它们对应的 $\delta p(\mathbf{D})+(1-\delta)p(\mathbf{D}|\theta^{(i)})$ 变为 $p(\mathbf{D})$ 。上面的方法产生了一个 $\hat{p}_3(\mathbf{D})$ 的近似：

$$\hat{p}_4(\mathbf{D}) = \frac{\delta m/(1-\delta)+\sum_{i=1}^m p(\mathbf{D}|\theta^{(i)})/[\delta\hat{p}_4(\mathbf{D})+(1-\delta)p(\mathbf{D}|\theta^{(i)})]}{\delta m/[(1-\delta)\hat{p}_4(\mathbf{D})]+\sum_{i=1}^m 1/[\delta\hat{p}_4(\mathbf{D})+(1-\delta)p(\mathbf{D}|\theta^{(i)})]} \quad 。 \tag{22}$$

$\hat{p}_3(\mathbf{D})$ 和 $\hat{p}_4(\mathbf{D})$ 的计算可以使用简单的迭代方法获得。通常只需要几步迭代我们就可以获得收敛后的结果。

注意到：

$$\begin{aligned} p(\mathbf{D})^{-1} &= \int \frac{\pi(\theta)}{p(\mathbf{D}|\theta)\pi(\theta)} p(\theta|\mathbf{D}) d\theta \\ &= \int \frac{1}{p(\mathbf{D}|\theta)} p(\theta|\mathbf{D}) d\theta \quad , \end{aligned} \tag{23}$$

这也为近似计算 $\hat{p}_2(\mathbf{D})$ 提供了一个更加直观的解释。对于一个任意的概率密度函数 $f(\theta)$，同样有：

$$p(\mathbf{D})^{-1} = \int \frac{f(\theta)}{p(\mathbf{D}|\theta)\pi(\theta)} p(\theta|\mathbf{D}) d\theta \quad , \tag{24}$$

由此我们可以获得(19)的一个修正形式（Gelfand & Dey(1994)）：

$$\hat{p}_5(\mathbf{D}) = \left\{ \frac{1}{m} \sum_{i=1}^{m} \frac{f(\theta^{(i)})}{p(\mathbf{D}|\theta^{(i)})\pi(\theta^{(i)})} \right\}^{-1} \quad 。 \tag{25}$$

如果 $f(\cdot)$ 的尾巴足够瘦小，$\hat{p}_5(\mathbf{D})$ 满足中心极限定理。当 $f(\cdot)$ 大约正比于 $p(\mathbf{D}|\theta)$ 时，$\hat{p}_5(\mathbf{D})$ 的高效性最易体现。

4.4 Comparison of Methods

5 The Choice of Priors

到底该选取什么样的先验分布并没有什么一般的定律，它是依据具体问题而定的。

对于先验分布，首要的问题是如何选取它以便代表已知信息。另一个重要的问题是贝叶斯因子对于先验选取的*敏感性（sensitivity）*。

最简单的处理先验选择的问题是直接忽略它，而使用 Schwarz criterion 等方法计算贝叶斯因子。虽然这种方法在“充分大”样本量时可以获得正确的结论，但关键在于“充分大”没有什么界定准则。另一方面，不同于贝叶斯点估计（如后验均值），贝叶斯因子的确倾向于对模型参数的先验分布选取敏感。

5.1 Prior Information

正如为数据选择模型分布，选择先验分布时通常也做一些简化。这种简化对于嵌套模型假设（ $H_0 : (\psi = \psi_0, \beta);\ H_1 : (\psi, \beta)$ ）尤为突出。一种常用的假设是：

$$\pi(\beta|H_0) = \int \pi(\beta, \psi|H_1) d\psi \quad 。 \tag{26}$$

如果进一步假设 H_1 的先验中 β 和 ψ 相互独立，那么由上式可得 $\pi(\beta|H_0) = \pi(\beta|H_1)$ ，所以我们只要分别为 β 和 ψ 选择一个先验分布即可。另一种代替假设(26)的常用假设是：

$$\pi(\beta|H_0) = \pi(\beta|\psi = \psi_0, H_1) \quad 。 \tag{27}$$

在数据建模中，为先验分布做简化要非常小心，因为它们可能影响结果并且这种影响可能不被察觉。所以这里检验（testing，用来做假设检验）和估计（estimation，用来估计感兴趣量的值或分布）是不相同的。估计中通常为先验选择方便的形式，因为我们知道如果样本量充分大的话，先验的影响会很小。对于检验就不是这样了。

5.2 Sensitivity Analysis

既然要估计贝叶斯因子对先验分布的敏感性，自然首先应该明确模型 H_1 和 H_2 中可能使用的先验种类。如果已知的足够信息可以产生由超参数决定的初始先验分布（如 $\mathrm{N}(\nu, \phi^2)$ ），我们可以给予超参数一定的扰动（如 $\mu \to \mu \pm \phi$ ）以便计算扰动下的贝叶斯因子值。

一个近似计算(3)的重要方法是使用(11)。此时如果先验分布从 $\pi(\theta|H_*)$ 变为 $\pi^{(\mathrm{NEW})}(\theta|H_*)$ ，则贝叶斯因子从原来的 B_{12} 变为：

$$B_{12}^{(\mathrm{NEW})} \approx B_{12} \cdot \left(\frac{\pi^{(\mathrm{NEW})}(\hat{\theta}_1|H_1)}{\pi^{(\mathrm{NEW})}(\hat{\theta}_2|H_2)} \right) \cdot \left(\frac{\pi(\hat{\theta}_2|H_2)}{\pi(\hat{\theta}_1|H_1)} \right) \quad , \tag{28}$$

其中近似误差为 $O(n^{-1})$ 。(28)的简单性使得它在可选先验分布较多时很有用。

对于嵌套模型，在一些条件假设下，贝叶斯因子对参数 β 分布的选取依赖很小。

当有很少的先验信息时，使用在似然值较大的区域取值较平的先验分布对贝叶斯因子的冲击较小，所以这种先验在此时比较合适。

先验分布的选取应该使得所选先验在选取模型时不要贡献太多的 evidence 。

5.3 Bayes Factors with Improper Priors

所谓的 *improper priors* 其实就是平（均匀取值）的分布。对于嵌套模型，只在 H_1 中为参数 ψ 使用 improper 先验会使得贝叶斯因子倾向于喜欢零假设 H_0 。

6 Accounting For Model Uncertainty

实际情况下模型建立时往往包括很多候选的模型，而不止两个。例如，在回归中，我们需要决定哪些观测数据是噪音，决定哪些变量将被使用，决定对已选变量使用何种变换。每种可能的选择组合都将定义一个不同的模型。常用的处理这个问题的策略是使用一系列的显著性检验（significance tests）。

但这种常用做法存在几个问题。（1）整体策略的抽样性质往往不同于单独检验的性质，这种做法无法很好地理解整体策略的这种抽样性质。（2）正在比较的这些模型往往

不相互嵌套。（3）当设置显著水平时，power consideration 通常没有被考虑进来，检验的 power 特征往往并不知道。（4）任何选择单一模型然后基于它做推断的方法都忽视了模型选择中的不确定性，这可能会很大程度地低估感兴趣量的不确定性。

如果我们使用贝叶斯方法计算所有候选模型（这些候选模型是利用贝叶斯因子选出的）对应的后验概率，所有前面的困难都能避免。然后我们可以以简单的方式在做组合推断时计入模型不确定性。

6.1 Basic Ideas

假设有 $K+1$ 个候选模型： $H_0, H_1, \ldots, H_K$ 。 $H_1, \ldots, H_K$ 中的每一个轮流与 H_0 进行比较，产生贝叶斯因子 $B_{10}, \ldots, B_{K0}$ 。那么 H_k 的后验概率为

$$p(H_k|\mathbf{D}) = \frac{\alpha_k B_{k0}}{\sum_{l=0}^{K} \alpha_l B_{l0}} \quad , \tag{29}$$

其中 $\alpha_k = p(H_k)/p(H_0)$ ， $B_{00} = \alpha_0 = 1$ 。通常我们可以取所有的 $\alpha_k = 1$ ，但有时候也可以选取其他值来反映各个模型相对合理性的先验信息。

对于一个在所有模型中都有很好定义的感兴趣的量 Δ ，我们可以计算它在给定模型 H_k 的条件下的后验密度：

$$p(\Delta|\mathbf{D}, H_k) = \int p(\Delta|\mathbf{D}, \theta_k, H_k) p(\theta_k|\mathbf{D}, H_k) d\theta_k \quad 。 \tag{30}$$

(30)能够在给定模型 H_k 的条件下关于 Δ 做推断，但我们可能使用 Δ 的没有条件的后验密度，即：

$$p(\Delta|\mathbf{D}) = \sum_{k=0}^{K} p(\Delta|\mathbf{D}, H_k) p(H_k|\mathbf{D}) \quad 。 \tag{31}$$

(31)计入了模型形式的不确定性。它的后验均值和方差为：

$$\mathbb{E}[\Delta|\mathbf{D}] = \sum_{k=0}^{K} \mathbb{E}[\Delta|\mathbf{D}, H_k] \cdot p(H_k|\mathbf{D}) \quad , \tag{32}$$

$$\mathrm{var}[\Delta|\mathbf{D}] = \sum_{k=0}^{K} \left(\mathrm{var}[\Delta|\mathbf{D}, H_k] + (\mathbb{E}[\Delta|\mathbf{D}, H_k])^2\right) \cdot p(H_k|\mathbf{D}) - \mathbb{E}[\Delta|\mathbf{D}]^2 \quad 。 \tag{33}$$

如果可能的话，在做决定之前都不应该选择单一模型，应该使用组合模型(31)计入模型的不确定性。

6.2 Occam's Window

尽管模型不确定性的重要性以及存在处理它的一般策略，要使上面介绍的方法得到广泛使用，至少还有三个主要障碍。第一个是计算贝叶斯因子的困难性。

第二个困难是(31)中的求和可能包含非常多的项。例如回归中包含 n 个实例， J 个候选独立变量。如果考虑所有可能的变量子集（ 2^J ），所有可能的噪音情况（大约 $\binom{n}{O_{\max}}$ ）以及每个变量对应的四种可能转换（ 4^{J+1} ），那么可能的模型数目大约为 $2^J \times \binom{n}{O_{\max}} \times 4^{J+1}$ ，其中 $O_{\max}$ 为数据集中噪音点的最大可能数目。就算在 $n = 40$ ， $J = 12$ 以及 $O_{\max} = 5$ 时，模型数量也达到了 10^{16} 的量级。

第三个困难是需要为每个模型的参数指定先验分布。解决这个困难的方法有很多。一个是使用 Schwarz criterion ，另一个是为一个或几个“大”模型（模型参数很多）指定先验分布，然后其他模型看成它们的嵌套模型，通过约束它们的先验而获得其他模型的先验分布。

本节中介绍一种被称为 *Occam's window*（Madigan & Raftery (1994)）的方法，它被用来在初始时选择一部分模型以便克服初始模型过多的困难。这样(31)中的平均只在更少得多的模型上进行。Madigan & Raftery (1994) 认为如果一个模型的后验值 $p(H_k|\mathbf{D})$ 比最可能模型的值小很多，它就不将再被考虑。所以，对于不属于集合

$$\mathcal{A}' = \left\{ H_k : p(H_k|\mathbf{D}) \geq \frac{1}{C} \max_l \{p(H_l|\mathbf{D})\} \right\} \qquad (C \gg 1) \tag{34}$$

的模型，计算(31)时将不予与考虑。 C 的值由数据分析者确定，例如 $C = 20$ 。类似于 Occam's razor ，在计算(31)时他们也排除了那些获得更低后验值的复杂模型，也即排除了以下集合中的模型：

$$\mathcal{B} = \{ H_k : \exists H_l \in \mathcal{A}',\ H_l \subset H_k,\ p(H_l|\mathbf{D}) > p(H_k|\mathbf{D}) \} \quad 。 \tag{35}$$

所以方程(31)由下式代替：

$$p(\Delta|\mathbf{D}) = \frac{\sum_{H_k \in \mathcal{A}} p(\Delta|H_k, \mathbf{D}) p(\mathbf{D}|H_k) p(H_k)}{\sum_{H_k \in \mathcal{A}} p(\mathbf{D}|H_k) p(H_k)} \quad , \tag{36}$$

其中 $\mathcal{A} = \mathcal{A}' \setminus \mathcal{B}$ 。

典型地(36)中的项数降至 25 或更少，而且经常可以降至一两项。大多数包括集合 $\mathcal{A}$ 的初始类的最终选取结果都相同，在这个意义下最终的解可以说独立于初始类的选取。

6.3 Markov Chain Monte Carlo Model Composition (MC^3)

Madigan & York (1992) 建议使用 MCMC 方法近似(31)。他们创建一个*不可约（irreducible）*马尔科夫链 $\{H(t)\}$ （$t = 1, 2, \ldots$），使得它的状态空间为 $\mathcal{H} = \{H_k\}$ ，平稳分布为 $\{p(H_k|\mathbf{D})\}$ 。则对于任何函数 $u(H_k)$ ，马尔科夫链的平均值

$$\hat{U} = \frac{1}{m} \sum_{t=1}^{m} u(H(t)) \tag{37}$$

当 $m \to \infty$ 时都以概率 1 收敛到 $\mathbb{E}[u(H)]$ 。为了计算(31)，他们选取 $u(H) = p(\Delta|H, \mathbf{D})$ 。

为了创建前面的马尔科夫链，对于每个模型 H ，定义它的*邻居* $\mathrm{nbd}(H)$ 为它自己以及与它只有一个不同参数的模型所组成的集合。定义*转移矩阵* $\mathbf{R}$ 为：

$$\mathbf{R}(H \to H') = \begin{cases} 0\ , & \text{如果} H' \notin \mathrm{nbd}(H)\ ; \\ \mathrm{const}\ , & \text{如果} H' \in \mathrm{nbd}(H)\ 。 \end{cases} \tag{38}$$

从 $\mathbf{R}$ 中抽取出 H' ，然后以概率

$$\min\left\{1, \frac{p(H'|\mathbf{D})}{p(H|\mathbf{D})}\right\} \tag{39}$$

接受 H'，否则此链还呆在状态 H 。Madigan & York (1992) 认为通常取链长约为 $10,000$ 就足够了。

6.4 Model Expansion

Draper (1995) 建议了模型扩展方法：初始时只选择一个模型，然后扩展到一组模型，这组模型包括初始的那个模型作为特例，但是放松了初始模型中的一些结构性假设。方程(31)然后被约束到这组扩展模型上为感兴趣的量做推断。

模型扩展方法对于计入模型中关于特别结构假设的不确定性很有帮助，但它并不是被设计来计入模型建立（当很多模型初始时被考虑）中内在的不确定性。

6.5 Evaluation of Methods

模型的预测能力可以从预测将来数据的能力上进行判断。从某种意义上说平均下来方程(31)可以担保比其中的任何单个被选模型获得更好的预测效果（Madigan & Raftery (1994)）。

在几个数据集上，使用 Occam's window 和 MC^3 方法获得的模型平均比其中的任何单个被选模型产生了更好的预测结果。单个“好”模型之间的差异比考虑模型不确定性获得的改进还要小。相比于 Occam's window 方法，MCMC 方法获得了更好的预测效果，但代价是它需要更大的计算量，并且产生的结果更不容易解释。

7 Applications, Revisited

8 Issues And Controversies

8.1 Why Test Sharp Hypotheses?

8.2 Bayes Factors Versus Non-Bayesian Significance Testing

现在已经有很多文献讨论 Bayesian 和 non-Bayesian 检验之间的争论，下面我们简单介绍几点：

1. 期待 P 值和零假设正确的后验概率值相似是不合理的。一般的感觉是贝叶斯因子比 P 值更保守。
2. 频率学家的检验在大样本量时几乎系统地倾向于拒绝零假设，而贝叶斯因子不会。
3. 贝叶斯因子可以在不关心数据的非预定分析的情况下被应用（Bayes factors may be applied without concerns about unscheduled analysis of the data）。
4. 贝叶斯因子可以容易地应用于非嵌套模型和嵌套模型。与之相比的是，non-Bayesian significance tests 应用于非嵌套模型就很困难。
5. Non-Bayesian significance tests 被发展用来比较两个模型，但实际的数据分析中往往包括很多模型。

8.3 Bayes Factors Versus the AIC

Akaike (1973) 建议在一组候选模型中选择最小化

$$\text{AIC} = -2(\text{log maximized likelihood}) + 2(\text{number of parameters}) \tag{40}$$

的模型作为最优模型。

既然 AIC 只是从多个候选模型中选择某个“最好”的模型，它无法计入参数和模型的不确定性。Shibata (1976) 和 Katz (1981) 显示 AIC 倾向于过高估计需要的参数数量，也就是说 AIC 偏向于喜欢更复杂的模型。

Schwarz criterion 显示具有最高后验概率的模型最小化

$$\text{BIC} = -2(\text{log maximized likelihood}) + \log N \cdot (\text{number of parameters}) \quad 。 \tag{41}$$

比较(40)和(41)我们就可以发现，相对于 AIC ，BIC 偏向于喜欢更简单的模型。

9 Bibliographical Remarks And Additional Work

10 Conclusion

使用贝叶斯因子做模型检验时的三个主要问题：（1）贝叶斯因子的计算；（2）模型的先验分布选取；（3）对感兴趣量做推断时，应计入模型的不确定性。贝叶斯因子的主要限制是它们对模型假设和先验选择的敏感性。

References

[1] R. Kass and A. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.