【统计理论与方法】

大数据背景下贝叶斯模型平均的理论突破与应用前景

高 磊[1],刘乐平[1],卢志义[2]
(1. 天津财经大学 大数据统计分析中心,天津 300222; 2. 天津商业大学 理学院,天津 300134)

摘要:大数据统计分析过程中常面临模型比较和选择的不确定性问题。贝叶斯模型平均(BMA)方法可以通过先验和后验概率度量模型不确定性,并利用后验概率对模型的结果进行加权平均,最终得到更稳健的估计结果。在回顾贝叶斯模型平均发展历程的基础上,介绍贝叶斯模型平均的基本原理,综述其在一些难点问题上的理论进展,并介绍大数据背景下贝叶斯模型平均的应用前景。贝叶斯模型平均与复杂数据分析方法相结合,可能成为大数据研究的新思路。

关键词:大数据;模型不确定性;贝叶斯模型平均;MCMC
中图分类号:C829.29;O212.8 文献标志码:A 文章编号:1007-3116(2016)06-0014-09

一、引 言

在大数据的统计实践中,研究人员常常构建一组模型,然后依据模型选择准则,从中挑选"最优"模型进行统计推断和预测。这里就有两个值得考虑的问题:第一,单从模型选择准则来分析,可能会得到几个模型对数据拟合均较好的结论。也就是说,这些模型难分伯仲,舍弃其中任何一个都令人可惜,Breiman 就把这种现象称为模型选择的"罗生门效应"①(Rashomon Effect)[1];第二,即便可以选出最优模型,由于数据样本的随机性,每次选择的最优模型也可能有所不同,这就是所谓的模型不确定性(Model Uncertainty)[2]。在统计分析中,这两个问题较为普遍,但经常被研究人员所忽略。

贝叶斯模型平均(Bayesian Model Averaging, BMA)方法以贝叶斯理论为基础,将备选模型作为随机变量,通过赋予先验概率和后验概率来度量其不确定性,并利用后验概率对备选模型的结果进行加权平均,最终得到更稳健的估计结果。不妨做一比喻,如果单个模型的结果是含有金子的渣块,那么研究人员应像淘金者,与其选择含金量最大的渣块,不如利用一种方法从所有的渣块中淘取更多的金子。BMA 就是一种从多个模型中"淘金"的方法,即把各模型的结果综合起来,发挥各模型的优势,并融合更多的信息。

BMA 方法改变了人们对模型比较和模型选择

收稿日期:2016-01-18;修复日期:2016-04-20
基金项目:国家自然科学基金项目《Solvency II 框架下非寿险准备金风险度量与控制研究》(71171139);《多重风险相依情形下的最优保险问题研究》(71371138);《Basel III 框架下商业银行监管资本套利识别研究》(71303169);《逆周期资本监管框架下考虑跳跃行为的信用风险度量研究》(71401069);天津财经大学研究生科研资助计划(2014TCB03)
作者简介:高 磊,男,山东德州人,博士生,研究方向:精算与风险管理;
刘乐平,男,江西萍乡人,经济学博士,教授,博士生导师,研究方向:贝叶斯数据分析,精算与风险管理;
卢志义,男,内蒙古包头人,经济学博士,副教授,硕士生导师,研究方向:精算与风险管理。

① 《罗生门》是一部日本电影,在电影中发生了一起刑事案件,一名男性被杀,另有一名女性被强奸。案件共有四名目击者,当他们在法庭作证时,面对同样的事件,却从自身利益出发讲述了完全不同的事情经过。Brieman 认为统计模型也具有"罗生门效应",即不同模型讲述了关于同一数据的不一样的故事,而且听起来都非常逼真。

的传统认识，是对经典建模理论的有益补充。30 多年来，随着计算技术的不断进步，特别是 MCMC (Markov Chain Monte Carlo)方法的发展，BMA 方法渐趋成熟，其应用也愈加广泛。笔者在回顾 BMA 方法发展历程的基础上，介绍 BMA 方法的基本原理，并综述大数据背景下贝叶斯模型平均的理论突破，介绍 BMA 的一些应用。

二、贝叶斯模型平均的发展历程

模型平均起源于 20 世纪 60 年代，而 BMA 是模型平均方法的一个重要分支①。1963 年，Barnard 在研究民用航空数据时首次提出模型综合的概念。1965 年，Roberts 考虑了一种结合两名专家观点的预测分布，该分布本质是两个模型后验分布的加权平均。1969 年，Bates 和 Granger 通过综合两个无偏预测来预测航空需求，肯定了模型综合方法在统计预测中的优势，他们的论文催生了 20 世纪 70 年代关于模型综合的研究。1978 年，Leamer 进一步完善了模型综合方法，首次提出了 BMA 分析的基本范式，Leamer 认为 BMA 方法就是从概率的角度度量模型的不确定性。继 Leamer 之后，BMA 的研究沉寂了一段时期。

20 世纪 80 年代末 90 年代初，MCMC 方法的发展极大地促进了现代贝叶斯统计学的复兴[3]，与此同时，忽视模型不确定性所带来的弊端也再次引发了学者的思考。在这种背景下，George、Drapper、Raftery 等学者重新开展了 BMA 方法的研究，BMA 迎来了理论发展的黄金时期。在 10 多年的时间里，学者们针对设定先验分布、计算边际似然和模型搜索等难点问题进行了深入研究，并取得一系列理论进展(见本文第三部分)。1999 年，Hoeting 等人在国际著名统计期刊《*Statistical Science*》上发表了综述文章[4]，全面回顾 90 年代 BMA 方法的理论进展，并对 21 世纪 BMA 的应用前景进行展望，这篇文章标志着 BMA 方法渐趋成熟，目前该文引用已达 2 727 次②。

进入 21 世纪，BMA 在国内外得到了迅猛的发展和应用。2005 年，Gneiting 和 Raftery 合作在世界顶级学术刊物《*Science*》的气象科学板块撰文，指出利用贝叶斯模型平均方法进行天气预测更为有效[5]。除气象学研究外，在大数据背景下，BMA 方法的应用领域还包括计量经济学、医学健康、水文地理、工程技术等(见本文第四部分)。

三、大数据背景下贝叶斯模型平均的理论突破

(一)BMA 的基本原理

贝叶斯模型平均是一种从多个模型中“淘金”的方法。若 Δ 是所感兴趣的“金子”(Δ 可能是系数的估计，也可能是未来的预测)，那么在观测数据 D 给定的条件下，通过 BMA 方法得到的后验分布是：

$$p(\Delta \mid D) = \sum_{k=1}^{K} p(\Delta \mid M_k, D)\, p(M_k \mid D) \quad (1)$$

其中 $M_1, M_2, \cdots, M_K$ 表示备选模型，$p(M_k \mid D)$ 表示备选模型 M_k 的后验概率，而 $p(\Delta \mid M_k, D)$ 表示在备选模型 M_k 下 Δ 的后验分布。因此，由 BMA 所得的后验分布 $p(\Delta \mid D)$ 是各备选模型下后验分布 $p(\Delta \mid M_k, D)$ 的加权平均，加权权重为各备选模型的后验概率。

根据 Δ 后验分布式(1)，可得后验均值和方差：

$$E(\Delta \mid D) = \sum_{k=1}^{K} \hat{\Delta}_k p(M_k \mid D) \quad (2)$$

$$\mathrm{Var}[\Delta \mid D] = \sum_{k=1}^{K} (\mathrm{Var}[\Delta \mid D, M_k] + \hat{\Delta}_k^2) p(M_k \mid D) - E[\Delta \mid D]^2 \quad (3)$$

其中 $\hat{\Delta}_k$ 表示备选模型 M_k 下 Δ 的后验均值，即 $\hat{\Delta}_k = E[\Delta \mid D, M_k]$。因此，由 BMA 所得的后验均值是各备选模型后验均值的加权平均，加权权重为各备选模型的后验概率。类似地，式(3) 右侧第一项是各备选模型后验二阶矩的加权平均，加权权重也是各备选模型的后验概率。

研究表明，由 BMA 得到的均值预测式(2) 会优于单个模型预测。一个直观的解释是，若各模型的预测结果均是无偏的，那么选择单个模型的预测结果就如同从多个无偏预测中随机抽取一个预测值，虽然结果无偏，但其方差的不确定性仍然很大；而利用合适的权重对各模型的预测值加权平均，不仅仍然可以得到无偏估计，还可以降低估计的方差，从而提高估计的准确性。可证明，在对数得分标准下，由 BMA 得到的的预测不仅优于单个模型的预测，而且

① 模型平均方法另一个重要分支是频率模型平均(Frequentist Model Averaging, FMA)。

② 截至 2016 年 1 月 1 日。

比其他加权平均结果要好。

备选模型的后验概率 $p(M_k \mid D)$ 非常重要，在式(1)、式(2)、式(3)中均有出现，表示对备选模型的“信赖”程度。备选模型的后验概率可根据贝叶斯公式得到：

$$p(M_k \mid D) = \frac{p(D \mid M_k)p(M_k)}{\sum_{l=1}^{K} p(D \mid M_l)p(M_l)} \tag{4}$$

其中 $p(M_k)$ 是备选模型 M_k 的先验概率，$p(D \mid M_k)$ 是在备选模型 M_k 下观测数据 D 的边际似然，即：

$$p(D \mid M_k) = \int p(D \mid \theta_k, M_k)p(\theta_k \mid M_k)\mathrm{d}\theta_k \tag{5}$$

其中 θ_k 表示模型 M_k 中的参数向量，$p(\theta_k \mid M_k)$ 表示模型 M_k 中参数 θ_k 的先验分布，而 $p(D \mid \theta_k, M_k)$ 则表示在给定模型 M_k 和参数 θ_k 下，观测数据 D 的似然。式(5)涉及积分运算，因此边际似然又被称为积分似然。

式(1)～式(5)含义清楚易于理解，涵盖了 BMA 的基本方面，但在 BMA 的应用中，仍有不少细节需要考虑，某些细节至关重要已然成为 BMA 的应用难点。从 20 世纪 90 年代至今，学者们围绕这些难点问题进行了深入研究并取得了一系列理论进展。按照 BMA 分析的流程，这些难点可归纳为以下三个方面：

1. 设定先验分布。应用 BMA 时，首先需要设定参数和模型的先验分布。参数先验分布出现在计算模型边际似然的式(5)中，边际似然是计算模型后验概率的关键，而模型的先验概率也会影响模型后验概率。因此，如果选择了不稳健的先验分布，不仅会得到失真的模型后验概率，而且还会降低 BMA 的预测能力。

2. 求解边际似然。边际似然直接影响模型后验概率。与似然函数不同，边际似然是似然函数在参数先验分布下的期望，涉及积分运算。积分运算常常较为复杂，尤其当模型的参数维度较多时，一般积分算法难以处理这种高维积分问题。不过，在贝叶斯线性回归模型中，通过把参数设置为共轭先验，也可得到边际似然解的解析形式，但共轭先验只是先验分布的一种特殊情形，在更为复杂的贝叶斯模型中，如果根据需要将参数设定为非共轭先验，那么积分运算仍然会非常困难。求解边际似然是 BMA 应用中必须克服的一个难点。

3. 搜索模型空间。利用边际似然求解的近似或模拟方法，可以得到单个模型的边际似然，但当备选模型的数量巨大时，要求出所有模型的边际似然乃至后验概率，在计算上是不可能完成的。例如陈伟等人在利用贝叶斯模型平均方法预测中国通货膨胀率时，考虑了 28 个解释变量，在单一模型为线性模型假设下，备选模型总数多达 268 435 456 个[6]①。实际上，当解释变量个数超过 20 个时就不能像式(1)那样对所有模型加权平均，而如何设计一种模型搜索策略，在模型空间中进行搜索、得到模型空间的一个子集、然后在这个子集基础上进行 BMA，则是学者关注的又一个难点问题。

(二)设定参数先验分布

回归模型是 BMA 方法讨论最为成熟的模型。首先介绍回归模型中常用的先验形式——Zellner's g 先验，然后介绍另外两种设定参数先验的方法。

1. Zellner's g 先验。设被解释变量为 Y，解释变量为 $X_1, X_2, \cdots, X_p$。完整的回归模型应包括 p 个解释变量，而其他备选模型则考虑 p 个解释变量的一个子集，用 X_k 表示 $X_1, X_2, \cdots, X_p$ 的子集，则备选模型 M_k 可表示为：

$$M_k: Y = X_k\beta_k + \varepsilon \tag{6}$$

其中 X_k 为设计矩阵，β_k 为相应的回归系数，ε 为误差向量，一般假设 $\varepsilon \sim N_n(0, \sigma^2 I)$。

Zellner's g 先验的设置方法是在 σ^2 给定条件下，令回归系数 β_k 服从均值为 β^0，协方差矩阵为 $g\sigma^2(X_k'X_k)^{-1}$ 的多元正态分布，即：

$$p(\beta_k \mid \sigma^2) = N(\beta^0, g\sigma^2(X_k'X_k)^{-1}) \tag{7}$$

然后将方差 σ^2 设定为无信息先验分布：

$$p(\sigma^2) \propto \frac{1}{\sigma^2} \tag{8}$$

式(7)和式(8)构成回归模型的 Zellner's g 先验。在 Zellner's g 先验中，一般令 $\beta^0 = 0$，因此 Zellner's g 先验只需指定超参数 g 即可。在 Zellner's g 先验假设下，β_k 和 σ^2 的联合后验密度可以分解为：

$$p(\beta_k, \sigma^2 \mid D) \propto p(\beta_k \mid D, \sigma^2)p(\sigma^2 \mid D) \tag{9}$$

这里分解的两项 $p(\beta_k \mid D, \sigma^2)$ 和 $p(\sigma^2 \mid D)$ 均是常见的分布形式，其中 $p(\beta_k \mid D, \sigma^2)$ 是多元正态分布，

① 线性回归模型中，每个解释变量都有两种选择：进入模型或在模型外，因此若有 5 个解释变量，则模型空间中共有 $2^5 = 32$ 个备选模型；若有 20 个解释变量，则模型空间中共有 $2^{20} = 268\ 435\ 456$ 个备选模型，可见备选模型的数量随解释变量个数增加呈指数式增长。

$p(\sigma^2 \mid D)$ 是逆伽马分布。

Zellner's g 先验由 Zellner(1986)提出且应用较为广泛，在 BMA 方法中使用 Zellner's g 先验的好处是明显的：首先，Zellner's g 先验形式简洁，只需设定一个参数，而且参数的后验分布是常见的分布形式；其次，在 Zellner's g 先验假设下，备选模型对空模型(不含解释变量，仅有截距项)的贝叶斯因子容易求出，因此方便进行模型比较；此外，在马尔科夫链蒙特卡洛模型综合(Markov Chain Monte Carlo Model Composition，MC^3)等模型搜索算法中，Zellner's g 先验可以提高算法计算效率。

2. 数据依赖先验(data－dependent prior)。参数先验是进行数据分析之前关于参数的信息，与观测数据联系很少，因此"数据依赖先验"的概念可能令人疑惑。然而，当研究人员关于参数的先验信息极少时，将先验分布设定为数据依赖先验仍是可行的。Wasserman 证明，数据依赖先验不仅具有良好的性质，而且比采用数据独立先验(data－independent prior)有更好的预测能力。此外，由于考虑了数据信息，数据依赖先验比无信息先验更为稳健。

3. 单位信息先验(Unit Information Prior，UIP)。Kass 和 Wasserman 提出的单位信息先验的概念为多元正态分布，其均值为参数极大似然估计，协方差矩阵是由一单位观测数据得到的 Fisher 信息的逆矩阵。用一个例子描述其基本思想，假设观测数据的分布是正态分布，即：

$$Y_i \sim N(\psi,\sigma^2) \quad (i=1,2,\cdots,n) \tag{10}$$

这里 σ 已知；ψ 未知，是待估参数，那么 ψ 的单位信息先验可设定为：

$$\psi \sim N(\psi_0,\tau^2) \tag{11}$$

其中 ψ_0 为观测数据 $Y_1,Y_2,\cdots,Y_n$ 的样本均值，而 $\tau=\sigma$，这表示式(11)包含的关于 ψ 的信息(由 τ^2 度量)，与一单位观测数据包含的关于 ψ 的信息(由 σ^2 度量)是相同的，这正是单位信息先验名称的由来。由该先验得到的贝叶斯因子与施瓦茨准则结果接近，但也面临与施瓦茨准则相同的问题，即其模型选择结果比较保守，偏向于较为简单的模型。

(三) 设定模型先验概率

Raftery 提出模型的均匀先验，即为所有模型指定相同的先验概率：

$$p(M_k)=\frac{1}{K}(k=1,2,\cdots,K) \tag{12}$$

其中 K 是模型空间中备选模型的总数。模型均匀先验表示对所有备选模型都一视同仁，不偏向也不歧视任何备选模型。在均匀先验下，模型后验概率可以进一步简化：

$$p(M_k \mid D)=\frac{p(D \mid M_k)p(M_k)}{\sum_{l=1}^{K}p(D \mid M_l)p(M_l)}=\frac{p(D \mid M_k)}{\sum_{l=1}^{K}p(D \mid M_l)} \tag{13}$$

可见，由于各模型先验概率相同，分子分母中先验概率一项可以消去，模型后验概率不受模型先验的影响，只由模型的边际似然决定。在 BMA 应用研究中，设定模型均匀先验较为流行，这是因为从形式上看式(12)简洁、方便，从应用上看这种方式也符合直觉，容易被数据分析客户接受。

Mitchell 和 Beauchamp 提出，在回归模型中从解释变量的角度设置模型先验：

$$p(M_k)=\prod_{j=1}^{p}\pi_j^{\delta_{kj}}(1-\pi_j)^{1-\delta_{kj}} \tag{14}$$

其中 δ_{kj} 为指示变量，$\delta_{kj}=1$ 表示解释变量 X_j 进入回归模型 M_k 中，$\delta_{kj}=0$ 表示解释变量 X_j 在回归模型 M_k 之外；π_j 表示解释变量 X_j 进入回归模型的先验概率，一般假设 $\pi_1=\pi_2=\cdots=\pi_p=\pi$。若 $\pi=0.5$，式(14)就等同于均匀先验式(12)；若 $\pi>0.5$，表示解释变量进入回归模型的可能性大，这种先验意味着对大模型的偏好；若 $\pi<0.5$，表示解释变量进入回归模型的可能性小，这种先验则意味着对大模型的惩罚。

(四) 求解边际似然

方便起见，将边际似然式(5)中关于模型的信息去掉，边际似然简化为：

$$p(D)=\int p(D \mid \theta)p(\theta)\mathrm{d}\theta \tag{15}$$

下面介绍在 BMA 应用中，两类常用的求解边际似然方法：近似算法和随机模拟算法。近似算法主要是拉普拉斯方法(Laplace's Method)，而随机模拟算法则包括蒙特卡洛模拟(Monte Carlo，MC)和马尔科夫链蒙特卡洛模拟(Markov Chain Monte Carlo，MCMC)两种方法。

1. 拉普拉斯方法。Tierney 和 Kadane 提出了拉普拉斯方法[7]。该方法分两步进行：

首先，用多元正态分布对后验分布进行近似。当观测数据 D 样本容量较大时，后验分布函数 $p(\theta \mid D)=A\cdot p(D \mid \theta)p(\theta)$($A$ 是归一化常量)在后验众数 $\tilde{\theta}$ 附近是一个尖峰，因此可将后验分布函数的对数形式 $l(\theta)=\log(A\cdot p(D \mid \theta)p(\theta))$，在 $\tilde{\theta}$ 附近进行

二阶泰勒级数展开，得到以下近似：

$$l(\theta) \approx l(\tilde{\theta}) + (\theta - \tilde{\theta})' l''(\tilde{\theta})(\theta - \tilde{\theta})/2 \quad (16)$$

这里 $l''(\tilde{\theta})$ 是 $l(\theta)$ 在 $\tilde{\theta}$ 处海森矩阵的取值。可以看出，式(16)中二次项 $(\theta - \tilde{\theta})' l''(\tilde{\theta})(\theta - \tilde{\theta})/2$ 与多元正态分布密度函数指数项部分类似，因此后验分布可由多元正态分布近似。式(16)两边同时取指数，可得：

$$p(\theta \mid D) = \exp(l(\theta)) \approx A' \exp\left(-\frac{(\theta - \tilde{\theta})'(-l''(\tilde{\theta}))(\theta - \tilde{\theta})}{2}\right) \quad (17)$$

这意味着 $p(\theta \mid D)$ 可由均值为 $\tilde{\theta}$、协方差矩阵为 $V = (-l''(\tilde{\theta}))^{-1}$ 的多元正态分布近似，归一化常量 $A' = (2\pi)^{-d/2}(-l''(\tilde{\theta}))^{1/2}$。

其次，运用基本边际似然等式(Basic marginal likelihood identity，BMI)求解边际似然：

$$p(D) = \frac{p(D \mid \theta)p(\theta)}{p(\theta \mid D)} \quad (18)$$

该式对于所有 θ 都成立，因此不妨取 $\theta = \tilde{\theta}$，并用多元正态分布近似 $p(\theta \mid D)$，代入式(18)可得：

$$p(D)_1 = \frac{p(D \mid \tilde{\theta})p(\tilde{\theta})}{p(\tilde{\theta} \mid D)} \approx (2\pi)^{d/2} p(\tilde{\theta}) f(D \mid \tilde{\theta})(-l''(\tilde{\theta}))^{-1/2} \quad (19)$$

以上是求解边际似然的拉普拉斯方法，$p(D)_1$ 就是边际似然的拉普拉斯近似。

2. Monte Carlo 方法。在式(15)中，如果将 θ 看作服从分布 $p(\theta)$ 的随机变量，而将 $p(D \mid \theta)$ 看作是随机变量 θ 的函数，那么式(15)相当于利用积分方法求 $p(D \mid \theta)$ 的期望。设 $\{\theta^{(i)}: i = 1,2,\cdots m\}$ 表示来自于 $p(\theta)$ 的样本，那么边际似然就可用样本均值进行估计：

$$\hat{p}(D)_2 = \frac{1}{m}\sum_{i=1}^{m} p(D \mid \theta^{(i)}) \quad (20)$$

这种方法的关键是利用 Monte Carlo 方法得到先验分布 $p(\theta)$ 的样本 $\{\theta^{(i)}: i = 1,2,\cdots,m\}$。如果 $p(\theta)$ 是常见的分布形式，如正态分布、伽马分布、贝塔分布等，利用统计软件中的随机数生成函数就可以得到 $p(\theta)$ 的样本；而如果先验分布 $p(\theta)$ 形式复杂，但与之近似的分布 $p^*(\theta)$ 的样本却容易随机模拟，这时可以将式(15)转化为：

$$p(D) = \int p(D \mid \theta)\frac{p(\theta)}{p^*(\theta)}p^*(\theta)d\theta \quad (21)$$

设 $\{\theta^{(i)}: i = 1,2,\cdots,m\}$ 表示来自于分布 $p^*(\theta)$ 的样本，那么边际似然就可以用样本均值进行估计：

$$\hat{p}(D)_3 = \frac{1}{m}\sum_{i=1}^{m} p(D \mid \theta^{(i)})\frac{p(\theta^{(i)})}{p^*(\theta^{(i)})} \quad (22)$$

其中 $p(\theta^{(i)})/p^*(\theta^{(i)})$ 表示样本点 $\theta^{(i)}$ 处的重要性权重，式(22)就是求解积分的重要性抽样方法(Importance Sampling，IS)。利用重要性抽样求解积分具有悠久的历史，但其效率很大程度上依赖于选择合适的建议分布。在 θ 维度不高的情况下，建议分布选择 T 分布则可以取得较好的估计效果。θ 维度较高时，有包括 Meng 和 Wong 的桥抽样(Bridge sampling)、Gelman 和 Meng 的路径抽样(Path sampling)、Chen 和 Shao 的比率重要性抽样(Ratio important sampling)等方法可供选择。

3. Markov Chain Monte Carlo 方法。Newton 和 Raftery 提出利用 MCMC 方法估计边际似然[8]。设利用 MCMC 方法随机模拟得到了后验分布 $p(\theta \mid D)$ 的样本 $\{\theta^{(i)}: i = 1,2,\cdots,m\}$，那么就可通过下式估计边际似然：

$$\hat{p}(D)_4 = \left\{\frac{1}{m}\sum_{i=1}^{m}\frac{1}{p(D \mid \theta^{(i)})}\right\}^{-1} \quad (23)$$

由此可见，边际似然是一种调和均值估计，属于一致估计，但由于似然函数倒数的方差并不总是有界的，所以该估计的有效性较差。

Gelfand 和 Dey 提出了 $\hat{p}(D)_4$ 的改进形式：

$$\hat{p}(D)_5 = \left\{\frac{1}{m}\sum_{i=1}^{m}\frac{f(\theta^{(i)})}{p(D \mid \theta^{(i)})p(\theta^{(i)})}\right\}^{-1} \quad (24)$$

其中 $\{\theta^{(i)}: i = 1,2,\cdots,m\}$ 是 MCMC 方法的后验分布样本；$f(\theta)$ 是专门设计的密度函数，其均值和协方差矩阵与后验分布相同，但具有比后验分布更薄的尾部。通过引入这样的 $f(\theta)$，$\hat{p}(D)_5$ 不仅是 $p(D)$ 的无偏一致估计，而且满足中心极限定理，估计比较稳定，但对于参数维度较多的情况，找到满足上述要求的 $f(\theta)$ 并非易事。

Chib 提出用 Gibbs 抽样结果估计边际似然[9]。Chib 重新考虑了基本边际似然等式，认为该式对于所有 θ 均成立，因此不妨取 θ 为极大似然估计 $\hat{\theta}$(或后验众数估计)：

$$p(D) = \frac{p(D \mid \hat{\theta})p(\hat{\theta})}{p(\hat{\theta} \mid D)} \quad (25)$$

$p(D \mid \hat{\theta})p(\hat{\theta})$ 的计算比较容易，真正的难点在于计算 $p(\hat{\theta} \mid D)$。Chib 发现，如果将参数 θ 分成多个参数块，而且每一参数块 Gibbs 抽样的条件分布解析形式是已知的，那么就可以进行多轮 Gibbs 抽样并利

用 Rao-Blackwellization 技术求出 $p(\hat{\theta} \mid D)$，其主要思路仍然是把求积分的复杂问题转化为求样本均值的简单问题；如果 θ 由两个参数块构成，那么仅需一次 Gibbs 抽样的样本即可；如果 θ 可划分为 B 个参数块，则共需 $(B-1)$ 轮 Gibbs 抽样。Chib 和 Jeliazkov 将这一方法扩展到 Metropolis Hasting 算法[10]。理论上，这种方法可以应用到几乎所有的 MCMC 算法，但是当参数 θ 被划分成多个参数块时，多轮 Gibbs 抽样会花费较长的时间，这可能会影响该方法的实际应用。

(五)模型搜索策略

利用近似或模拟等方法可以得到单个模型的边际似然，但当备选模型的数量巨大时，要求出所有模型的边际似然乃至后验概率，在计算上是不可能完成的。在这种情况下，可以采用模型搜索策略，搜索重要的模型构成模型空间的一个子集，然后在这个子集基础上进行贝叶斯模型平均。下面介绍三种模型搜索策略：Occam 窗口方法、逆跳马尔可夫链蒙特卡罗方法(Reversible Jump Markov Chain Monte Carlo，RJMCMC)、马尔可夫链蒙特卡罗模型综合方法(Markov Chain Monte Carlo Model Composition，MC^3)。

1. Occam 窗口方法。Madigan 和 Raftery 提出用 Occam 窗口方法选择一个模型子集[11]。令 A 表示模型空间 $\{M_1, M_2, \cdots, M_K\}$。在筛选模型时，有以下两条准则：

第一，将后验概率非常低的模型删掉，模型后验概率的高低是相对的，如果具有最大后验概率的模型 M_{k^*} 与模型 M_k 后验概率相比超过一个阈值，比如 20 倍，就认为 M_k 后验概率非常低，考虑将这个模型删掉。也就是说，如果模型 M_k 属于集合：

$$A_1 = \{M_k : \frac{\max_l \{\mathrm{pr}(M_l \mid D)\}}{\mathrm{pr}(M_k \mid D)} \geqslant C\} \tag{26}$$

就将模型 M_k 从模型空间 A 中删掉，这里 C 是由研究者选择的一个阈值。

第二，将后验概率较低的复杂模型删掉。这与 Occam 剃刀原理相似，即“如无必要，勿增实体”。对于一个复杂模型，如果其子模型具有更高的后验概率，就保留子模型而将原模型删掉。也就是说，如果模型 M_k 属于集合：

$$A_2 = \{M_k : M_l \subset M_k, \frac{\mathrm{pr}(M_l \mid D)}{\mathrm{pr}(M_k \mid D)} > 1\} \tag{27}$$

就将模型 M_k 从模型空间 A 中删掉。经过处理，模型空间中备选模型的数量大大减少，式(1)简化为：

$$\mathrm{pr}(\Delta \mid D) = \sum_{M_k \in A_3} \mathrm{pr}(\Delta \mid M_k, D)\mathrm{pr}(M_k \mid D) \tag{28}$$

这里 $A_3 = A/(A_1 \cup A_2)$。

2. 逆跳 MCMC 方法。如果说 MCMC 方法促进了现代贝叶斯统计学的复兴，那么 Green 提出的逆跳 MCMC 方法则被视为贝叶斯分析的革命。由逆跳 MCMC 方法构建的马氏链不仅可以在单个模型的参数空间内进行转移，还可以在不同模型、不同维度参数空间之间实现跳跃，从而为 BMA 模型搜索提供强大工具。设逆跳 MCMC 当前模型状态为 k，参数状态为 θ_k，θ_k 的维度为 d_k，那么从当前状态 (k, θ_k) 向下一状态转移的步骤如下：

步骤 1　从模型建议分布 $w(k, k^*)$ 中，生成一个建议模型 k^*。

步骤 2　从建议分布 $q(\mu \mid \theta_k, k, k^*)$ 中，生成随机向量 μ。

步骤 3　令 $(\theta_k^{**}, u^*) = T(\theta_k, u)$，其中 T 为转换函数，且 T 和其反函数 T^{-1} 相同，θ_k^{**} 为模型 k^* 的建议参数状态。

步骤 4　计算由当前状态 (k, θ_k) 向建议状态 (k^*, θ_k^{**}) 转移的接受概率：

$$\alpha = \min\left\{1, \frac{p(M_{k^*}, \theta_k^{**} \mid D)}{p(M_k, \theta_k \mid D)} \frac{p(M_{k^*})}{p(M_k)} \frac{w(k, k^*)}{w(k^*, k)} \frac{q(u^* \mid \theta_k^{**}, k, k^*)}{q(u \mid \theta_k, k, k^*)} \left| \frac{\partial T(\theta_k, u)}{\partial(\theta_k, u)} \right|\right\} \tag{29}$$

如果建议分布 $q(\mu \mid \theta_k, k, k^*)$ 是 $\theta_k^{**} \mid M_{k^*}$ 的后验分布，此时接受概率中的雅可比行列式就等于 1，这样接受概率就简化为：

$$\alpha = \min\left\{1, \frac{p(M_{k^*}, \theta_k^{**} \mid D)}{p(M_k, \theta_k \mid D)} \frac{p(M_{k^*})}{p(M_k)} \frac{w(k, k^*)}{w(k^*, k)} \frac{q(u^* \mid \theta_k^{**}, k, k^*)}{q(u \mid \theta_k, k, k^*)}\right\} \tag{30}$$

3. MC^3 方法。Madigan 和 York 提出马尔可夫链蒙特卡罗模型综合算法对模型进行抽签[12]。MC^3 方法倾向于抽取后验概率较高的模型，在一定数量的抽签后，能保证抽签结果收敛于基于所有模型的结果。与逆跳 MCMC 算法相似，MC^3 方法也是构造一条关于模型的马式链：$M^{(1)}, M^{(2)}, \cdots, M^{(N)}$，并且这条马氏链的平稳分布就是模型的后验概率分布。为了构造这样的马氏链，要为任意一个模型定义相邻的模型空间，用以从中抽取备选模型。以线性模型为例，第 $s+1$ 次从如下模型空间中等概率地抽取备选模型：当前模型 $M^{(s)}$、当前模型 $M^{(s)}$ 删减一个解释变

量的模型、当前模型 $M^{(s)}$ 增加一个解释变量的模型。备选模型生成后，以如下接受概率判断是否接受备选模型 M^*：

$$\alpha(M^{(s)},M^*)=\min\left\{1,\frac{p(D\mid M^*)p(M^*)}{p(D\mid M^{(s)})p(M^{(s)})}\right\} \tag{31}$$

其中 $p(D\mid M^*)$ 和 $p(D\mid M^{(s)})$ 是边际似然，可以用 Laplace、MC 或 MCMC 等方法计算得到。如果假设所有模型先验概率相等，那么式(31)就简化为计算两个模型的贝叶斯因子：$p(D\mid M^*)/p(D\mid M^{(s)})$。George 和 McCulloch 提出了与 MC^3 相似的模型搜索策略，即随机搜索变量选择(Stochastic search variable selection，SSVS)。在对模型抽签时，MC^3 会移除一个解释变量，但在 SSVS 中，所有解释变量被赋予一个概率，一个解释变量不会真正移除，而是以很大的概率趋于零。

四、大数据背景下贝叶斯模型平均的应用前景

在大数据背景下，借助于便捷的统计软件工具，BMA 在国内外得到了广泛应用，特别是考虑到大数据价值的稀疏性，范剑青提出利用多个模型拟合数据，然后按照系统的观点利用模型平均方法将各模型结果综合起来，提取大数据的内在价值①。目前，BMA 的应用领域包括医学健康、计量经济学、工程技术、气象预报、水文地理等。

在医学健康方面：Volinsky 等人研究了美国成年人的中风死亡因素，发现应用 BMA 于 Cox 比例风险模型(Proportional hazard model)，可以避免逐步回归方法忽视的模型不确定性，改善对中风的预测，同时改进潜在中风患者的风险评价；Annest 等在癌症与基因相关性的研究中，发现通过 BMA 可以选择数量较少的预测基因，但仍能对癌症的复发和转移进行有效预测；Bobb 等人采用 1987 年至 2005 年美国 105 个城市的相关数据，构建了一组时间序列模型，估计高温天气对人类死亡率的影响，然后利用 BMA 将这些模型的结果进行综合，从而系统地解决了模型选择的不确定性问题；Carroll 等人利用 BMA 方法研究了美国乔治亚州不同区域肠癌的影响因素，发现在乔治亚州的北部郡县，中产阶级的收入和非裔人群比例是肠癌的重要预测变量；在乔治亚州的南部，贫困线下人口比例和非裔人群比例是肠癌的重要影响因素[13]。

在计量经济学方面：BMA 应用成果丰硕，Fernández 等人利用 BMA 研究了跨国经济增长回归模型，发现拥有不同协变量的回归模型后验概率相差不大，证明模型不确定性的确存在，其实证结果支持萨拉伊马丁 的“乐观”结论，即多组协变量在解释跨国经济增长中具有重要作用；2008 年，Wright 在《计量经济学》发表文章，提出了对汇率预测的随机游走模型进行贝叶斯模型平均，研究发现模型平均的预测效果比单一模型要好；Wright 又使用 BMA 对单一时间序列预测模型进行了综合，发现这种方法对美国通胀率的预测优于简单平均的方法，也优于单一时间序列预测模型；Jacobson 和 Karlsson 使用类似的方法预测了瑞典的通胀；Eicher 等利用 BMA 方法，综合考虑几种关于外商直接投资(Foreign Direct Investment，FDI)的模型，并且进一步将 BMA 扩展为 HeckitBMA，解决了模型选择偏倚问题。

中国学者陈伟和牛霖琳运用 BMA 对中国通货膨胀率进行建模，并对样本外通胀进行预测，发现在均方根误差标准下，BMA 方法的预测优于 AR 模型、主成分分析模型、菲利普斯曲线模型、利率期限结构模型等单一模型[6]；王亮和刘金金采用 BMA 方法，使用 1970—2007 年的省际数据，对影响中国经济增长的因素进行了分析，发现高等教育发展阶段、工业化推进速度、对外开放程度、东部区位优势、消费能力和对内开放水平等 6 个解释变量对中国经济增长具有长期、持续和稳健的影响[14]；朱慧明等人采用逆跳 MCMC 方法选择分位数自回归模型的阶次，沪深 300 指数的实证研究显示，贝叶斯方法可以有效地识别分位数回归的阶次并进行参数估计；司明和孙大超采用 BMA 方法分析发达国家主权债务危机成因，发现金融危机冲击、经济增长率下降、失业率升高、人口老龄化和政府预算收入降低是债务危机爆发的主要原因；高丽君采用中小企业信用数据，利用传统生存模型、Bootstrap 生存模型和 BMA 生存模型估计中小企业信用违约情况，研究发现 BMA 生存模型结果准确率较高，Bootstrap 生存模型次之，传统生存模型准确率最低；李佳蓓等人对多元线性回归问题中的变量选择方法进行了研

① 2015 年 12 月 20 日，范剑青在中国人民大学统计与大数据研究院暨大数据论坛做了题为“大数据人才培养：复旦方案”的报告。这里的观点是笔者根据报告的部分内容整理得到的。

究，改进了现有的贝叶斯自适应抽样方法，数据仿真发现改进后的方法预测效果比改进前更好[15]。

此外，在工程技术领域，Raftery 和 Kárny 在传统 BMA 方法基础上，提出了动态模型平均技术（Dynamic Model Averaging，DMA），并将其运用到冷轧机输出的在线预测中；王华伟等人采用 BMA 方法，研究了不同失效模式对航空发动机可靠性的影响；在气象预测领域，Raftery 等将 BMA 引入到对各种气象参数的预测中，目前 BMA 已经在气温、气压、风速、能见度等预测中得到广泛应用。Fang 和 Li 在运用气候大数据对过去 1 000 年气候变化模拟时，利用 BMA 方法综合考虑了不同气候模型的模拟结果，发现 BMA 方法能够发挥各模型的优势，得到更可靠的气候变化模拟结果；在水文地理方面，梁忠民等人发现基于 BMA 的水文模型合成预报，不仅可以提供精度较高的均值预测，而且可以通过预测分布评价预测的不确定性。

五、结束语

从 1978 年 Leamer 提出贝叶斯模型平均方法至今，已有逾 30 年的历史，在这 30 多年里，贝叶斯模型平均方法渐趋成熟，其应用也愈加广泛。笔者回顾了贝叶斯模型平均的发展历程，介绍了贝叶斯模型平均的基本原理，综述了大数据背景下贝叶斯模型平均的理论突破，并介绍了大数据背景下贝叶斯模型平均在各领域中的应用。

贝叶斯模型平均方法不偏好也不摒弃各个模型，而是对各模型结果进行综合，以期发挥各模型优势，融合更多信息。贝叶斯模型平均的魅力不仅在其对模型结果的综合，还在于这种方法本身所蕴藏的贝叶斯智慧。贝叶斯模型平均改变了人们对模型比较和模型选择的传统认识，是对经典建模理论的有益补充。本文的目的是系统地介绍贝叶斯模型平均方法的理论进展，为国内学者应用贝叶斯模型平均方法提供参考。笔者相信在大数据背景下，将贝叶斯模型平均方法应用到中国社会、经济各领域的数据分析中，将会得到更多有用的信息和有价值的结论。

参考文献：

[1] Breiman L. Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)[J]. Statistical Science, 2001 (3).

[2] Clyde M, George E I. Model Uncertainty[J]. Statistical Science, 2004(1).

[3] 刘乐平，高磊，杨娜. MCMC 方法的发展与现代贝叶斯的复兴[J]. 统计与信息论坛，2014 (2).

[4] Hoeting J A, Madigan D, Raftery A E, et al. Bayesian Model Averaging: A Tutorial[J]. Statistical Science, 1999(4).

[5] Gneiting T, Raftery A E. Weather Forecasting with Ensemble Methods[J]. Science, 2005(10).

[6] 陈伟，牛霖琳. 基于贝叶斯模型平均方法的中国通货膨胀的建模及预测[J]. 金融研究，2013(11).

[7] Tierney L, Kadane J B. Accurate Approximations for Posterior Moments and Marginal Densities[J]. Journal of the American Statistical Association, 1986(3).

[8] Newton M A, Raftery A E. Approximate Bayesian Inference with the Weighted Likelihood Bootstrap[J]. Journal of the Royal Statistical Society. Series B (Methodological), 1994(1).

[9] Chib S. Marginal Likelihood from the Gibbs Output[J]. Journal of the American Statistical Association, 1995(12).

[10] Chib S, Jeliazkov I. Marginal Likelihood From the Metropolis-Hastings Output[J]. Journal of the American Statistical Association, 2001(3).

[11] Madigan D, Raftery A E. Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window[J]. Journal of the American Statistical Association, 1994(12).

[12] Madigan D, York J, Allard D. Bayesian Graphical Models for Discrete Data[J]. International Statistical Review/Revue Internationale de Statistique, 1995(8).

[13] Carroll R, Lawson A B, Faes C, et al. Bayesian Model Selection Methods in Modeling Small Area Colon Cancer Incidence[J]. Annals of Epidemiology, 2016(1).

[14] 王亮，刘金全. 中国经济增长的决定因素分析——基于贝叶斯模型平均（BMA）方法的实证研究[J]. 统计与信息论坛，2010(9).

[15] 李佳蓓，朱永忠，王明刚. 贝叶斯变量选择及模型平均的研究[J]. 统计与信息论坛，2015(8).

【统计理论与方法】

Alpha 正态分布及其在环境污染中的应用

陈明明[a],马江洪[b],姬楠楠[b]
(长安大学 a. 经济与管理学院;b. 理学院,陕西 西安 710064)

摘要:目前,对实际数据的处理常采用一些对称分布,如正态分布和 t 分布等,而这种对称分布所给出的结果往往并不能令人满意。偏分布常用来处理有偏重尾数据,基于传统正态分布,提出一种处理偏态和重尾数据的 alpha 正态分布,并研究其参数估计方法及基本性质。将所提分布应用于环境污染数据,通过拟合检验 alpha 正态分布给出了很好的结果。

关键词:偏正态分布;alpha 正态分布;MLE 估计;环境污染
中图分类号:O212.1;F205 **文献标志码**:A **文章编号**:1007-3116(2016)06-0022-06

一、引 言

众所周知,通常假设诸多领域中的数据都服从正态分布,以期得到所需的结论。但是,在实际生活中所得到的数据往往呈现出有偏重尾的现象,此时将其假设为正态分布并不合理,这时就需要更灵活的统计模型来描述这些数据。对数正态分布作为正态分布的一种变换,其密度函数为 $f_{LN}(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma x}\exp[-\frac{(\ln x-\mu)^2}{2\sigma^2}]$, $x>0$,常被用来分析正实数集上的有偏数据,它的一个重要应用就是在污染物浓度中的应用。范绍佳利用中国最早建站的北

收稿日期:2015-12-31
基金项目:国家自然科学基金项目《基于信息瓶颈编码原理的深度学习研究》(11501049);国家自然科学基金项目《模糊假设的统计检验理论和方法研究》(11261044)
作者简介:陈明明,女,山东枣庄人,博士生,研究方向:运输统计分析。
马江洪,男,陕西绥德人,教授,理学博士,研究方向:数据挖掘的统计学方法。
姬楠楠,女,陕西渭南人,讲师,理学博士,研究方向:深度学习。

On Theoretical Breakthrough and Application Prospect of BMA in Context of Big Data

GAO Lei[1], LIU Le-ping[1], LU Zhi-yi[2]
(1. Center for Big Data Analysis, Tianjin University of Finance and Economics, Tianjin 300222, China;
2. School of Science, Tianjin University of Commerce, Tianjin 300134, China)

Abstract: Model comparison and selection uncertainty issue is very common in the big data analysis. The Bayesian model averaging (BMA) treats model as stochastic variable and assigns prior and posterior probability for it in order to account for model uncertainty. BMA weights the results of each model by their posterior model probability, and in the end obtatin more robust results. In this paper, we briefly describe the origins and developments of BMA, introduce the paradigm of BMA, and then discuss new progresses of BMA. Some important aspects of application are given in the context of big data. BMA combined with complex data analysis methods will provide new insights in our big data research methods.

Key words: big data; model uncertainty; Bayesian model averaging; MCMC

(责任编辑:郭诗梦)