遗传 Hereditas (Beijing) 2014 年 2 月, 36(2): 111－118
www.chinagene.cn

综 述

基因组育种值估计的贝叶斯方法

王重龙[1,2], 丁向东[2], 刘剑锋[2], 殷宗俊[3], 张勤[2]

1. 安徽省农业科学院畜牧兽医研究所, 合肥 230031;
2. 中国农业大学动物科技学院, 北京 100193;
3. 安徽农业大学动物科技学院, 合肥 230036

摘要: 基因组育种值估计是基因组选择的重要环节, 基因组育种值的准确性是基因组选择成功应用的关键, 而其准确性在很大程度上取决于估计方法。目前研究和应用最多的基因组育种值估计方法是贝叶斯(Bayes)和最佳线性无偏预测(BLUP)两大类方法。文章系统介绍了目前已提出的各种 Bayes 方法, 并总结了该类方法的估计效果和各方面的改进。模拟数据和实际数据研究结果都表明, Bayes 类方法估计基因组育种值的准确性优于 BLUP 类方法, 特别对于存在较大效应 QTL 的性状其优势更明显。由于 Bayes 方法的理论和计算过程相对复杂, 目前其在实际育种中的运用不如 BLUP 类方法普遍, 但随着快速算法的开发和计算机硬件的改进, 计算问题有望得到解决; 另外, 随着对基因组和性状遗传结构研究的深入开展, 能为 Bayes 方法提供更为准确的先验信息, 从而使 Bayes 方法估计基因组育种值准确性的优势更加突出, 应用将会更加广泛。

关键词: 基因组选择; 基因组育种值; 贝叶斯方法

Bayesian methods for genomic breeding value estimation

Chonglong Wang[1,2], Xiangdong Ding[2], Jianfeng Liu[2], Zongjun Yin[3], Qin Zhang[2]

1. *Institute of Animal Husbandry and Veterinary Medicine, Anhui Academy of Agricultural Sciences, Hefei* 230031, *China*;
2. *College of Animal Science and Technology, China Agricultural University, Beijing* 100193, *China*;
3. *College of Animal Science and Technology, Anhui Agricultural University, Hefei* 230036, *China*

Abstract: Estimation of genomic breeding values is the key step in genomic selection. The successful application of genomic selection depends on the accuracy of genomic estimated breeding values, which is mostly determined by the estimation method. Bayes-type and BLUP-type methods are the two main methods which have been

收稿日期: 2013–06–07; **修回日期:** 2013–12–12

基金项目: 农业部 948 计划(编号：2011-G2A), 教育部博士学科点专项科研基金项目(编号：20110008110001), 国家高技术研究发展计划(863 计划)项目(编号：2011AA100302), 国家自然科学基金项目(编号：31371258, 31171200, 31272418), 国家农业科技成果转化资金项目(编号：2011GB2C300017), 国家生猪产业技术体系(编号：CARS-36), 科技富民强县专项行动计划, 黎平黄牛品种资源保护与开发利用研究(编号：黔农育专字(2010)016 号), 安徽省现代农业项目, 安徽省生猪产业技术体系, 安徽省农业科学院成果推广项目(编号：13E0403), 安徽省农业科学院院长杰出青年创新基金项目(编号：13B0405)和安徽省农业科学院科技创新团队建设项目(编号：13C0405)资助

作者简介: 王重龙, 博士, 副研究员, 研究方向：猪遗传育种与健康养殖。E-mail: ahwchl@163.com

通讯作者: 张勤, 博士, 教授, 博士生导师, 研究方向：动物分子数量遗传学。E-mail: qzhang@cau.edu.cn

DOI: 10.3724/SP.J.1005.2014.0111

网络出版时间: 2013-12-24 16:56:02

URL: http://www.cnki.net/kcms/detail/11.1913.R.20131224.1656.001.html

widely studied and used. Here, we systematically introduce the currently proposed Bayesian methods, and summarize their effectiveness and improvements. Results from both simulated and real data showed that the accuracies of Bayesian methods are higher than those of BLUP methods, especially for the traits which are influenced by QTL with large effect. Because the theories and computation of Bayesian methods are relatively complicated, their use in practical breeding is less common than BLUP methods. However, with the development of fast algorithms and the improvement of computer hardware, the computational problem of Bayesian methods is expected to be solved. In addition, further studies on the genetic architecture of traits will provide Bayesian methods more accurate prior information, which will make their advantage in accuracy of genomic estimated breeding values more prominent. Therefore, the application of Bayesian methods will be more extensive.

Keywords: genomic selection; genomic estimated breeding values (GEBVs); Bayesian methods

基因组选择(Genomic selection, GS)已成为动植物遗传育种领域的最新研究热点。基因组育种值的估计是基因组选择的重要环节，估计基因组育种值(Genomic estimated breeding values, GEBVs)的准确性是基因组选择成功应用的关键，而其准确性在很大程度上受估计方法的影响。在估计基因组育种值的过程中，由于需要估计效应的标记数目通常远远多于表型记录数，因此需要估计的模型效应变量数(p)远远超过有观察值的样本数(n)，即“大 p，小 n”问题，会导致多重共线性和过度参数化。为解决这个问题一系列估计方法已被提出和研究，如最小二乘法(Least-squares, LS)[1]、偏最小二乘法(Partial least squares, PLSR)和主成分回归法(Principal component regression, PCR)[2]、随机回归 BLUP(Random regression BLUP, RRBLUP)[1]、GBLUP[3]、TABLUP[4]、BayesA 和 BayesB[1]、BayesCπ 和 BayesDπ[5]、Bayesian SSVS[6]、Bayesian LASSO[7]、EN[8]、半参数方法(Semiparametric procedures)[9]和机器学习方法(Machine learning methods)[10]等。目前研究和使用最多的为 BLUP 和 Bayes 两大类方法。本文对基因组育种值估计的贝叶斯(Bayes)方法的研究进展进行了综述。

1 Bayes 方法先验假设及模型

Bayes 方法基于如下先验假设：影响性状的所有 QTL 中，只有少数具有较大效应，大多数具有微小效应，而在全基因组的 SNP 中，只有少数与这些 QTL 相连锁，表现出或大或小的效应，多数 SNP 与 QTL 不连锁，不表现出效应。基于此，该方法的模型分为 2 个水平，即数据水平和标记效应方差水平。

①数据水平模型：$\mathbf{y}=\mathbf{Xb}+\sum_{k=1}^{q}\mathbf{z}_k g_k+\mathbf{e}$.，其中 $\mathbf{y}$ 为 n 个个体的表型观察值向量；$\mathbf{b}$ 为 f 维“固定”效应向量(包括群体均值)，$\mathbf{X}$ 为“固定”效应的 $n\times f$ 维关联矩阵；g_k 为第 k 个 SNP 的效应值，q 为 SNP 个数，$\mathbf{z_k}$ 为 n 个个体在第 k 个 SNP 的基因型向量，z_{ki} 取值−1、0 或 1 对应于该 SNP 的 3 种基因型(00、01 或 11)；$\mathbf{e}$ 为剩余随机残差向量并且 $\mathbf{e}\sim N(0,\sigma_e^2)$。于是 $p(\mathbf{g},\mathbf{b}\,|\,\mathbf{y})\propto p(\mathbf{y}\,|\,\mathbf{g},\mathbf{b})p(\mathbf{g},\mathbf{b})$，其中 $p(\mathbf{g},\mathbf{b})$是 $\mathbf{g}$ 和 $\mathbf{b}$ 的先验分布密度函数，$p(\mathbf{y}|\mathbf{g},\mathbf{b})$是 $\mathbf{y}$ 的似然函数。②标记效应方差水平模型：$p(\sigma_{g_k}^2\,|\,g_k)\propto p(g_k\,|\,\sigma_{g_k}^2)p(\sigma_{g_k}^2)$，其中 $\sigma_{g_k}^2$ 为第 k 个 SNP 的效应值的方差，$p(\sigma_{g_k}^2)$ 是 $\sigma_{g_k}^2$ 的先验分布密度函数，$p(g_k\,|\,\sigma_{g_k}^2)$ 是 g_k 的似然函数。

2 Bayes 估计方法分类

根据对 SNP 效应及其方差的先验分布的不同假设，可将 Bayes 方法分为 BayesA、BayesB[1]、BayesCπ、BayesDπ[5]、Bayesian SSVS[6]和 Bayesian LASSO[7]等。

2.1 BayesA

BayesA[1]的先验分布假设：(1) $p(g_k\,|\,\sigma_{g_k}^2)=N(0,\sigma_{g_k}^2)$，即 SNP 效应 g_k 服从正态分布，并且不同 g_k 的方差不同为 $\sigma_{g_k}^2$；(2) $p(\sigma_{g_k}^2)=x^{-2}(v,S)$，即 SNP

效应方差 $\sigma_{g_k}^2$ 服从自由度为 v、尺度参数为 S 的逆卡方分布; (3) $p(\sigma_e^2)=x^{-2}(-2,0)$，即残差效应方差 σ_e^2 服从自由度为–2、尺度参数为 0 的逆卡方分布; (4) $p(\mathbf{b})\propto$ 常量，即“固定”效应及群体均值服从均匀分布。

似然函数:

$$p(\mathbf{y}\mid\mathbf{g},\mathbf{b},\sigma_e^2)\propto(\sigma_e^2)^{-\frac{n}{2}}$$
$$\exp\left\{-\frac{1}{2\sigma_e^2}\sum_{i=1}^{n}(y_i-\sum_{j=1}^{f}x_{ij}b_j-\sum_{k=1}^{q}z_{ki}g_k)^2\right\}$$

于是可由联合后验分布:

$$p(\mathbf{g},\mathbf{b},\sigma_g^2,\sigma_e^2\mid\mathbf{y})\propto p(\mathbf{y}\mid\mathbf{g},\mathbf{b},\sigma_e^2)p(\mathbf{g},\mathbf{b},\sigma_g^2,\sigma_e^2)$$
$$=p(\mathbf{y}\mid\mathbf{g},\mathbf{b},\sigma_e^2)p(\mathbf{b})p(\mathbf{g}\mid\sigma_g^2)p(\sigma_g^2)p(\sigma_e^2)$$

获得各变量的完全条件后验分布。运用 MCMC (Markov Chain Monte Carlo)算法，通过各变量的完全条件后验分布进行 Gibbs 抽样，以获得各变量的估计值。

Gibbs 抽样过程简要如下:

第 1 步，初始化。

给所有未知参数赋初值 $\boldsymbol{\theta}^{(0)}=[b_1^{(0)},\cdots,b_f^{(0)},g_1^{(0)},\cdots,g_q^{(0)},\sigma_{g_1}^{2\,(0)},\cdots,\sigma_{g_q}^{2\,(0)},\sigma_e^{2(0)}]'$。

第 2 步，更新 b_j。

从 b_j 的完全条件后验分布 $N\left((\mathbf{X}_j'\mathbf{X}_j)^{-1}(\mathbf{X}_j'\mathbf{y}-\mathbf{X}_j'\mathbf{X}_{-j}\mathbf{b}_{-j}-\mathbf{X}_j'\mathbf{Z}\mathbf{g}),(\mathbf{X}_j'\mathbf{X}_j)^{-1}\sigma_e^2\right)$ 抽样，其中 $\mathbf{X}_j$ 是矩阵 $\mathbf{X}$ 中与 b_j 对应的列，$\mathbf{X}_{-j}$ 是 $\mathbf{X}$ 中去除 $\mathbf{X}_j$ 后的剩余部分，$\mathbf{b}_{-j}$ 是 $\mathbf{b}$ 中去除 b_j 后的剩余部分。

第 3 步，更新 g_k。

从 g_k 的完全条件后验分布 $N\left((\mathbf{Z}_k'\mathbf{Z}_k+\sigma_e^2/\sigma_{g_k}^2)^{-1}(\mathbf{Z}_k'\mathbf{y}-\mathbf{Z}_k'\mathbf{X}\mathbf{b}-\mathbf{Z}_k'\mathbf{Z}_{-k}\mathbf{g}_{-k}),(\mathbf{Z}_k'\mathbf{Z}_k+\sigma_e^2/\sigma_{g_k}^2)^{-1}\sigma_e^2\right)$ 抽样，其中 $\mathbf{Z}_k$ 是矩阵 $\mathbf{Z}$ 中与 g_k 对应的列，$\mathbf{Z}_{-k}$ 是 $\mathbf{Z}$ 中去除 $\mathbf{Z}_k$ 后的剩余部分，$\mathbf{g}_{-k}$ 是 $\mathbf{g}$ 中去除 g_k 后的剩余部分。

第 4 步，更新 σ_e^2。

从 σ_e^2 的完全条件后验分布 $x^{-2}\left[n-2,(\mathbf{y}-\mathbf{X}\mathbf{b}-\mathbf{Z}\mathbf{g})'(\mathbf{y}-\mathbf{X}\mathbf{b}-\mathbf{Z}\mathbf{g})/(n-2)\right]$ 抽样。

第 5 步，更新 $\sigma_{g_k}^2$ $(k=1,2,\cdots,q)$。

从 $\sigma_{g_k}^2$ 的完全条件后验分布 $x^{-2}\left[v+1,(vS+g_k'g_k)/(v+1)\right]$ 抽样。

重复 2~5 步，直到收敛达平稳分布，并获得足够的样本。

2.2 BayesB

BayesB[1]与 BayesA 的区别在于对 SNP 效应的先验假设不同。BayesA 假设所有 SNP 都有效应，并且每个 SNP 效应都有不同的方差。BayesB 假设大多数 SNP 没有效应，少数 SNP 有效应并各自有不同的效应方差。因此, BayesB 对 $\sigma_{g_k}^2$ 先验分布的假设为: 以概率 π，$\sigma_{g_k}^2$ =0; 以概率 1–π，$p(\sigma_{g_k}^2)=x^{-2}(v,S)$，这里的 π 是人为给定的。其他变量的先验假设同 BayesA。

应用于 BayesB 的 MCMC 算法组成，除了对 $\sigma_{g_k}^2$ 的抽样是采用 Metropolis-Hastings 算法，而不是简单地从逆卡方分布中抽取外，其余与 BayesA 相同。

2.3 BayesCπ和 BayesDπ

BayesCπ和 BayesDπ[5]是针对 BayesA 和 BayesB 的两个缺陷提出的改进方法。其中, BayesCπ除了用于基因组育种值估计，还用于全基因组关联分析[11,12]。BayesA 和 BayesB 的第一个统计缺陷是, SNP 效应方差的先验分布为逆卡方分布，其自由度小，尺度参数经常是从某一特定遗传假设的加性遗传方差推导而来[13,14]。不论测定的基因型或表型的数量多少，SNP 效应的完全条件后验分布仅比其先验分布增加了 1 个自由度。这与贝叶斯的理念冲突，并导致 SNP 效应的压缩程度严重依赖于位点特异方差先验分布的尺度参数[14]。当 SNP 密度增加时，这个问题变得更加严重。有两种办法可以克服这个缺陷：一是用一个共同效应方差去代替各 SNP 特异方差; 二是将先验分布中的尺度参数作为未知参数(有它自己的先验分布)进行估计。BayesCπ采用了第一种策略，对于所有效应为非零的 SNP(概率为 1–π)设置一个共同方差; BayesDπ采用了第二种策略，并假设尺度参数 S 的先验分布为 Gamma(1,1)[5]。

BayesA 和 BayesB 的另一个缺陷是把某个 SNP 效应为零的概率值π作为已知参数。在 BayesA 中，π=0，即所有 SNP 都有非零效应; 而在 BayesB 中，π>>0，以适应很多 SNP 效应为零的先验知识。因为 SNP 效应的压缩程度受π值的影响，所以应该把π作为未知参数，由数据信息推断获得。BayesCπ、BayesDπ方法中都采用了这个策略，并假设π的先验

分布为 Uniform(0,1)[5]。

2.4 Bayesian SSVS

Bayesian SSVS，即贝叶斯随机搜索变量选择方法(Bayesian stochastic search variable selection, Bayesian SSVS)[6]，与 BayesA 或 BayesB 比较，其最重要的特点是引入了一个指示变量γ到分层模型中。这样使信息提取与变量选择联系起来。潜在变量取值 1 或 0，代表该 SNP 是否有显著效应而应包括(或不包括)在模型中。如此，SNP 效应的先验分布是一个以相应的γ为条件的混合正态分布：$g_k \mid \gamma_k, \sigma^2_{g_k} \sim (1-\gamma_k)N(0, \sigma^2_{g_k}/100) + \gamma_k N(0, \sigma^2_{g_k})$，其方差的先验分布同样是逆卡方分布，即 $\sigma^2_{g_k} \sim x^{-2}(v, S)$。指示变量$\gamma_k$的先验分布为一个贝努利分布，即$\gamma_k \sim bernoulli(p_k)$，其中 p_k 是γ_k 取值为 1 的先验概率，它反映了多少 QTL 影响目标性状，可量化为与 QTL 连锁的 SNP 标记数量除以总的 SNP 标记数。在应用基因组选择时，期望的 QTL 比例可由目标性状及先前 QTL 研究的知识估计。相应的指示变量γ_k 的完全条件后验分布为：

$$p(\gamma_k = 1 \mid \mathbf{b}, \boldsymbol{\gamma}_{-k}, \sigma^2_e,) \sim bernoulli\left(\frac{p(g_k \mid \boldsymbol{\gamma}_{-k}, \gamma_k = 1)p_k}{p(g_k \mid \boldsymbol{\gamma}_{-k}, \gamma_k = 1)p_k + p(g_k \mid \boldsymbol{\gamma}_{-k}, \gamma_k = 0)(1-p_k)}\right),$$

它反映了每个 SNP 出现在模型中的频率。频繁出现在模型中的 SNP 标记的后验概率高，也最有可能与一个 QTL 连锁。

2.5 Bayesian LASSO

Bayesian LASSO，即贝叶斯最小绝对压缩和选择操作方法(Bayesian least absolute shrinkage and selection operator, Bayesian LASSO)[7]是通过最小化残差平方和及约束回归系数绝对值的和获得回归系数估计值，即：

$$\min_{\beta,\lambda}[(\mathbf{y} - \sum \mathbf{X}_j\beta_j)'(\mathbf{y} - \sum \mathbf{X}_j\beta_j) + \lambda \sum |\beta_j|]$$

其中 $t \geq \sum |\beta_j| (t \geq 0)$，$\lambda \geq 0$ 是一个拉格朗日乘子，与边界值 t 有隐含关系并控制压缩程度。通常采用的先验分布为双指数分布(又称拉普拉斯分布)：

$p(\mathbf{g}) = \prod_{k=1}^{q} \frac{\lambda}{2} \exp(-\lambda \mid g_k \mid)$，它是一个两水平的层级模型分布，由 $p(g_k \mid \sigma^2_{g_k}) = N(0, \sigma^2_{g_k})$ 与 $p(\sigma^2_{g_k}) = Expon(\lambda^2/2)$ 或 $p(\sigma^2_{g_k}) = gamma(1, \lambda^2/2)$ 混合组成[15~17]。

3 Bayes 方法相互间及其与 BLUP 方法的估计效果比较

BLUP 方法和 Bayes 方法是估计 GEBV 的两类主流方法，因此有许多研究对这两类方法的效果进行比较。效果评介指标主要有：①GEBV 与真实育种值(TBV)间的相关系数($r_{TBV,\ GEBV}$)，代表其准确性，其平方称为可靠性；②TBV 对 GEBV 的回归系数($b_{TBV,\ GEBV}$)，代表其无偏性，若 $b_{TBV,\ GEBV}$ =1 表明无偏，否则有偏；③GEBV 的均方误，MSE；④计算时间。

3.1 模拟数据

Meuwissen 等[1]用模拟数据的研究结果表明，BayesA 和 BayesB 的估计准确性比 RRBLUP 方法分别提高约 9%和 16%; BayesB 的无偏性优于 RRBLUP 方法，BayesA 的无偏性不及 RRBLUP 方法，但三者的 $b_{TBV,GEBV}$ 都小于 1; BayesA 和 BayesB 因为使用了 MCMC 算法，都很耗时间，但当时模拟研究使用的是 Pentium500PC，若使用大型计算机，对实际育种数据的处理也还是可行的。详细结果见表 1。

Usai 等[18]分析了第 12 届 QTL-MAS Workshop 模拟数据[19]，结果表明：BayesA 和 Bayesian LASSO 的准确性比 RRBLUP 分别提高约 12%和 19%; Bayesian LASSO 稍微低估了真实育种值; Bayesian LASSO 需要的计算时间要高于 RRBLUP 和 BayesA。详细结果见表 2。

表 1 BayesA、BayesB 和 RRBLUP 方法估计 GEBVs 的结果比较[1]

方法	$r_{TBV,\ GEBV} \pm SE$	$b_{TBV,\ GEBV} \pm SE$	计算时间
RRBLUP	0.732±0.030	0.896±0.045	–
BayesA	0.798	0.827	2 weeks
BayesB	0.848±0.012	0.946±0.018	1 day

表 2 BayesA、Bayesian LASSO 和 RRBLUP 方法分析第 12 届 QTL-MAS Workshop 模拟数据的结果比较[18]

方法	$r_{TBV,\ GEBV}$	$b_{TBV,\ GEBV}$	计算时间
RRBLUP	0.748	0.868	00:23:50
BayesA	0.836	0.916	04:36:10
Bayesian LASSO	0.894	1.148	07:06:20

Sun 等[11]分析了第 14 届 QTL-MAS Workshop 模拟数据[20]，结果表明：BayesCπ准确性最高，BayesB 准确性受π值影响，当π取值合适时，BayesB 和 BayesCπ的准确性近似，都比 GBLUP 提高 6%左右。详细结果见表 3。

表 3　BayesB、BayesCπ和 GBLUP 方法分析第 14 届 QTL-MAS Workshop 模拟数据的结果比较[11]

方法	$r_{TBV,GEBV}$	$b_{TBV,GEBV}$
GBLUP	0.610	0.949
BayesB,π=0.75	0.632	0.950
BayesB,π=0.95	0.640	0.960
BayesB,π=0.99	0.646	0.967
BayesCπ	0.650	0.952

Lund 等[19]、Bastiaansen 等[21]和 Pszczola 等[22]分别总结了第 12、13 和 14 届 QTL-MAS Workshop 关于基因组选择的研究结果，都表明 Bayes 类方法优于 BLUP 类方法，但后者在运算效率上有很大的优势。出现差别的主要原因在于两者先验假设的不同，BLUP 方法的先验假设是微效多基因模型，即所有的 SNP 效应都是微小的，且服从相同的正态分布，而 Bayes 方法的先验假设是少数 QTL 具有较大效应，大多数 QTL 具有微小效应，其先验分布如前文所述。大多数模拟数据考虑影响性状的 QTL 仅有 50 个或更少，其遗传结构更近似于 Bayes 方法的先验假设，这可能是 Bayes 方法在模拟数据研究中的效果优于 BLUP 方法的最主要原因。Daetwyler 等[23]特别针对这个问题，开展了遗传结构对估计方法效果影响的模拟研究，其结果表明，对于特定的样本规模和性状遗传力，不论影响某个性状 QTL 数目的多少，BLUP 方法都有一个恒定的准确性。而 BayesB 方法在影响某个性状 QTL 的数目较少时，具有较高的准确性，而随着 QTL 数目的增多，其优势下降，当 QTL 数目很多时，BayesB 方法的准确性略低于 BLUP 方法。BLUP 方法在 QTL 数目很少时仍然保持一个较高准确性的原因，可能主要在于 QTL 附近的许多 SNP 都捕获了一部分微小效应，从而使 SNP 效应的分布能较好地近似于 BLUP 方法的先验假设。

3.2　真实数据

实际数据研究结果[6,24~26]表明，Bayes 方法准确性相对于 BLUP 方法的优势，虽然没有如同模拟数据中大，但总体上仍有一定的优势。Hayes 等[27]总结了世界上较早开展 GS 的几个国家的实际数据研究结果。澳大利亚的研究结果表明，除了繁殖性状，Bayes 方法对其他性状估计育种值的可靠性比 BLUP 方法高 2%~7%。新西兰的研究结果表明，Bayes 方法的可靠性比 BLUP 方法高 2%~3%。

VanRaden 等[28]在北美地区荷斯坦牛群体中，以 1952 年~1998 年出生的 3 576 头公牛作为参考群体，1999 年~2002 年出生的 1 759 头公牛作为验证群体，使用 Illumina 公司牛 50k 芯片进行 SNP 基因型检测，最终 38 416 个 SNPs 通过质量控制过程，分别采用 BLUP 方法(Linear)和 Bayes 方法(Nonlinear)进行 27 个性状的 GEBVs 估计。虽然整体上 Bayes 方法的估计可靠性仅比 BLUP 方法高 1%，但在乳脂率性状上的优势高达 8%，这可能是由于 *DGAT1* 基因[29]对乳脂率性状有大效应。

4　Bayes 方法的改进

从上述的探讨中可以看出，在大多数性状上 Bayes 方法的准确性优于 BLUP 方法，在某些性状上优势明显。但传统 Bayes 方法的第一个缺陷是运算时间太长，其次是标记效应先验分布的超参数不合适，第三是模型中未考虑剩余微效多基因效应。下文将从这几个方面介绍 Bayes 方法的改进。

4.1　BayesA 的 EM 算法

Hayashi 等[30]为提高 BayesA 估计 GEBVs 的计算效率，将 EM 算法应用其中。为实现 EM 过程，将 $\sigma^2_{g_k}$ $(k=1,2,\cdots,q)$ 作为缺失数据，其条件后验期望值($\hat{\sigma}^2_{gk}$)(给定 g_k 的当前估计值 $\hat{g}_k$)为：

$$\hat{\sigma}^2_{g_k}=\frac{\hat{g}^2_k+S}{v+1} \tag{I}$$

用 $\hat{\sigma}^2_{g_k}$ 代替 $\sigma^2_{g_k}$，得到完全数据的对数后验分布函数关于 $\sigma^2_{g_k}$ 的条件数学期望：

$$\ln p(\boldsymbol{\theta}|\mathbf{y},\mathbf{U})\propto-\frac{n}{2}\ln(\sigma^2_e)^{-\frac{n}{2}}-\frac{\sum_{i=1}^{n}(y_i-\sum_{j=1}^{f}x_{ij}b_j-\sum_{k=1}^{q}z_{ik}g_k)^2}{2\sigma^2_e}-\sum_{k=1}^{q}\left\{(v/2+1)\ln(\hat{\sigma}^2_{g_k})+\frac{g^2_k+S}{2\hat{\sigma}^2_{g_k}}\right\}$$

求该期望关于 g_k、b_j 和 σ_e^2 的最大值，获得它们的一个新的估计值：

$$\hat{g}_k = \frac{\sum_{i=1}^{n} z_{ik}(y_i - \sum_{j=1}^{f} x_{ij}\hat{b}_j - \sum_{l\neq k} z_{il}\hat{g}_l)}{\sum_{i=1}^{n} z_{ik}^2 + \hat{\sigma}_e^2 / \sigma_{g_k}^2} \quad (k=1,2,\cdots,q) \quad \text{(II)}$$

$$\hat{b}_j = \frac{\sum_{i=1}^{n} x_{ij}(y_i - \sum_{h\neq j} x_{ih}\hat{b}_h - \sum_{k=1}^{q} z_{ik}\hat{g}_k)}{\sum_{i=1}^{n} x_{ij}^2} \quad (j=1,2,\cdots,f) \quad \text{(III)}$$

$$\hat{\sigma}_e^2 = \frac{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{f} x_{ij}\hat{b}_j - \sum_{k=1}^{q} z_{ik}\hat{g}_k)^2}{n} \quad \text{(IV)}$$

应用于 BayesA 方法的 EM 算法总结如下：

首先赋 g_k(k=1, 2,⋯, q)初值，然后对以下的 E 步和 M 步进行迭代直至收敛：

E 步：用公式(I)计算 $\sigma_{g_k}^2$ 的条件期望值 $\hat{\sigma}_{g_k}^2$；

M 步：用 $\hat{\sigma}_{g_k}^2$ 代替 $\sigma_{g_k}^2$ 得到完全数据的对数后验分布函数关于缺失数据 $\sigma_{g_k}^2$ 的条件数学期望，求该期望关于参数 g_k、b_j 和 σ_e^2 的最大值，利用公式(II)、(III)和(IV)获得 g_k、b_j 和 σ_e^2 新的估计值 $\hat{g}_k$、$\hat{b}_j$ 和 $\hat{\sigma}_e^2$。

在应用上述 EM 算法的基础上，如同 Bayesian SSVS，还可引入一个取值为 0 或 1 的指示变量 γ_k，其模型相应修改为：

$$\mathbf{y} = \mathbf{Xb} + \sum_{k=1}^{q} \gamma_k \mathbf{z}_k g_k + \mathbf{e}$$

该新方法被称为 wBSR。模拟结果表明，基于 EM 的 wBSR 方法估计准确性优于基于 MCMC 的 BayesA 方法，且计算时间由>3 小时缩短到<2 分钟。

4.2 BayesB 的 ICE 算法

Meuwissen 等[31]为提高 BayesB 方法的运算效率，提出了一种新的快速迭代算法，即 ICE(Iterative Conditional Expectation)算法。该算法首先考虑一个 SNP 效应的估计模型：

$$\mathbf{y} = \mathbf{z}g + \mathbf{e}$$

该模型与前文描述的 BayesB 方法采用的模型本质上是一致的，此处只是为了简化模拟过程而没有考虑非遗传“固定”效应及群体均值。同样，此处 **y** 为观察值向量，**z** 为标准化的 SNP 基因型向量，g 为 SNP 效应，**e** 为剩余残差向量。

SNP 效应 g 的先验分布为：

$$p(g) = \begin{cases} 1/2\pi\lambda \exp(-\lambda|g|) & \text{for } g \neq 0 \\ (1-\pi) & \text{for } g = 0 \end{cases}$$

引入一个单位脉冲德耳塔函数(Dirac delta function)$\delta(g)$后，上述先验分布可表示为：

$$p(g) = 1/2\pi\lambda \exp(-\lambda|g|) + (1-\pi)\ \delta(\mathrm{g})$$

g 的后验分布为：

$$p(g|\mathbf{y}) = \frac{\varphi(Y;g,\sigma^2)p(g)}{\int_{-\infty}^{\infty}\varphi(Y;g,\sigma^2)p(g)d\ g}$$

其中 $\phi()$ 是标准正态分布密度函数，$Y = (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{y} = \mathbf{z}'\mathbf{y}/n$；$\sigma^2 = (\mathbf{z}'\mathbf{z})^{-1}\sigma_e^2 = \sigma_e^2/n$。

g 条件期望为：

$$E[g|\mathbf{y}] = \int g\ p(g|\mathbf{y})dg = \frac{\int_{-\infty}^{\infty} g\ \varphi(Y;g,\sigma^2)p(g)d\ g}{\int_{-\infty}^{\infty}\varphi(Y;g,\sigma^2)p(g)d\ g}$$

它具有闭合形式(Closed form)，可解析求解。

扩展到 m 个 SNPs, ICE 过程如下：

赋 g_i(i=1,2,⋯,m)初始值，例如 $\hat{g}$ =0，然后进入以下迭代步骤直至收敛。

对于 i=1,2,⋯,m：

第一步：计算较正的观察值 $\mathbf{y}_{-i} = \mathbf{y} - \sum_{j\neq i} z_j \hat{g}_j$，于是 $Y_i = \mathbf{z}_i'\mathbf{y}_{-i}/n$ 且 $\sigma^2 = \sigma_e^2/n$；

第二步：计算 $\hat{g}_i = E[g_i|Y_i]$，从而更新第 i 个 SNP 的效应。

该新方法被称为 fBayesB。模拟结果表明，fBayesB 方法估计准确性仅比基于 MCMC 的 BayesB 方法低 0.011(s.e.0.005)，两者估计值偏离真值的方向相反，但两者相应的计算时间分别为 2~5 分钟和 47 小时。

4.3 Full hierarchical BayesA

在传统 BayesA 方法中，标记效应方差 $\sigma_{g_k}^2$ 的先验分布为逆卡方分布 $x^{-2}(v, S)$，其中自由度 v 和尺度参数 S 被给予一个不合适的固定值。这两个超参数 v 和 S 控制标记效应估计的压缩程度和强度，因此会对推断产生重要影响。

Jia 等[32]将 v 和 S 作为未知参数，利用数据信息，采用 Metropolis 算法进行估计，该新方法被称为 Full hierarchical BayesA。模拟研究表明，对于受多基因

控制的性状, 新模型要明显优于传统 BayesA, 且多性状分析的优势比单性状分析大。新模型多性状分析准确性比传统固定先验值模型提高 17~20 个百分点。鉴于性状的遗传结构未知, 在育种实践中应用该新方法是稳健的。

4.4　模型中加入剩余微效多基因效应

在利用 SNP 芯片估计基因组育种值时, 人们期望所有与性状相关的遗传变异都能被 SNP 捕获到, 但仍存在被遗漏的可能, 因此 Hayes 等[27]建议在模型中加入剩余微效多基因效应,

$$\mathbf{y}=\mathbf{1}\mu+\mathbf{a}+\sum_{j=1}^{q}\mathbf{z}_j g_j+\mathbf{e}$$

其中 $\mathbf{a}$ 为剩余微效多基因效应向量。此时 GEBV= $\hat{\mathbf{a}}+\sum_{j=1}^{q}\mathbf{z}_j\hat{g}_j$。

Solberg 等[33]研究了基于此模型的 GEBVs 准确性随世代改变问题, 其研究结果表明, 在直接的后续世代中, 模型中包含剩余微效多基因效应未对 GEBVs 估计准确性产生影响, 但随着世代的延伸, 模型中包含剩余微效多基因效应能保持高一点的准确性。另外, 模型中包含剩余微效多基因效应时, 其 GEBVs 的无偏性会得到显著改善, 并且跨世代稳定, 几乎一直保持在 0.98~1.00。特别对于较低密度的芯片, 模型中包含剩余微效多基因效应更为重要。

5　结 语

基因组育种值估计是基因组选择的重要环节。在估计基因组育种值的过程中, 由于标记数常常多于表型记录数, 即“大 *p*, 小 *n*”问题, 会导致多重共线性和过度参数化, 为解决这个问题一系列估计方法已被提出和研究。在这些方法中, 以 BLUP 和 Bayes 方法应用最为广泛。模拟数据和实际数据研究结果都表明, 总体上说 Bayes 方法估计 GEBVs 的准确性优于 BLUP 方法, 特别对于存在较大效应 QTL 的性状其优势更明显。但目前 Bayes 方法在实际育种中的运用不如 BLUP 方法普遍, 最主要原因可能在于 Bayes 方法计算效率相对较低, 另外可能由于 Bayes 方法过程相对复杂, 而人们更加熟悉 BLUP 方法。但随着其快速算法的开发和计算机硬件的改进, 计算问题有望得到解决。另外, 随着对基因组和性状遗传结构研究的深入开展, 将能为 Bayes 方法提供更为准确的先验信息, 从而使 Bayes 方法估计基因组育种值的准确性优势更加突出, Bayes 方法将会得到更为广泛的应用。

参考文献

[1] Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 2001, 157(4): 1819–1829. [DOI]

[2] Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE. Reducing dimensionality for prediction of genome-wide breeding values. *Genet Sel Evol*, 2009, 41(1): 29. [DOI]

[3] VanRaden PM. Efficient methods to compute genomic predictions. *J Dairy Sci*, 2008, 91(11): 4414–4423. [DOI]

[4] Zhang Z, Liu J, Ding X, Bijma P, de Koning DJ, Zhang Q. Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. *PLoS ONE*, 2010, 5(9): e12648. [DOI]

[5] Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, 2011, 12(1): 186. [DOI]

[6] Verbyla KL, Hayes BJ, Bowman PJ, Goddard ME. Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. *Genet Res* (*Camb*), 2009, 91(5): 307–311. [DOI]

[7] Yi N, Xu S. Bayesian LASSO for quantitative trait loci mapping. *Genetics*, 2008, 179(2): 1045–1055. [DOI]

[8] Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol*, 2005, 67(2): 301–320. [DOI]

[9] Gianola D, Fernando RL, Stella A. Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics*, 2006, 173(3): 1761–1776. [DOI]

[10] Long N, Gianola D, Rosa GJM, Weigel KA, Avendano S. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. *J Anim Breed Genet*, 2007, 124(6): 377–389. [DOI]

[11] Sun XC, Habier D, Fernando RL, Garrick DJ, Dekkers JCM. Genomic breeding value prediction and QTL mapping of QTLMAS2010 data using Bayesian Methods. *BMC Proceedings*, 2011, 5(Suppl. 3): S13. [DOI]

[12] 刘小磊, 杨松柏, Max F Rothschild, ZHANG Zhi-Wu, 樊斌. 利用紧缩线性模型和贝叶斯模型对猪总产仔数和产活仔数性状的全基因组关联研究. 遗传, 2012, 34(10):

1261–1270. [DOI]

[13] Fernando RL, Habier D, Stricker C, Dekkers JCM, Totir LR. Genomic selection. *Acta Agric Scand A Anim Sci*, 2007, 57(4): 192–195. [DOI]

[14] Gianola D, de los Campos G, Hill WG, Manfredi E, Fernando R. Additive genetic variability and the Bayesian alphabet. *Genetics*, 2009, 183(1): 347–363. [DOI]

[15] Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Series B Stat Methodol*, 1996, 58(1): 267–288. [DOI]

[16] Yuan M, Lin Y. Efficient empirical Bayes variable selection and estimation in linear models. *J Am Stat Assoc*, 2005, 100(472): 1215–1225. [DOI]

[17] Park T, Casella G. The Bayesian Lasso. Technical report. Gainesville, FL: University of Florida, 2008. [DOI]

[18] Usai MG, Goddard ME, Hayes BJ. LASSO with cross- validation for genomic selection. *Genet Res* (*Camb*), 2009, 91(6): 427–436. [DOI]

[19] Lund MS, Sahana G, de Koning DJ, Su G, Carlborg O. Comparison of analyses of the QTLMAS XII common dataset. I: Genomic selection. *BMC Proceedings*, 2009, 3(Suppl. 1): S1. [DOI]

[20] Szydlowski M, Paczyńska P. QTLMAS 2010: simulated dataset. *BMC Proceedings*, 2011, 5(Suppl. 3):S3. [DOI]

[21] Bastiaansen JWM, Bink MCAM, Coster A, Maliepaard C, Calus MPL. Comparison of analyses of the QTLMAS XIII common dataset. I: genomic selection. *BMC Proceedings*, 2010, 4(Suppl. 1): S1. [DOI]

[22] Pszczola M, Strabel T, Wolc A, Mucha S, Szydlowski M. Comparison of analyses of the QTLMAS XIV common dataset. I: genomic selection. *BMC Proceedings*, 2011, 5(Suppl. 3): S1. [DOI]

[23] Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA. The impact of genetic architecture on Genome-wide evaluation methods. *Genetics*, 2010, 185(3): 1021–1031. [DOI]

[24] Moser G, Tier B, Crump RE, Khatkar MS, Raadsma HW. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genet Sel Evol*, 2009, 41(1): 56. [DOI]

[25] Chen CY, Misztal I, Aguilar I, Tsuruta S, Meuwissen THE, Aggrey SE, Muir WM. Genome wide marker assisted selection in chicken: making the most of all data, pedigree, phenotypic,and genomic in a simple one step procedure. In: Proceeding of the 9th world congress on genetics applied to livestock production. Germany. 2010, 0288. [DOI]

[26] Habier D, Tetens J, Seefried F-R, Lichtner P, Thaller G. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genet Sel Evol*, 2010, 42(1): 5. [DOI]

[27] Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: Genomic selection in dairy cattle: progress and challenges. *J Dairy Sci*, 2009, 92(2): 433–443. [DOI]

[28] VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS. Invited review: Reliability of genomic predictions for North American Holstein bulls. *J Dairy Sci*, 2009, 92(1): 16–24. [DOI]

[29] Grisart B, Farnir F, Karim L, Cambisano N, Kim J-J, Kvasz A, Mni M, Simon P, Frere JM, Coppieters W, Georges M. Genetic and functional confirmation of the causality of the *DGAT1 K232A* quantitative trait nucleotide in affecting milk yield and composition. *Proc Natl Acad Sci USA*, 2004, 101(8): 2398–2403. [DOI]

[30] Hayashi T, Iwata H. EM algorithm for Bayesian estimation of genomic breeding values. *BMC Genetics*, 2010, 11: 3. [DOI]

[31] Meuwissen THE, Solberg TR, Shepherd R, Woolliams JA. A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. *Genet Sel Evol*, 2009, 41(1): 2. [DOI]

[32] Jia Y, Jannink J-L. Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics*, 2012, 192(4): 1513–1522. [DOI]

[33] Solberg TR, Sonesson AK, Woolliams JA, Ødegard J, Meuwissen THE. Persistence of accuracy of genome-wide breeding values over generations when including a polygenic effect. *Genet Sel Evol*, 2009, 41(1): 53. [DOI]

(责任编委：杜立新)