中等稀疏条件下基因交互作用的两步 Bayes 方法

杨青龙¹, 刘媛², 刘妍岩^{2*}, 邹珺³

1. 中南财经政法大学统计与数学学院, 武汉 430073;

2. 武汉大学数学与统计学院, 武汉 430072;

3. 华中农业大学植物科学技术学院华中农业大学作物遗传改良国家重点实验室, 武汉 430070

E-mail: yangqinglong@zuel.edu.cn, larin_27@163.com, liuyy@whu.edu.cn, zoujun@mail.hzau.edu.cn

收稿日期: 2019-10-09; 接受日期: 2020-03-31; 网络出版日期: 2020-11-17; * 通信作者

国家自然科学基金 (批准号: 11971362, 11971324, 11771366 和 11671131) 和中南财经政法大学中央高校基本科研业务费 (批准号: 2722020JCT030) 资助项目

摘要 植物遗传与基因组研究表明许多重要的农艺性状有影响的基因位点不是稀疏的, 受到大量微效基因的影响, 并且还存在基因交互项的影响. 本文基于重要油料作物油菜的花期数据, 研究中等稀疏条件下的基因选择问题, 提出了一种两步 Bayes 模型选择方法. 考虑基因间的交互作用, 模型的维数急剧增长, 加上数据结构特别, 通常的变量选择方法效果不好. 本文提出两步变量选择的方法: 首先利用 Kolmogorov 特征扫描方法剔除那些明显不重要的变量, 达到降维的目的; 其次, 在选出的位点中考虑交互作用. 为了克服 Bayes 方法计算速度慢的问题, 本文在模型中引入指示变量, 通过估计指示变量的后验分布选择模型. 模拟结果表明本文提出的方法在预测精度和计算稳定性上有良好的表现, 与不加指示变量的 Bayes 方法相比, 在预测精度上有很大的提高. 最后, 利用本文提出的方法分析一个油菜花期数据, 发现了一些交互效应的基因位点.

关键词 交互作用 高维数据 基因组选择 Bayes 方法 油菜花期

MSC (2010) 主题分类 62C10, 62J07

1 引言

在动植物育种研究中, 基因组选择 (genomic selection) 正在发挥重要作用. 基因组选择是一种基于全基因组范围内的大量标记信息, 通过统计模型预测个体在特定性状上的表现, 并加以选择育种的方法 (参见文献 [1]). 准确揭示基因型与表现型之间的关系, 不仅可以提供相对准确的育种方案, 加速育种, 减少世代间隔, 而且可以通过基因指导配偶选择, 更好地利用现有遗传资源, 达到提高遗传增益速率的目的 (参见文献 [2]). 基于全基因组分子标记, 用基因组选择来加速育种循环和根据关联映射 (association mapping) 寻找数量性状位点 (quantitative trait locus, QTL) 是统计方法在生物学应用中比较成功的两个案例. 数量性状是指个体的表现性状为连续变化的, 如植物开花时间、植株高度和种

英文引用格式: Yang Q L, Liu Y, Liu Y Y, et al. Two-step Bayesian method for detecting gene-gene interactions with moderate sparsity (in Chinese). Sci Sin Math, 2021, 51: 393–410, doi: 10.1360/SSM-2019-0243

子含油量、奶牛的产奶量和脂肪含量等 (参见文献 [3]). 随着基因技术的高速发展, 动物可获取的基因相关数据有基因表达值和基因型等数据, 基因表达值为连续型变量, 基因型数据根据实际情况分为显性纯合子、杂合子和隐性纯合子. 对于植物而言, 基因型数据与动物有较大差异. 在植物基因组选择问题中, 植物基因型可进行人为控制, 研究对象可为纯合体也可杂合体, 通过花药培养等技术可以快速得到大量的纯合个体, 如油菜的双单倍体 (double haploid, DH) 群体. 此外, 动物性状通常受少量位点的影响, 属于稀疏问题, 而影响植物性状的有很多位点并且效应都不强, 属于中等稀疏问题. 此外, 植物性状还受到许多基因与基因之间交互效应的影响. 由于基因标记数目远多于植株数目, 若进一步考虑交互效应将使得备选变量维数非常高, 属于超高维变量选择问题, 加上植物性基因数据的特殊结构, 给统计分析带来了挑战.

目前, 基于全基因组关联分析 (genome-wide association studies, GWAS) 寻找对性状有显著贡献的位点和根据关联映射寻找基因变异点是两种最常用的方法. 例如, Meuwissen 等^[4] 最先提出使用基因组选择方法对动物和植物进行育种. 同时, 基因组选择方法也被应用到多种植物的 QTL 定位上, 如大米、油菜和玉米等 (参见文献 [5–8]). 尽管 GWAS 方法计算简单, 但是由于样本量通常不会太大, 导致其效率和精度均不够理想. 对于高维数据的变量选择问题, 统计学中已有大量比较成熟的方法, 如惩罚函数的方法 (LASSO (least absolute shrinkage and selection operator)^[9]、SCAD (smoothly clipped absolute deviation)^[10] 和 EN (elastic net)^[11] 等)、Bayes 方法 (如 BayesA 和 BayesB) 和 GBLUP (genomic best linear unbiased prediction)^[12] 等. 惩罚的方法也被应用到 GWAS 中 (如文献 [13–15]), 但基于惩罚的方法存在如下缺点: 第一, 基于惩罚方法的变量选择最多只能选出跟样本个数同样多的变量, 由于植物基因选择是中等稀疏的, 因此, 惩罚方法有可能漏掉一些对性状有影响的变量; 第二, 当自变量存在较强的相关性时, 惩罚方法表现不够令人满意, 甚至不如 GBLUP; 第三, 对于超高维变量选择问题, 惩罚方法在计算精度、收敛速度和稳定性方面存在明显的缺陷. 因此, 一些研究者提出通过将不重要的变量剔除实现降维, 然后再进行变量选择的方法. 这方面的工作首先是由 Fan 和 Lv^[16] 提出的边际筛选 (iterative sure independent screening, ISIS) 方法, 利用边际相关性过滤一些不重要变量, 从而实现降维的目的. 许多学者随后通过定义不同的边际相关性度量指标对 SIS (sure independent screening) 方法进行了拓广. 例如, Chang 等^[17] 使用边际经验似然比对变量重要性进行排序. Fan 等^[18] 将 SIS 推广到非参数模型. He 等^[19] 提出分位数自适应筛选 (quantile-adaptive model-free variable screening, QA) 方法. Li 等^[20] 提出用 Kendall's τ 相关系数来代替传统的 Pearson 相关系数以提高稳健性, 该方法在具有单调联系方程的半参数单指标模型中效果显著. 距离相关系数指标筛选 (feature screening via distance correlation learning, DCS)^[21] 用距离相关系数替代 Pearson 相关系数来进行指标筛选. KF (Kolmogorov filter)^[22] 和融合 KF^[23] 方法用 Kolmogorov 检验统计量来筛选变量. He 和 Lin^[24] 结合 ISIS 和惩罚方法提出了多步的 GWAS 方法. 近期一些学者将再生核 Hilbert 空间方法 (reproducing kernel Hilbert spaces, RKHS)^[12] 和机器学习方法^[25] 应用到基因组选择中. 例如, Da 等^[26] 建立了包含加性效应和显性关系的线性混合模型, 将模型分为 $n > p$ 和 $p > n$ 两种情形, 并在预测过程中, 使用了 GBLUP 和限制似然估计方法, 并对两者的预测精度进行阐述. Xu 等^[5] 基于线性模型对水稻进行基因组选择, 并将 GBLUP、SSVS (stochastic search variable selection) 与 LASSO 进行比较, 结果表明 LASSO 方法的预测效果最好, 其次是 GBLUP, SSVS 预测效果最差. Würschum 等^[7] 基于线性模型运用岭回归最佳线性无偏预测和 BayesB 方法对甘蓝型油菜的开花期和植株高度等 6 个性状进行预测. 基因组选择中的线性混合模型通常假设所有基因方差相同, Speed 和 Balding^[27] 对模型进行拓展, 允许基因效应存在异方差, 并运用限制极大似然方法进行估计以得到 GBLUP.

上述研究大多没有考虑基因交互作用的影响, 而基因交互作用被广泛认为对表型性状的遗传变异

具有不可忽略的影响. 如果在模型中引入交互项, 使得变量维数急剧增加, 已有的统计方法会存在计算精度低、收敛性不佳以及不稳定等问题, 因此降维是重点. 目前关于基因交互作用影响的研究较少. 已有的研究主要是利用两步或者多步方法. 例如, Yang 等^[28]利用自适应 LASSO 方法选择上位效应, Yi 等^[29]基于广义 Bayes 模型选择上位效应, 不过这两种方法都假设响应变量是二分变量, 不适用于数量特征的数据. Hao 和 Zhang^[30]采用 iFOR 贪婪前向选择方法, 提出的 iFORT 两步方法, 第一步向前筛选主效应基因, 第二步筛选基因间交互作用. Li 等^[31]用变量筛选方法先做单基因 (加性效应或者显性效应) 筛选, 再做单基因交互项的筛选, 最后做上位交互作用的选择. Kong 等^[32]在第一步先对自变量和响应变量平方处理, 再用距离相关法筛选自变量, 第二步用 G LASSO (graphical LASSO) 选择已筛出基因间的交互项. 由于植物纯合个体基因型数据的自变量是二分型 $(-1, 1)$ 的, 上述迭代方法在应用的时候不稳定, 从而增加了计算的困难. 目前, 针对植物基因数据的交互作用的选择方法十分有限. Li 等^[6]结合岭回归最佳线性无偏估计、再生核 Hilbert 空间、Bayes 方法 (BayesLASSO、BayesA 和 BayesB)、随机森林和支持向量回归方法对单环境下油菜花期数据进行分析, 并运用交叉验证在多种环境下对以上方法进行比较. 他们基于随机森林方法对变量进行重要性排序, 筛选前 10% 位点用两两交互作用构建模型. 随机森林方法倾向于高估相关自变量的重要性, 筛选出的位点准确性有待考察.

本文研究基于甘蓝型油菜双亲本杂交所衍生的双单倍体遗传分离群体及其亲本的基因型和花期数据 (参见文献 [33, 34]), 该数据包含 182 株 TN-DH 系的 2,041 个位点基因型信息, 基因型数据以 A 和 B 标记, 分别表示子代与父本或母本基因型相同 (父本和母本都是纯合子), 在实际分析过程中会用数值 -1 和 1 代替. 该数据集属于典型的小 n 大 p 问题, 基因位点个数远远多于样本数. 开花期是油菜重要的生长发育性状, 与油菜的产量和品质等重要性状都密切相关, 研究者普遍认为与该性状有关的基因不是稀疏的, 并且基因的交互作用与该性状相关. 直接考虑交互作用将会有四百多万备选变量, 并且备选变量均是二分变量, 已有的方法很难直接应用. 在此背景下, 针对具有连续型的数量特征和二分协变量的数据, 假设交互作用存在中等稀疏问题, 本文提出两阶段 Bayes 方法建立预测模型. 第一步采用适合离散型自变量的独立筛选方法 (参见文献 [23]) 筛除不重要的基因位点, 然后基于保留的位点及两两之间的交互项利用 Bayes 方法建立预测模型.

由于 Bayes 方法在变量维数超高的情形下存在计算速度慢且预测精度不高的缺点, 我们在模型中对每个备选项引入指示变量, 利用 Bayes 方法估计指示变量的后验分布, 基于预测的后验概率来选择模型. 相比已有的方法, 所提出的方法的优势在于: 第一, 适用于中等稀疏下的离散型基因数据; 第二, 本文中模型考虑了交互项, 通过加入指示变量实现模型选择和估计; 第三, 引入指示变量使得计算变得更加简单快速, 同时进行变量选择. 模拟计算的结果表明所提出的方法稳定性好并且容易实现.

2 两步 Bayes 方法

本节详细介绍两步 Bayes 方法, 首先运用融合 KF 指标扫描方法对基因位点进行筛选, 然后基于保留的基因位点分别建立带指示变量的模型, 通过指示变量的后验分布实现变量选择, 在参数估计过程中剔除对表现型无影响的变量, 预测精度更高.

2.1 第一步: Kolmogorov 扫描

假设 Y 表示表现型特征, $X = (X_1, X_2, \dots, X_q)$ 为基因型数据, 共包含 q 个基因位点. 根据基因

型与父本或母本一致, X_j 分别取值为 1 或 -1. 若给定 $X_j = 1$ 和 $X_j = -1$ 的条件下, Y 的条件分布完全相同, 则认为 X_j 与 Y 是相互独立的. 令 $F_{+j}(y)$ 和 $F_{-j}(y)$ 分别表示给定 $X_j = 1$ 和 $X_j = -1$ 条件下, Y 的条件分布函数, 根据统计量 $K_j = \sup_{-\infty \leq y \leq \infty} |F_{+j}(y) - F_{-j}(y)|$ 确定基因位点和表现型的相关关系, 统计量越大, 则认为该基因位点越重要 (参见文献 [23]). 当基于样本进行计算时, 该统计量计算公式为

$$K_{nj} = \sup_y |\hat{F}_{+j}(y) - \hat{F}_{-j}(y)|,$$

其中 $\hat{F}_{+j}(y)$ 和 $\hat{F}_{-j}(y)$ 分别表示 $F_{+j}(y)$ 和 $F_{-j}(y)$ 的估计. 依据 K_{nj} 的取值从大到小对基因位点进行排序, 截取前 p 个进行建模, 其他基因位点认为是对表现型无影响的位点予以剔除.

2.2 第二步: Bayes 模型选择

基于上述指标筛选保留的 p ($p < q$) 个位点基因型数据和表现型数据建立模型. 本文认为植株的表现型不仅受到单个位点效应的影响, 也受到两两之间交互作用的影响. 根据以上设定, 假设表现型和基因型数据满足以下模型:

$$y_i = \mu + \sum_{j=1}^p x_{ij}\beta_j + \sum_{k=1}^M a_{ik}\alpha_k + e_i, \quad (2.1)$$

其中 μ 表示植株表现型的均值, y_i 代表第 i 棵植株的表现型, β_j 表示第 j 个位点的基因效应, x_{ij} 表示第 i 棵植株第 j 个位点的基因型, 取值为 -1 和 1, 分别表示与父本或母本表现型一致, $i = 1, 2, \dots, n$, $j = 1, 2, \dots, p$. 为了便于表述, 将基因位点之间两两交互的效应编号为 $1, 2, \dots, M$, 其中 $M = C_p^2$, a_{ik} ($k = 1, 2, \dots, M$) 表示第 i 棵植株的第 k 个交互项, 取值为 1 和 -1, 分别表示两个位点是否与同一亲本相同. e_i ($i = 1, 2, \dots, n$) 表示模型中的随机扰动项, 服从正态分布 $N(0, \sigma_e^2)$.

基因组选择问题属于高维回归问题, 一般而言基因位点数大于植株数, 即 $p > n$, 而模型中加入了基因交互作用, 使得变量维数远高于样本个数. 如何定义一个阈值确定重要的变量是关键问题, 以往的研究多是通过经验, 但是对于中等稀疏并且效应不高的变量, 很容易被剔除. 为解决此问题, 我们在模型中加入示性变量进行变量选择. 如果该变量取值为 1, 则表示对应的效应对响应变量有影响; 否则, 取值为 0 表示该效应与响应变量无关. 改进后的模型如下:

$$y_i = \mu + \sum_{j=1}^p x_{ij}\beta_j\gamma_j + \sum_{k=1}^M a_{ik}\alpha_k\eta_k + e_i, \quad i = 1, \dots, n, \quad j = 1, \dots, p, \quad (2.2)$$

其中 γ_j 表示第 j 个位点的示性变量, 即当第 j 个位点包含在模型中时, γ_j 取值为 1, 否则取 0; η_k 表示第 k 个交互效应的示性变量, 当第 k 项交互效应包含在模型中时, η_k 取值为 1, 否则为 0.

上面的模型表示成矩阵形式如下:

$$Y = \mu \mathbf{1}_n + \sum_{j=1}^p X_j \beta_j \gamma_j + \sum_{k=1}^M a_k \alpha_k \eta_k + e, \quad (2.3)$$

其中 $Y = (y_1, y_2, \dots, y_n)^T$, $\mathbf{1}_n = (1, 1, \dots, 1)^T$, $X_j = (x_{1j}, x_{2j}, \dots, x_{nj})^T$, $e = (e_1, e_2, \dots, e_n)^T$, $a_k = (a_{1k}, a_{2k}, \dots, a_{nk})^T$.

定义 $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_M)^T$, $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T$, $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_p)^T$, $\eta = (\eta_1, \eta_2, \dots, \eta_M)^T$, 采用的先验分布为

$$\begin{aligned}\mu &\propto p(1), \\ \beta &\sim N(0, \sigma^2 I_p), \quad p(\sigma^2) \propto \frac{1}{\sigma^2}, \\ \alpha &\sim N(0, \sigma_a^2 I_M), \quad p(\sigma_a^2) \propto \frac{1}{\sigma_a^2}, \\ \gamma_j &\sim B(1, a), \quad a \sim U(0, 1), \\ \eta_k &\sim B(1, b), \quad b \sim U(0, 1), \\ p(\sigma_e^2) &\propto \frac{1}{\sigma_e^2}.\end{aligned}$$

此时, 可以得到表现型的期望和方差如下:

$$\begin{aligned}E(Y \mid \mu, \beta, \sigma^2, \alpha, \sigma_a^2, \sigma_e^2, \gamma, \eta) &= E\left(\mu \mathbf{1}_n + \sum_{j=1}^p X_j \beta_j \gamma_j + \sum_{k=1}^M a_k \alpha_k \eta_k + e \mid \mu, \beta, \sigma^2, \alpha, \sigma_a^2, \sigma_e^2, \gamma, \eta\right) \\ &= \mu \mathbf{1}_n + \sum_{j=1}^p X_j \beta_j \gamma_j + \sum_{k=1}^M a_k \alpha_k \eta_k, \\ \text{Var}(Y \mid \mu, \beta, \sigma^2, \alpha, \sigma_a^2, \sigma_e^2, \gamma, \eta) &= \text{Var}\left(\mu \mathbf{1}_n + \sum_{j=1}^p X_j \beta_j \gamma_j + \sum_{k=1}^M a_k \alpha_k \eta_k + e \mid \mu, \beta, \sigma^2, \alpha, \sigma_a^2, \sigma_e^2, \gamma, \eta\right) \\ &= \text{Var}(e) \\ &= \sigma_e^2 I_n.\end{aligned}$$

在模型 (2.3) 的框架下, 表现型 Y 的条件分布为正态分布

$$N\left(\mu \mathbf{1}_n + \sum_{j=1}^p X_j \beta_j \gamma_j + \sum_{k=1}^M a_k \alpha_k \eta_k, \sigma_e^2 I_n\right).$$

根据 Bayes 公式, 可以得到所有参数的联合后验分布:

$$\begin{aligned}f(\mu, \beta, \sigma^2, \alpha, \sigma_a^2, \sigma_e^2, \gamma, \eta \mid y) \\ \propto f(y \mid \mu, \beta, \sigma^2, \alpha, \sigma_a^2, \sigma_e^2, \gamma, \eta) f(\beta \mid \sigma^2) f(\sigma^2) f(\alpha \mid \sigma_a^2) f(\sigma_a^2) f(\sigma_e^2) f(\gamma \mid a) f(\eta \mid b),\end{aligned}$$

这里用 $f(a \mid b)$ 表示给定 b 的条件下, a 的条件密度函数.

接下来, 可以利用 Gibbs 抽样构造 Markov 链给出参数的估计, 具体步骤如下:

第 1 步 指定各参数初始值;

第 2 步 更新 μ ,

$$\mu \sim N\left(\frac{\mathbf{1}_n^T(Y - \mu \mathbf{1}_n - \sum_{j=1}^p x_j \beta_j \gamma_j - \sum_{k=1}^M a_k \alpha_k \eta_k)}{n}, \frac{\sigma_e^2}{n}\right);$$

第 3 步 更新 β_j , $j = 1, 2, \dots, p$, 如果 $\gamma_j = 0$, 从先验分布中生成 β_j , 否则,

$$\beta_j \sim N\left(\frac{X_j^T(Y - \mu \mathbf{1}_n - \sum_{m \neq j} x_m \beta_m \gamma_m - \sum_{k=1}^M a_k \alpha_k \eta_k) \sigma^2}{\sigma_e^2 + \sigma^2 X_j^T X_j}, \frac{\sigma^2 \sigma_e^2}{\sigma_e^2 + \sigma^2 X_j^T X_j}\right);$$

第 4 步 更新 $\alpha_k, k = 1, 2, \dots, M$, 如果 $\eta_k = 0$, 从先验分布中生成 α_k , 否则,

$$\alpha_k \sim N\left(\frac{a_j^T(Y - \mu \mathbf{1}_n - \sum_{j=1}^p x_j \beta_j \gamma_j - \sum_{t \neq k} a_k \alpha_k \eta_k) \sigma_a^2}{\sigma_e^2 + \sigma_a^2 a_k^T a_k}, \frac{\sigma_a^2 \sigma_e^2}{\sigma_e^2 + \sigma_a^2 X_j^T X_j}\right);$$

第 5 步 更新 σ_e^2 ,

$$\sigma_e^2 \sim IG\left(\frac{n}{2}, \frac{\|Y - \mu \mathbf{1}_n - \sum_{j=1}^p x_j \beta_j \gamma_j - \sum_{k=1}^M a_k \alpha_k \eta_k\|^2}{2}\right);$$

第 6 步 更新 $\sigma^2, \sigma^2 \sim IG(\frac{p}{2}, \frac{\sum_{j=1}^p \beta_j^2}{2})$;

第 7 步 更新 $\sigma_a^2, \sigma_a^2 \sim IG(\frac{M}{2}, \frac{\sum_{k=1}^M \eta_k^2}{2})$;

第 8 步 更新 $\gamma_j, j = 1, 2, \dots, p, \gamma_j \sim B(1, \frac{a_{1j}}{a_{0j} + a_{1j}})$, 其中

$$a_{1j} = \exp\left\{\frac{\|Y - \mu \mathbf{1}_n - \sum_{m \neq j} x_m \beta_m \gamma_m - \sum_{k=1}^M a_k \alpha_k \eta_k - x_j \beta_j\|^2}{2\sigma_e^2}\right\} p,$$

$$a_{0j} = (1 - p) \exp\left\{\frac{\|Y - \mu \mathbf{1}_n - \sum_{m \neq j} x_m \beta_m \gamma_m - \sum_{k=1}^M a_k \alpha_k \eta_k\|^2}{2\sigma_e^2}\right\};$$

第 9 步 更新 $\eta_k, k = 1, 2, \dots, M, \eta_k \sim B(1, \frac{b_{1k}}{b_{0k} + b_{1k}})$, 其中

$$b_{1k} = \exp\left\{\frac{\|Y - \mu \mathbf{1}_n - \sum_{j=1}^p x_j \beta_j \gamma_j - \sum_{t \neq k} a_t \alpha_t \eta_t - a_k \eta_k\|^2}{2\sigma_e^2}\right\} q,$$

$$b_{0k} = (1 - q) \exp\left\{\frac{\|Y - \mu \mathbf{1}_n - \sum_{j=1}^p x_j \beta_j \gamma_j - \sum_{t \neq k} a_t \alpha_t \eta_t\|^2}{2\sigma_e^2}\right\}.$$

2.3 评价指标

借鉴二分类问题研究中的评价指标 (参见文献 [35]), 在单次指标筛选模拟过程中假设基因位点分为有效应和无效应, 指标筛选过程中基因位点被选入模型即预测该位点有效应, 未被选入模型即认为该位点没有效应. 在指标筛选过程中共有以下 4 种情况:

TP: 有效应的基因位点被选入模型;

FN: 有效应的基因位点未被选入模型;

TN: 没有效应的基因位点未被选入模型;

FP: 没有效应的基因位点被选入模型,

其中 T、F、N 和 P 分别表示 T = True、F = False、P = Positive 和 N = Negative.

准确率 (precision) 表示正确预测为有效应基因位点的个数占全部预测为有效应基因位点个数的比例, 召回率 (recall rate) 表示正确预测为有效应基因位点的个数占全部有效应基因位点个数的比例, 定义分别为

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}.$$

指标筛选的目标是初步对高维数据进行降维, 把一部分不重要的指标从模型中过滤掉, 与变量选择方法不同的是, 指标筛选保留下的指标有重要的也有不重要的, 因此, 准确率中的 FP 会偏大, 从而导致准确率较低. 在指标筛选过程中, 我们更关注的是有效应的基因位点在经过指标筛选之后是否被选进模型, 也就是说, 相对于准确率, 更加关注召回率 R .

进一步, 为了考察指标筛选方法, 本文在多次模拟过程中定义完整率指标. 完整率表示所有真实位点全部被选进模型占总模拟次数的比例, 假设共模拟 N 次, 第 j ($j = 1, 2, \dots, N$) 次模拟有 TP_j 个有效应的基因位点被选入模型, FN_j 个有效应的基因位点未被选入模型, TN_j 个没有效应的基因位点未被选入模型, FP_j 个没有效应的基因位点被选入模型. 3 个指标的计算公式依次为

$$\text{完整率} = \frac{\sum_{j=1}^N \mathbf{1}_{(FN_j=0)}}{N}, \quad \text{召回率} = \frac{\sum_{j=1}^N TP_j}{\sum_{j=1}^N (TP_j + FN_j)}, \quad \text{准确率} = \frac{\sum_{j=1}^N TP_j}{\sum_{j=1}^N (TP_j + FP_j)}.$$

在农作物基因数据研究中, 真实值和预测值的相关系数经常用来评价模型预测的效果 (参见文献 [33]). 其计算公式为

$$\text{corr}(y, \hat{y}) = \frac{\text{cov}(y, \hat{y})}{\sqrt{\text{var}(y)\text{var}(\hat{y})}},$$

其中 y 是表现型真实值, $\hat{y}$ 是表现型预测值. 在模拟计算中, 我们采用上述相关系数指标评价方法的预测能力.

3 模拟计算

本节在不同参数设置下, 通过模拟计算评估所提出的方法的表现, 并与在模型中没有加入指示变量的 Bayes 方法进行了比较. 第一步均采用第 2.1 小节的 Kolmogorov 扫描方法进行降维, 第二步分别采用两种 Bayes 方法建立模型, 先验分布假定同第 2.2 小节.

基因型数据使用 R 包 `qtl` 生成, 根据每种参数设定下具体的非零基因比例和非零交互项比例生成系数. 由于指标筛选只对基因位点进行, 而交互项不会进行筛选, 因此在设定系数的过程中遵循单个基因位点效应较大, 而交互项效应较小的原则. 根据以上原则, 单个基因位点系数处于 $(0, 1)$ (或者 $(-1, 0)$) 之间的设置为 1 (或 -1), 基因交互项系数中, 大于 0.5 (或者小于 -0.5) 的设定为 0.5 (或 -0.5), 在具体的非零交互项比例设定下, 如果非零交互效应不足 10 个, 则随机选取 10 个为非零效应; 如果非零效应多于 100 个, 则随机选 100 个作为非零位点. 模型中的随机干扰项来自正态分布, 正态分布的方差在每个部分单独设定, 表现型数据根据模型设定计算得出. 下文依次在不同的保留指标个数、样本数和基因个数、基因位点和干扰项方差、非零交互项比例及非零位点比例设定下, 对指标筛选方法的表现进行模拟. 每种参数设定下模拟 50 次.

3.1 指标筛选精度评价

考虑指标筛选过程中保留指标个数与基因位点个数和样本个数有关, 当基因位点个数加倍时, 有效应位点的个数也加倍, 因此保留指标个数相应也需要加倍. 类似地, 对样本数进行推断, 当样本数增加时, 样本所包含的信息会增加, 但是样本数加倍的情况下, 保留指标个数缩短一半会极大减少保留指标个数, 导致漏掉部分有效应的基因位点, 因此取保留指标个数与取对数的样本数成反比. 因此, 指标筛选保留的指标个数以 $k \frac{q}{\log(n)}$ 为基础, k 分别取 1、2 和 3. k 取值越大, 指标筛选保留的基因位点个数越多, 模型越复杂, 计算量也相应增加, 将 k 不同取值下的指标筛选结果进行汇总, 可以得到表 1 所示的结果.

由表 1 可以看出, 随着 k 的增大, 完整率和召回率逐渐增大, 准确率逐渐变小. 在 50 次模拟过程中, 3 种参数设置下要把全部有效应位点选入模型需要的模型大小的中位数分别为 24、122 和 474. 当

$k = 1$ 时, 基因位点个数为 500 的召回率只有 54.80%; 当 $k = 3$ 时, 保留的指标数超过总基因个数的一半. 综合考虑模型的准确性和计算量, 在权衡各指标的前提下, 以下取 $k = 2$, 即第一步保留指标个数定为 $\frac{2q}{\log(n)}$.

不同样本数和基因个数设置下的结果见表 2, 可以看出, 当基因位点个数 q 固定时, 随着样本数 n 的增加, 完整率、召回率和准确率均呈现出逐渐增加的趋势. 当基因位点个数为 100 和样本量为 500 时, 完整率和召回率均为 100%, 说明此时有效应的基因位点可以全部被选出来.

当样本大小 n 固定时, 随着备选基因位点个数 q 的增加, 完整率、准确率和召回率均呈现出逐渐下降的趋势. 当样本大小固定时, 随着基因位点个数的增加, 完整率急剧下降, 说明当基因位点个数较多时, 指标筛选方法不能完全选出全部有效应的位点. 召回率下降幅度也较大, 样本数为 100 时, 召回率从 94.40% 下降到 79.50%; 样本数为 500 时, 召回率从 100% 下降到 97.10%. 由于指标筛选方法包含部分没有效应的位点, 因此准确率变化幅度相对较小.

综上所述, 当备选基因位点个数较少、样本量较多时, 指标筛选方法可以准确选出全部有效应的基因位点; 当备选基因位点个数较少、样本量也较少时, 指标筛选方法可在部分模拟过程中完全选出有效应位点, 且准确率较高; 当基因位点个数较多、样本数也较多时, 虽然不能在模拟中完全选出有效应的位点, 但从召回率看, 指标筛选方法能选出大多数有效应的基因位点; 当基因位点个数较多、样本量较少时, 指标筛选方法不能完全选出有效应的位点, 但从召回率看, 可以选出 80% 以上有效应的基因位点.

表 1 第一步保留不同指标个数 ($p = \frac{kq}{\log(n)}$) 时的模拟结果 ($n = 200$) ($\sigma^2 = 1$, $\sigma_e^2 = 1$, $P_1 = 10\%$, $P_2 = 1\%$)

k	q	p	M_0	完整率 (%)	召回率 (%)	准确率 (%)	D_0
1	100	19	190	34.00	91.20	48.00	24
	200	38	741	0.00	82.20	43.26	122
	500	95	4,560	0.00	54.80	28.84	474
2	100	38	741	78.00	97.80	25.74	24
	200	76	2,926	20.00	91.90	24.18	122
	500	189	17,955	0.00	73.44	19.43	474
3	100	57	1,653	94.00	99.40	17.44	24
	200	114	6,555	42.00	95.90	16.83	122
	500	284	40,470	0.00	85.44	15.04	474

P_1 : 非零基因位点比例; P_2 : 非零交互项比例; $p = \frac{kq}{\log(n)}$: 第一步指标筛选后留下的基因位点个数; M_0 : 加入基因交互项后模型中指标个数; D_0 : 多次模拟中包含全部非零效应最少要选择指标数的中位数.

表 2 第一步保留 $p = \frac{2q}{\log(n)}$ 个指标的模拟结果 ($\sigma^2 = 1$, $\sigma_e^2 = 1$, $P_1 = 10\%$, $P_2 = 1\%$)

q	n	p	M_0	完整率 (%)	召回率 (%)	准确率 (%)
100	100	44	990	54.00	94.40	21.46
	200	38	741	78.00	97.80	25.74
	500	33	561	100.00	100.00	30.30
200	100	87	3,828	4.00	79.50	18.28
	200	76	2,926	20.00	91.90	24.18
	500	65	2,145	54.00	97.10	29.88

表格注解同表 1

表 3 和 4 展示随机扰动项方差和基因位点效应方差相同和不同情形下模拟的结果. 从中可以看出, 当随机扰动项方差与基因位点效应方差相同时, 随着方差的逐渐增大, 完整率、召回率和准确率呈现出逐渐降低的趋势. 根据 3σ 原理, 当随机扰动项方差和基因位点效应方差为 0.01 时, 产生的随机数分布于 $(-0.3, 0.3)$, 由此可推断, 在此种设定下随机干扰项取值较小, 单个基因位点的效应取值全部为 1, 基因交互项效应取值来源于正态分布 $N(0, 0.01)$. 此种情形表明, 当基因位点效应较大、而基因交互项效应较小时, 指标筛选方法效果较好.

当随机扰动项方差和基因位点效应方差不相同时, 固定基因位点效应方差为 0.10, 分别取随机扰动项方差为 0.25、0.50、1.00 和 4.00, 随机扰动项取值 0.25 和 0.50 指标筛选结果相差不大. 随着随机扰动项方差的增大, 完整率、召回率和准确率逐渐降低, 与上文中变化规律相同, 完整率变化幅度最大, 其次是召回率, 准确率只有小幅下降.

对比随机干扰项方差和基因位点及交互项方差取值 (0.01, 0.01) 和 (0.10, 0.50) 两种情况, 可以发现两种设置下指标筛选方法精度相差不大. 考虑到基因位点效应和随机扰动项的取值范围, 在下一步模拟过程中设置随机干扰项方差和基因位点及交互项方差分别为 0.10 和 0.50.

表 3 基因效应方差与干扰项方差相等时指标筛选精度汇总表 ($P_1 = 10\%$, $P_2 = 1\%$)

q	$\sigma^2 = \sigma_e^2$	n	p	M_0	完整率 (%)	召回率 (%)	准确率 (%)
100	0.01	100	44	990	76.00	97.40	22.14
		200	38	741	98.00	99.80	26.26
		500	33	561	100.00	100.00	30.30
	0.10	100	44	990	70.00	96.60	21.95
		200	38	741	100.00	100.00	26.32
		500	33	561	100.00	100.00	30.30
	0.50	100	44	990	62.00	95.80	21.77
		200	38	741	92.00	99.20	26.11
		500	33	561	100.00	100.00	30.30
	1.00	100	44	990	54.00	94.40	21.45
		200	38	741	78.00	97.80	25.74
		500	33	561	100.00	100.00	30.30
200	0.01	100	87	3,828	12.00	88.70	20.39
		200	76	2,926	50.00	96.60	25.42
		500	65	2,145	100.00	100.00	30.77
	0.10	100	87	3,828	6.00	87.80	20.18
		200	76	2,926	50.00	96.50	25.39
		500	65	2,145	98.00	99.90	30.74
	0.50	100	87	3,828	2.00	85.40	19.63
		200	76	2,926	36.00	95.00	25.00
		500	65	2,145	90.00	99.50	30.62
	1.00	100	87	3,828	4.00	79.50	18.28
		200	76	2,926	20.00	91.90	24.18
		500	65	2,145	54.00	97.10	29.88

表格注解同表 1

表 4 基因效应方差与干扰项方差不等时指标筛选精度汇总表 ($\sigma^2 = 0.10, P_1 = 10\%, P_2 = 1\%$)

q	σ_e^2	n	p	M_0	完整率 (%)	召回率 (%)	准确率 (%)
100	0.25	100	44	990	72.00	96.80	22.00
		200	38	741	100.00	100.00	26.32
		500	33	561	100.00	100.00	30.30
	0.50	100	44	990	74.00	97.00	22.05
		200	38	741	100.00	100.00	26.32
		500	33	561	100.00	100.00	30.30
	1.00	100	44	990	64.00	96.00	21.82
		200	38	741	98.00	99.80	26.26
		500	33	561	100.00	100.00	30.30
	4.00	100	44	990	40.00	92.60	21.05
		200	38	741	96.00	99.60	26.21
		500	33	561	100.00	100.00	30.30
200	0.25	100	87	3,828	4.00	87.70	20.16
		200	76	2,926	52.00	96.60	25.42
		500	65	2,145	98.00	99.90	30.74
	0.50	100	87	3,828	12.00	87.70	20.16
		200	76	2,926	44.00	96.20	25.32
		500	65	2,145	98.00	99.90	30.74
	1.00	100	87	3,828	6.00	86.00	19.77
		200	76	2,926	40.00	95.50	25.13
		500	65	2,145	98.00	99.90	30.74
	4.00	100	87	3,828	2.00	85.00	19.54
		200	76	2,926	26.00	94.30	24.82
		500	65	2,145	96.00	99.70	30.68

表格注解同表 1

在选定基因位点非零比例为 10% 的前提下, 对基因交互项中的非零比例进行调整, 分别设置非零比例为 0.01、0.10、0.20 和 0.50, 同时设置最大值和最小值分别为 10 和 100. 其他参数根据上文中的最优结果进行设置, 在每种参数设置下模拟 50 次, 将结果进行汇总可以得到表 5.

从表 5 中数据可以得知, 当 q 等于 100、非零交互项比例小于等于 0.2 时, 非零交互项都为 10 项. 因此, 在以上几种设置中指标筛选结果没有差异. 在表 5 中对比 $n = 200$ 时的指标筛选结果可以发现, 随着非零交互项比例的增加, 完整率、召回率和准确率有轻微下降趋势. 整体来看, 非零交互项比例对指标筛选结果影响较小.

将不同基因位点非零比例下指标筛选精度进行汇总可以得到表 6 所示的结果. 从表 6 可以看出, 当基因位点非零比例为 40% 时, 所有参数设置下的完整率全部为 0. 随着非零基因位点比例的增加, 指标筛选的完整率和召回率有明显下降, 准确率显著提高. 当非零基因位点比例为 0.005 时, 基因位点个数 q 等于 100 的参数设置下完整率和召回率均为 100%, 但准确率极低, 说明指标筛选方法确定的保留指标个数较长, 包含太多没有效应的基因位点, 同时也说明指标筛选方法能将非零基因位点全部选出来.

表 5 不同非零交互项比例对模拟结果的影响比较结果 ($\sigma^2 = 0.10$, $\sigma_e^2 = 0.50$, $P_1 = 10\%$)

q	$P_2(\#)$	n	p	M_0	完整率 (%)	召回率 (%)	准确率 (%)
100	0.010(10)	100	44	990	74.00	97.00	22.05
		200	38	741	100.00	100.00	26.32
		500	33	561	100.00	100.00	30.30
	0.100(10)	100	44	990	74.00	97.00	22.05
		200	38	741	100.00	100.00	26.32
		500	33	561	100.00	100.00	30.30
	0.200(10)	100	44	990	74.00	97.00	22.05
		200	38	741	100.00	100.00	26.32
		500	33	561	100.00	100.00	30.30
	0.500(23)	100	44	990	68.00	96.60	21.96
		200	38	741	98.00	99.80	26.26
		500	33	561	100.00	100.00	30.30
200	0.010(10)	100	87	3,828	12.00	87.70	20.16
		200	76	2,926	44.00	96.20	25.32
		500	65	2,145	98.00	99.90	30.74
	0.100(19)	100	87	3,828	4.00	85.80	19.72
		200	76	2,926	36.00	95.70	25.18
		500	65	2,145	94.00	99.70	30.68
	0.200(38)	100	87	3,828	4.00	84.80	19.49
		200	76	2,926	32.00	94.60	24.90
		500	65	2,145	98.00	99.90	30.74
	0.500(95)	100	87	3,828	4.00	81.70	18.78
		200	76	2,926	8.00	91.30	24.03
		500	65	2,145	92.00	99.60	30.65

(#) 表示非零交互项的个数; 其他表格注解同表 1

3.2 模型预测效果

结合上述指标筛选结果, 第一步保留指标个数定为 $\frac{2q}{\log(n)}$, 第二步采用以下几种方法选择并且估计模型:

- (1) 提出的两步 Bayes 方法;
- (2) 依据模型 (2.1) 的 Bayes 方法, 先验分布与 (1) 相同;
- (3) LASSO 方法;
- (4) EN 方法^[11].

每种参数设置下模拟 100 次. 预留 10% 的样本做预测, 预测结果列在表 7.

从表 7 中可以看出, 当样本数 $n = 500$ 时, LASSO 方法预测效果最好, 其次是 EN 和我们的方法, 不带指示变量的 Bayes 方法预测效果最差. 当样本量达到 $n = 1,000$ 时, 本文提出的两步 Bayes 方法与 LASSO、EN 方法预测效果是可比的, 不加入指示变量的 Bayes 方法表现稍差一些. 从估计的精度上看, 我们的预测精度最好. 我们绘制基因位点个数为 100, 样本数为 500 和 1,000 的位点真实值和估

表 6 不同非零基因比例的模拟结果比较 ($\sigma^2 = 0.10, \sigma_e^2 = 0.50, P_2 = 1\%$)

q	P_1	n	p	M_0	完整率 (%)	召回率 (%)	准确率 (%)
100	0.005	100	44	990	100.00	100.00	11.36
		200	38	741	100.00	100.00	13.16
		500	33	561	100.00	100.00	15.15
	0.100	100	44	990	74.00	97.00	22.05
		200	38	741	100.00	100.00	26.32
		500	33	561	100.00	100.00	30.30
	0.200	100	44	990	0.00	83.70	38.05
		200	38	741	38.00	95.00	50.00
		500	33	561	96.00	99.80	60.49
	0.400	100	44	990	0.00	64.20	58.36
		200	38	741	0.00	71.00	74.74
		500	33	561	0.00	79.25	96.06
200	0.005	100	87	3,828	76.00	97.40	11.20
		200	76	2,926	94.00	99.40	13.08
		500	65	2,145	100.00	100.00	15.39
	0.100	100	87	3,828	12.00	87.70	20.16
		200	76	2,926	44.00	96.20	25.32
		500	65	2,145	98.00	99.90	30.74
	0.200	100	87	3,828	0.00	71.05	32.67
		200	76	2,926	0.00	80.10	42.16
		500	65	2,145	14.00	95.10	58.52
	0.400	100	87	3,828	0.00	54.55	50.16
		200	76	2,926	0.00	58.05	61.11
		500	65	2,145	0.00	66.88	82.31

表格注解同表 1

表 7 模型预测效果比较 ($\sigma^2 = 0.10, \sigma_e^2 = 0.50$)

预测精度	n	q	p	M_0	Bayes1	Bayes2	EN	LASSO
corr	500	100	33	561	0.7300	0.9680	0.9720	0.9740
		200	65	2,145	0.3398	0.4759	0.9751	0.9838
	1,000	100	29	435	0.9501	0.9757	0.9757	0.9759
		200	58	1,711	0.6228	0.9790	0.9853	0.9865
MSE	500	100	33	561	5.4900	0.7100	0.7200	0.8500
		200	65	2,145	25.4200	18.7300	0.9000	1.7800
	1,000	100	29	435	1.1100	0.5500	0.9700	0.9900
		200	58	1,711	13.6600	0.8600	1.0500	1.1800

Bayes1: 基于模型 (2.1) 的两步 Bayes 方法 (没有指示变量); Bayes2: 基于模型 (2.2) 的两步 Bayes 方法 (加入指示变量); 其他表格注解同表 1

计的差异, 可以得到图 1 和 2 所示的结果. 显然, LASSO 和 EN 低估了那些大效应, 两种 Bayes 方法估计精度都很好.

图 1 (网络版彩图) 参数估计结果 ($q = 100, n = 500$) 图 2 (网络版彩图) 参数估计结果 ($q = 200, n = 1,000$)

4 真实数据分析

本节基于华中农业大学提供的甘蓝型油菜 TN-DH 群体 182 个 DH 系在分布于全基因组上 2,041 个遗传位点的基因型数据和 10 个环境下的花期表型数据进行分析 (参见文献 [33]), 花期数据的环境信息如表 8 所示. TN-DH 群体来自欧洲冬油菜品种 Tapidor 和中国半冬性油菜品种 Ningyou7 号杂

交获得的 F1, 对 F1 进行小孢子培养和秋水仙碱加倍获得了包含有 202 个 DH 系的分离群体. 该群体在花期上具有丰富的遗传变异, 曾种植于中国春性、半冬性和冬性油菜种植环境, 获得了 10 个环境的花期数据 (参见文献 [33,34]). 对该群体 182 个 DH 系利用油菜 60K-SNP 芯片进行基因型分析获得了 13,000 多个高质量的 SNP (single nucleotide polymorphisms) 标记, 对其进行归类 and 遗传做图分析, 构建了分布于甘蓝型油菜全基因组包含 2,041 个遗传位点的遗传图谱. 本研究利用这 182 个株系在 2,041 个位点的基因型开展分析. 首先, 对数据进行基本介绍和质量控制; 然后对基因位点进行筛选, 考虑到实际数据的样本和基因位点相差悬殊, 保留 100 个位点进行建模; 最后, 基于保留的基因位点及其交互项分别运用两种 Bayes 方法进行预测, 并与 LASSO 和 EN 两种变量选择方法进行比较.

4.1 数据介绍

本文提出的方法将会应用到甘蓝型油菜花期数据中 (参见文献 [33,34]). 该油菜花期数据包括 2,041 个 SNP, 以及 Tapidor 与 Ningyou7 的 202 个杂交系. 油菜的基因型都是纯合子, 如果某个位点的 SNP 与 Tapidor 一致, 则记为 A, 与 Ningyou7 一致则记为 B. 该油菜花表型数据开花期来自 10 个不同的环境, 5 个不同的区域, 表 8 汇总了真实数据的环境信息.

在对油菜花数据进行分析之前, 先分为 3 个步骤对数据进行筛选:

- (1) 对于个体来说, 如果其所有的 SNP 超过 20% 是缺失的, 按照农业专家的建议从数据中剔除;
- (2) 对于每个位点来说, 如果超过 10% 的个体缺失该 SNP, 则应当从数据中剔除该位点;

(3) 对于单个位点来说, 如果其 A 的频率或者 B 的频率小于 0.05, 则应当从数据中剔除该位点. 经过数据筛选之后, 共保留 1,810 个样本, 2,041 个基因位点. 通过数据筛选的油菜表现型数据和基因型数据仍存在缺失数据, 所以需要对数据进行填充. 对于表现型数据, 用当前环境下的所有表现型的均值代替样本表现型数据的缺失值; 对于基因型数据, 与母本相同记为 1, 与父本相同记为 -1, 然后用 rrBLUP 包中的 A.mat 函数进行插补, 根据插补法得到结果, 小于 0 的值记为 -1, 大于 0 的值记为 1.

绘制 10 种不同环境下花期数据的分布, 可以得到图 3. 从箱形图可以看出, 生长在陕西 (环境编码为 N3-N7) 的油菜开花期较晚, 生长在浙江 (环境编码为 E7) 和湖北 (环境编码为 S3-S7) 的油菜开花期较早; 2007 年收获的油菜 (环境编码为 E7, N7, S7) 花期数据分布比较分散, 其次是 2004 年, 其他年份收获的油菜花期数据相对比较集中.

表 8 油菜花期数据考察的环境信息汇总表

地理环境	生长时间	环境编码
浙江省杭州市, 120°12'E/30°16'N/40 m	2006 年 10 月-2007 年 5 月	E7
陕西省大荔县, 109°56'E/34°52'N/800 m	2002 年 9 月-2003 年 6 月	N3
陕西省大荔县, 109°56'E/34°52'N/800 m	2003 年 9 月-2004 年 6 月	N4
陕西省大荔县, 109°56'E/34°52'N/800 m	2005 年 9 月-2006 年 6 月	N6
陕西省大荔县, 109°56'E/34°52'N/800 m	2006 年 9 月-2007 年 6 月	N7
湖北省武汉市, 114°21'E/30°37'N/40 m	2002 年 10 月-2003 年 5 月	S3
湖北省武汉市, 114°21'E/30°37'N/40 m	2003 年 10 月-2004 年 5 月	S4
湖北省江陵县, 113°25'E/30°30'N/40 m	2004 年 10 月-2005 年 5 月	S5
湖北省大冶市, 114°48'E/30°06'N/100 m	2005 年 10 月-2006 年 5 月	S6
湖北省武汉市, 114°21'E/30°37'N/40 m	2006 年 10 月-2007 年 5 月	S7

图 3 不同环境下花期数据分布范围

4.2 指标筛选和模型预测

根据实际数据预处理结果, 每种环境下保留 181 个样本和 2,041 个位点, 由于实际数据中, 基因位点个数较多, 而样本量较少, 考虑到模型的复杂度和计算量, 在筛选指标时选择保留 100 个位点建立模型. 经验证, 筛选出的基因位点 99.6% 具有非零基因效应, 10 种环境筛选出的共同基因有 8 个, 全部具有非零效应. 分别将每个环境下筛选得到的 100 个指标两两组合, 大约有 45.10% 的交互项具有非零效应, 10 种环境无共同交互效应, 说明交互效应受环境影响较大.

为了评价预测效果, 随机选取 80% (145 个) 的样本作为训练集建立模型, 剩下 20% (36 个) 作为测试集进行预测, 为保证结果稳定性, 共进行 20 次预测. 分别使用 Bayes 和惩罚变量选择方法 (LASSO 和 EN) 对模型进行预测, MCMC (Markov chain Monte Carlo) 方法设置 Markov 链的长度为 15,000, 丢掉的链长度为 5,000, 每 2 个值保留 1 个. 根据实际数据预测结果, 将 20 次建模结果进行平均, 结果见表 9.

表 9 真实数据预测: 相关系数 ($q = 2,041, n = 181, p = 100, M_0 = 5,050$)

环境	Bayes1	Bayes2	EN	LASSO
E7	0.4745	0.5796	0.5980	0.6052
N3	0.3854	0.5644	0.5712	0.5628
N4	0.5156	0.7124	0.7455	0.7298
N6	0.2928	0.5630	0.7118	0.6964
N7	0.3712	0.4561	0.4038	0.3760
S3	0.3221	0.6330	0.6879	0.6676
S4	0.5453	0.6427	0.6210	0.6133
S5	0.1526	0.5406	0.5446	0.5319
S6	0.2769	0.5660	0.6019	0.5968
S7	0.5065	0.5734	0.5921	0.5652
均值	0.3843	0.5831	0.6078	0.5945

表格注解同表 1 和 7

对比上表中 4 种方法预测下的相关系数可以发现, EN 预测效果最好, 其次是 LASSO 和带指示变量的 Bayes 方法, 不带指示变量的模型预测效果最差. 特别地, 在 N7 环境下, LASSO 和 EN 在 20 次预测过程中分别失效 6 次, 其他 14 次的相关系数平均值分别为 0.4038 和 0.3760, 低于带指示变量的 Bayes 方法. 综上可知, 本文提出的带指示变量的 Bayes 方法更稳健, 预测效果可与变量选择方法 (LASSO 和 EN) 媲美. 因此, 本文所提出的方法是一种比较有效的预测方法.

5 讨论和展望

本文主要研究了中等稀疏条件下基因组选择问题, 在建立模型过程中, 不仅考虑了基因位点对表现型数据的影响, 也考虑了位点两两之间交互项对表现型数据的影响. 基因组选择问题本来就属于高维回归问题, 加入基因交互效应后, 模型结构更复杂, 计算量更大. 为了减少计算量, 本文提出两步估计法, 在建模之前首先进行降维, 使用指标筛选方法对基因位点进行过滤, 滤掉部分不重要的位点, 进行建模预测. 考虑到植物基因型全部为纯合子、基因型数据取离散值的特点, 本文采取融合 KF 方法进行指标筛选, 以期达到降维的效果.

在进行指标过滤之后需要进行建模, 本文考虑加入指示变量和不加指示变量两种形式. 不加指示变量考虑了全部基因位点和交互项的微小效应, 加入指示变量相当于在模型参数估计过程中增加变量选择. 在参数估计过程中, 分别使用 MCMC 和变量选择方法 (LASSO 和 EN) 对参数进行估计, 并分别对表现型数据进行预测, 最后通过预测值与真实值的相关系数来评判预测效果.

本文提出的两步估计法预测效果较好. 文中大量模拟结果表明, 融合 KF 指标筛选方法能有效获取真实基因位点, 有效减少了模型计算量; 根据模型预测效果可以看出, 加入指示变量的模型优于不加指示变量的模型, 说明加入指示变量能有效剔除一部分不重要变量, 提高模型预测效果. 相对来讲, LASSO 和 EN 是预测效果较好的方法, 但一方面, 随着样本量的增加, 本文提出的模型使用 MCMC 方法估计参数的预测效果有显著提高, 当样本量足够多时, 预测效果超过 LASSO; 另一方面, LASSO 和 EN 在维数较高时存在失效的情况, 而本文提出的 Bayes 方法比较稳健, 综合考虑, 本文提出的方法是一种比较有效的预测方法. 考虑到加入交互项参数较多, 从减少计算量方面考虑, 本文假设基因位点方差相同. 至于异方差的情形, 本文所提出的方法可以在适当修改后仍可使用. 多个基因位点控制同一个性状是很普遍的现象, 高阶交互作用大概率是存在的, 加入高阶交互后模型维度会急剧增加, 考虑计算资源有限, 文中仅对二阶交互作用进行研究, 若计算资源能够支持, 推断本文方法预测效果会更优.

致谢 作者对编委和审稿人的宝贵意见表示衷心感谢.

参考文献

- 1 Lipka A E, Kandianis C B, Hudson M E, et al. From association to prediction: Statistical methods for the dissection and selection of complex traits in plants. *Curr Opin Plant Biol*, 2015, 24: 110–118
- 2 Daetwyler H D, Calus M P, Pong-Wong R, et al. Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. *Genetics*, 2013, 193: 347–365
- 3 Jiménez-Montero J A, González-Recio O, Alenda R. Comparison of methods for the implementation of genome-assisted evaluation of Spanish dairy cattle. *J Dairy Sci*, 2013, 96: 625–634
- 4 Meuwissen T H, Hayes B J, Goddard M E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 2001, 157: 1819–1829
- 5 Xu S, Zhu D, Zhang Q. Predicting hybrid performance in rice using genomic best linear unbiased prediction. *Proc Natl Acad Sci USA*, 2014, 111: 12456–12461

- 6 Li L, Long Y, Zhang L, et al. Genome wide analysis of flowering time trait in multiple environments via high-throughput genotyping technique in *Brassica napus* L. PLoS ONE, 2015, 10: e0119425
- 7 Würschum T, Abel S, Zhao Y, et al. Potential of genomic selection in rapeseed (*Brassica napus* L.) breeding. Plant Breed, 2014, 133: 45–51
- 8 Bernardo R, Yu J. Prospects for genomewide selection for quantitative traits in maize. Crop Sci, 2007, 47: 1082–1090
- 9 Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B Stat Methodol, 1996, 58: 267–288
- 10 Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Amer Statist Assoc, 2001, 96: 1348–1360
- 11 Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol, 2005, 67: 301–320
- 12 De Los Campos G, Gianola D, Rosa G J M, et al. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet Res, 2010, 92: 295–308
- 13 Ayers K L, Cordell H J. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. Genet Epidemiol, 2010, 34: 879–891
- 14 Cho S, Kim H, Oh S, et al. Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. BMC Proc, 2009, 3: S25
- 15 Wu Z, Zhao H. Statistical power of model selection strategies for genome-wide association studies. PLoS Genet, 2009, 5: e1000582
- 16 Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Ser B Stat Methodol, 2008, 70: 849–911
- 17 Chang J, Tang C Y, Wu Y. Marginal empirical likelihood and sure independence feature screening. Ann Statist, 2013, 41: 2123–2148
- 18 Fan J, Feng Y, Song R. Nonparametric independence screening in sparse ultra-high-dimensional additive models. J Amer Statist Assoc, 2011, 106: 544–557
- 19 He X, Wang L, Hong H G. Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. Ann Statist, 2013, 41: 342–369
- 20 Li G, Peng H, Zhang J, et al. Robust rank correlation based screening. Ann Statist, 2012, 40: 1846–1877
- 21 Li R, Zhong W, Zhu L. Feature screening via distance correlation learning. J Amer Statist Assoc, 2012, 107: 1129–1139
- 22 Mai Q, Zou H. The Kolmogorov filter for variable screening in high-dimensional binary classification. Biometrika, 2013, 100: 229–234
- 23 Mai Q, Zou H. The fused Kolmogorov filter: A nonparametric model-free screening method. Ann Statist, 2015, 43: 1471–1497
- 24 He Q, Lin D. A variable selection method for genome-wide association studies. Bioinformatics, 2011, 27: 1–8
- 25 Ogutu J O, Piepho H P, Schulz-Streeck T. A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proc, 2011, 5: S11
- 26 Da Y, Wang C, Wang S, et al. Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. PLoS ONE, 2014, 9: e87666
- 27 Speed D, Balding D J. MultiBLUP: Improved SNP-based prediction for complex traits. Genome Res, 2014, 24: 1550–1557
- 28 Yang C, Wan X, Yang Q, et al. Identifying main effects and epistatic interactions from large-scale SNP data via adaptive group Lasso. BMC Bioinf, 2010, 11: S18
- 29 Yi N, Kaklamani V G, Pasche B. Bayesian analysis of genetic interactions in case-control studies, with application to adiponectin genes and colorectal cancer risk. Ann Human Genets, 2011, 75: 90–104
- 30 Hao N, Zhang H H. Interaction screening for ultrahigh-dimensional data. J Amer Statist Assoc, 2014, 109: 1285–1301
- 31 Li J, Zhong W, Li R, et al. A fast algorithm for detecting gene-gene interactions in genome-wide association studies. Ann Appl Stat, 2014, 8: 2292–2318
- 32 Kong Y, Li D, Fan Y, et al. Interaction pursuit in high-dimensional multi-response regression via distance correlation. Ann Statist, 2017, 45: 897–922
- 33 Luo Z, Wang M, Long Y, et al. Incorporating pleiotropic quantitative trait loci in dissection of complex traits: Seed yield in rapeseed as an example. Theor Appl Genet, 2017, 130: 1569–1585
- 34 Qiu D, Morgan C, Shi J, et al. A comparative linkage map of oilseed rape and its use for QTL analysis of seed oil and erucic acid content. Theor Appl Genet, 2006, 114: 67–80
- 35 Li H. Statistical Learning Methods (in Chinese). Beijing: Tsinghua University Press, 2012 [李航. 统计学习方法. 北京: 清华大学出版社, 2012]

Two-step Bayesian method for detecting gene-gene interactions with moderate sparsity

Qinglong Yang, Yuan Liu, Yanyan Liu & Jun Zou

Abstract Based on flowering time trait in *Brassica napus*, we propose a two-step Bayesian analysis method for model selection under moderate sparsity. The phenotype is quantitative and the genotype takes two values 1 and -1 . By considering the gene-gene interaction, the model dimensionality increases dramatically so that the existing variable selection methods cannot handle this problem. In this paper, we propose a two-step selection method. First, we apply the fused Kolmogorov filter to screen out some obviously unimportant variables. Then, based on the remaining variables, we add interaction effect to the model. To overcome the heavy computation burden of the Bayesian method, we add indicator variables to every candidate effects and choose the active effect via posterior distribution of the indicator variable. The simulation results show that our proposed method performs better delivering higher prediction accuracy than those without the indicator function. Furthermore, we apply the proposed model to analyze a flowering time data in the *Brassica napus* population in multiple environments.

Keywords interaction effect, high-dimensional data, genome-wide association, Bayesian method, *Brassica napus*' flowering time

MSC(2010) 62C10, 62J07

doi: 10.1360/SSM-2019-0243