



高维回归模型的 Fiducial 变量选择

赵勇超¹, 梁华², 李新民^{1*}

1. 青岛大学数学与统计学院, 青岛 266071;

2. Department of Statistics, The George Washington University, Washington, DC 20052, USA

E-mail: zhaoyc.kido@163.com, hliang@gwu.edu, xmli@qdu.edu

收稿日期: 2022-04-26; 接受日期: 2022-12-10; 网络出版日期: 2023-03-31; * 通信作者

国家自然科学基金 (批准号: 11871294) 资助项目

摘要 本文提出一种基于 Fiducial 推断的模型不确定度量, 给出基于 Fiducial 边际似然的高维线性回归变量选择方法改进 ε -容许集 (ε -admissible set, EAS) 方法. 在传统的稀疏性假设下, 本文证明该方法所选择的最优模型具有相合性. 此外, 本文还建议使用一种快速的算法来挑选阈值参数. 本文通过模拟研究验证了所提出方法的优越性, 并使用所提出的方法分析了一个真实数据案例.

关键词 变量选择 ε -容许集 Fiducial 边际似然函数 模型不确定度量 分组独立 Metropolis-Hastings 算法

MSC (2020) 主题分类 62J99, 62F30

1 引言

变量选择是统计学习中的一个重要步骤, 主要分为两大类: (1) 以优化理论为支撑的正则化变量选择; (2) 构建候选模型的概率描述. 第二类中最具代表性的是 Bayes 变量选择.

正则化变量选择因其计算速度快和对超高维度数据具有良好的适应性而得到迅速发展. 最具代表性的方法包括最小绝对值收敛和选择算子 (least absolute shrinkage and selection operator, LASSO)^[22]、平滑剪切绝对偏差 (smoothly clipped absolute deviation, SCAD) 惩罚^[6]、弹性网 (elastic net, EN)^[29]、自适应 LASSO^[28]、极小极大凹惩罚 (minimax concave penalty, MCP)^[26] 和 spike-and-slab LASSO^[19] 等. 众所周知, 正则化方法不能给出所选模型是最优模型的概率保证, 并且这些方法普遍缺乏一些重要的统计性质 (参见文献 [11]).

Bayes 变量选择方法包括基于边际似然的最高模型后验准则和偏差信息准则 (deviance information criterion, DIC)^[21]. 前者通过 Bayes 后验概率对候选模型进行概率描述, 并选择后验概率最大的模型作为最优模型. 该方法克服了正则化方法的不足. 与 Bayes 信息准则 (Bayesian information criterion, BIC)^[20] 相关的 Bayes 因子^[2-5, 15] 也可以通过后验概率来构造. DIC 包含模型的拟合优度和对模型复杂度的惩罚两部分. 该准则可以通过基于抽样的方法来估计, 并且在各种 Bayes 软件包中很容易获

英文引用格式: Zhao Y C, Liang H, Li X M. Fiducial variable selection for the high-dimensional regression model (in Chinese). Sci Sin Math, 2023, 53: 839–858, doi: 10.1360/SCM-2022-0245

得. Johnson 和 Rossell^[14] 通过模型参数的非局部先验密度将 Bayes 变量选择方法推广到高维情形. 然而, Bayes 变量选择方法仍然面临着诸多挑战. 一方面是模型和参数先验分布的选择问题; 另一方面, Maity 等^[18] 指出, DIC 不能对数据生成模型和过拟合模型进行渐近区分.

为了弥补这些不足, Hannig 等^[9] 提出了基于 Fiducial 推断的 Fiducial 变量选择范式, 以便在不给定主观先验信息的情形下获得参数空间的有意义的概率描述. 然而当 p 较大时, 由于此方法需要遍历整个候选模型集合, 导致计算量极大. 此外, 在高维情形下需要考虑的候选模型太多, 通常会导致包括真实模型在内的所有候选模型均被分配一个任意小的模型概率^[7], 以至于真实模型所反映的信号十分微弱, 并不能达到变量选择的目的. 为了减少候选模型的个数, 从而减轻计算负担, Williams 和 Hannig^[23] 引入了容许集的概念, 极大地压缩了候选模型集合. 这项工作带来了一种新的 Fiducial 变量选择方法, 被称为 ε -容许集 (ε -admissible set, EAS). 随后 EAS 方法得到了迅速的发展, Koner 和 Williams^[16] 不仅将 EAS 方法推广到了多响应变量的高维线性回归问题中, 还改善了 EAS 的实现算法, 并且描述了如何挑选一个调优参数 ε 的更新版本. 最近, Williams 等^[25] 进一步将 EAS 方法推广到了向量自回归模型.

EAS 需要对真实模型的模型长度 p_o 进行估计. 模拟研究表明 EAS 对 p_o 的选择是敏感的: 当 p_o 被低估时, EAS 的表现非常好; 但当 p_o 被准确估计或高估时, EAS 的错误识别率会增加, 从而导致过拟合现象. 这可能会引起一些担忧, 因为在高维问题的研究中, 真实模型的稀疏度往往是被高估的. 此外当样本量比较大时, EAS 的表现性能良好; 而当样本量比较小时, EAS 的效果还有待提高. 值得注意的是, 当 $p > n$ 时, 高维回归模型的回归系数并不存在唯一的最小二乘估计. 事实上, 现有文献的各种变量选择策略主要都是施加一定的稀疏性假设, 使得高维问题在数学上可处理. 尽管稀疏性假设实际上并不一定正确, 但就目前而言这是必要的. 为了方便在数学框架下讨论高维变量选择问题, 本文在一定的稀疏性假设下进行研究.

本文尝试使用充分统计量代替样本来构建参数的 Fiducial 分布, 并提出一种基于 Fiducial 推断的模型不确定性度量新方法, 该方法同时考虑了参数的 Fiducial 分布和似然函数. 进一步地, 将所提出的模型概率度量整合到 EAS 中, 提出了一个 Fiducial 变量选择准则. 该准则可以很好地处理高维甚至超高维的情形. 第 2 节提出一种新的模型不确定性度量方法, 以此构建 Fiducial 边际似然函数, 并简要介绍 EAS 方法, 最后给出一种新的 Fiducial 变量选择准则 (Fiducial model selection criteria, FMC). 第 3 节提出一种有效的算法来解决所提出的变量选择准则的实现问题, 并讨论阈值参数 p_o 的选择问题. 第 4 节相比于已有变量选择方法, 通过数值模拟试验来研究本文所提出的方法的数值表现. 第 5 节通过分析核黄素产量的真实数据案例进一步说明本文所提出的方法. 第 6 节对全文进行总结并讨论未来的研究方向. 相关假设和技术证明在附录 A 中.

2 FMC 变量选择方法

本节首先基于充分统计量的结构方程给出回归模型参数的 Fiducial 分布, 然后引入一种模型不确定性度量, 并将其与 EAS 相结合, 提出一个基于 Fiducial 边际似然的变量选择准则.

考虑 n 个独立的观测 (x_i, y_i) ($i = 1, 2, \dots, n$), 其中 $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$. 令 $y = (y_1, \dots, y_n)^T$ 以及 $X = (x_1, \dots, x_n)^T$, 其中 x_i 为 p 维列向量. $S = \{M \mid M \subset \{1, 2, \dots, p\}\}$ 是所有模型所构成的集合, 则集合 S 中包含 2^p 个候选模型. 对于候选模型 M , 记 $|M|$ 是模型 M 的模型长度. 本文考虑回归系数为 β_M 、误差项为 $\varepsilon_M \sim N_n(0, \sigma_M^2 I_n)$ 的多元正态线性模型

$$Y = X_M \beta_M + \varepsilon_M, \quad (2.1)$$

其中设计矩阵 X_M 定义为仅由 M 对应的 X 的那些列组成的矩阵, β_M 是由 M 对应的 β 的那些元素所构成的 $|M|$ 维向量. 令 $\theta_M = (\beta_M, \sigma_M) \in \Theta_M$, 其中 Θ_M 是第 M 个模型的参数空间. 令 $M_o \in S$ 为真实模型, 变量选择的目标是在 2^p 个候选模型中识别出真实模型 M_o .

2.1 基于充分统计量的结构方程

Fiducial 推断的目的是在没有先验信息的情形下构造参数空间的有意义的概率描述. 考虑结构方程

$$Y = G(M, U, \theta_M), \quad M \in S, \quad \theta_M \in \Theta_M, \quad (2.2)$$

其中 U 的分布是完全已知的. 如果 $Y = y$ 被观测, 并且关于 θ_M 的逆 G^{-1} 存在, 则 θ_M 的分布可以通过反解结构方程 (2.2) 获得, 即 $\theta_M = G^{-1}(M, y, u)$. 如果 U 被重复抽样, 通过结构方程的逆可以得到一组关于 θ_M 的随机样本, 该随机样本称为 θ_M 的 Fiducial 样本. 记 $r(\theta_M | y)$ 表示 θ_M 对应的 Fiducial 密度. 若对似然函数 $f(y, \theta)$ 施加一定的光滑性假设, 则 $r(\theta_M | y)$ 可以表示为 (参见文献 [9])

$$r(\theta_M | y) = \frac{f(y, \theta_M) J_M(y, \theta_M)}{\int_{\Theta_M} f(y, \theta_M) J(y, \theta_M) d\theta_M}, \quad (2.3)$$

其中, $J_M(y, \theta_M) = D(\frac{d}{d\theta_M} G(M, u, \theta_M) |_{u=H(M, y, \theta_M)})$, $D(A) = (\det A^T A)^{1/2}$, y 和 u 分别满足 $y = G(M, u, \theta_M)$ 和 $u = H(M, y, \theta_M)$. 基于 (2.3) 的推断无论是在理论上还是在模拟上都取得了很好的结果 (参见文献 [9]). 给定 (2.3), Hannig^[8] 定义了模型 M 的概率度量

$$r(M | y) = \frac{\int_{\Theta_M} r(\theta_M | y) d\theta_M}{\sum_{M' \in S} \int_{\Theta_{M'}} r(\theta_{M'} | y) d\theta_{M'}} = \frac{\int_{\Theta_M} f(y, \theta_M) J(y, \theta_M) d\theta_M}{\sum_{M' \in S} \int_{\Theta_{M'}} f(y, \theta_{M'}) J(y, \theta_{M'}) d\theta_{M'}}, \quad (2.4)$$

将其命名为模型 M 的 Fiducial 密度. 针对模型 (2.1), 上述结果可以简化为

$$r(\theta_M | y) \propto |X_M^T X_M|^{\frac{1}{2}} \text{RSS}_M^{\frac{1}{2}} \sigma_M^{-(n+1)} \exp \left\{ -\frac{\text{RSS}_M}{2\sigma_M^2} \right\}, \quad (2.5)$$

从而模型 M 的 Fiducial 密度为 $r(M | y) \propto \pi^{\frac{|M|}{2}} \Gamma(\frac{n-|M|}{2}) \text{RSS}_M^{\frac{n-|M|-1}{2}}$, 其中 $\text{RSS}_M = \|Y - X_M \hat{\beta}_M\|^2$. 直接使用 $\hat{M} = \arg \max_{M \in S} r(M | y)$ 作为变量选择准则会导致过拟合现象, 即会倾向于选择一个更大的模型. 一个自然的想法是通过引入适当的惩罚来解决这个问题, 如最小描述长度原则^[10] 或者 EAS 方法^[23] (后面会详细介绍).

值得注意的是, 在 (2.4) 中使用的表达式 $r(\theta_M | y)$ 是通过样本推导出来的, 并且涉及偏导数 $\frac{d}{d\theta_M} G(M, u, \theta_M)$ 的计算. 为了避免这种计算, 本文从充分统计量的结构方程入手推导 θ_M 的 Fiducial 分布.

对于模型 (2.1), (β_M, σ_M^2) 的充分统计量为

$$(\hat{\beta}_M, \hat{\sigma}_M^2) = \left((X_M^T X_M)^{-1} X_M^T Y, \frac{\text{RSS}_M}{n - |M|} \right). \quad (2.6)$$

注意到 $\hat{\beta}_M \sim N_{|M|}(\beta_M, \sigma_M^2 (X_M^T X_M)^{-1})$, $(n - |M|) \hat{\sigma}_M^2 \sim \sigma_M^2 \chi_{n-|M|}^2$. 令 $Z \sim N_{|M|}(0, I_{|M|})$, $U \sim \chi_{n-|M|}^2$, 并且两者相互独立, 则有

$$\hat{\beta}_M = \beta_M + \sigma_M (X_M^T X_M)^{-\frac{1}{2}} Z, \quad (n - |M|) \hat{\sigma}_M^2 = \sigma_M^2 U.$$

当给定观测值 y 和 (z, u) 时, 该结构方程的逆存在唯一解

$$\beta_M = \hat{\beta}_M - \sigma_M (X_M^T X_M)^{-\frac{1}{2}} z \quad \text{和} \quad \sigma_M^2 = \frac{(n - |M|) \hat{\sigma}_M^2}{u}. \quad (2.7)$$

于是参数 β_M 和 σ_M 的密度函数为

$$r_{\text{new}}(\beta_M | \sigma_M) = (2\pi\sigma_M^2)^{-\frac{|M|}{2}} |(X_M^T X_M)^{-\frac{1}{2}}| \exp \left\{ -\frac{\|X_M(\beta_M - \hat{\beta}_M)\|_2^2}{2\sigma_M^2} \right\},$$

$$r_{\text{new}}(\sigma_M) = \frac{\text{RSS}_M^{\frac{n-|M|}{2}}}{2^{\frac{n-|M|}{2}-1} \Gamma(\frac{n-|M|}{2})} \sigma_M^{-(n-|M|+1)} \exp \left\{ -\frac{\text{RSS}_M}{2\sigma_M^2} \right\}.$$

因此参数向量 $\theta_M = (\beta_M, \sigma_M^2)$ 的密度为

$$r_{\text{new}}(\theta_M | y) = \frac{\pi^{-\frac{|M|}{2}}}{2^{\frac{n}{2}-1} \Gamma(\frac{n-|M|}{2})} \text{RSS}_M^{\frac{n-|M|}{2}} |X_M^T X_M|^{\frac{1}{2}} \sigma_M^{-(n+1)} \exp \left\{ -\frac{\|y - X_M \beta_M\|_2^2}{2\sigma_M^2} \right\}. \quad (2.8)$$

上述密度称为 θ_M 的 Fiducial 密度, 与 (2.5) 略有不同. 它可以用于推断或构造变量选择准则, 并发挥着类似于 Hannig 等^[9] 提出的 (2.5) 的作用.

2.2 一种度量模型不确定性的新方法

在 Bayes 推断中, Bayes 边际似然被用来构建 Bayes 变量选择准则. 对于模型 (2.1), 其 Bayes 边际概率为

$$P(y | M) \triangleq \int_{\Theta_M} f(y | \theta, M) \cdot \pi(\theta_M) d\theta_M,$$

其中 $f(y | \theta, M)$ 和 $\pi(\theta_M)$ 是模型 (2.1) 的似然函数和先验分布. 则模型 (2.1) 的后验概率由下式给出:

$$P(M | y) = \frac{P(y | M) \pi(M)}{\sum_{M' \in S} P(y | M') \pi(M')} \propto P(y | M) \pi(M),$$

其中 $\pi(M)$ 是模型 (2.1) 在模型空间 S 中的先验概率. 由此可知, Bayes 变量选择的表现性能在很大程度上依赖于先验概率 $\pi(M)$. 本文将 Bayes 边际似然中的 $\pi(M)$ 替换为 $r_{\text{new}}(\theta_M | y)$, 后者称为 Fiducial 边际似然函数, 定义如下:

定义 2.1 令 $f(y | \theta, M)$ 为第 M 个模型下观测样本的似然函数, $r_{\text{new}}(\theta_M | y)$ 为 θ_M 的 Fiducial 密度, 则

$$P_r(y | M) \triangleq \int_{\Theta_M} f(y | \theta, M) \cdot r_{\text{new}}(\theta_M | y) d\theta_M \quad (2.9)$$

称为 Fiducial 边际似然 (Fiducial marginal likelihood, FML) 函数.

假设所有模型的先验分布 $\pi(M)$ 为均匀分布, 则有

$$P_r(M | y) = \frac{P_r(y | M) \pi(M)}{\sum_{M' \in S} P_r(y | M') \pi(M')} \propto P_r(y | M). \quad (2.10)$$

Fiducial 边际似然函数和 Bayes 边际似然的不同在于, 用 $r_{\text{new}}(\theta_M | y)$ 替换了 $P(M | y)$ 中的 $\pi(M)$, 即参数的先验分布被参数的 Fiducial 分布所代替. 由于 $r_{\text{new}}(\theta_M | y)$ 允许试验者以客观的角度

获取参数的信息, 并且这使得研究其大样本性质成为可能. 进一步, 因为 $P_r(M | y)$ 反映了观测样本更可能来自哪个模型, 所以 $P_r(M | y)$ 可以被用于推断或者构建变量选择准则. 类似的思想被 Williams 等^[24] 用于构造广义 Fiducial 因子, 它可以作为 Bayes 因子的有效替代. 但二者关注的问题不同: 本文旨在解决试验者基于观测数据寻找最优拟合模型的问题, 而文献 [24] 则是解决基于观测数据进行预测的问题, 所以二者有所区别. 进一步地, 与直接使用 Fiducial 分布 $r(M | y)$ 的变量选择准则相比, 基于 Fiducial 边际似然函数的准则综合了 Fiducial 分布 (2.8) 和似然函数的优点, 从而更具优势. 基于这种模型不确定性度量 (2.10), 本文提出一种新的 Fiducial 变量选择准则.

2.3 EAS 方法的介绍

为了压缩候选模型集, 应使任何一个候选模型都是非冗余的, 以便候选的真实模型包含解释或预测观测数据所需的最少信息. 这个动机启发了 Williams 和 Hannig^[23] 提出如下的 ε -容许集的定义.

定义 2.2 给定 $\varepsilon > 0$, 当一个带有指标集 $M \subset \{1, 2, \dots, p\}$ 的系数向量 β_M 满足 $h(\beta_M) = 1$ 时, 该系数向量称为 ε -容许的, 其中

$$h(\beta_M) := I\left\{\frac{1}{2}\|X^T(X_M\beta_M - Xb_{\min})\|_2^2 \geq \varepsilon\right\},$$

$b_{\min}$ 是条件 $\|b\|_0 \leq |M| - 1$ 下优化问题 $\min_{b \in \mathbb{R}^p} \frac{1}{2}\|X^T(X_M\beta_M - Xb)\|_2^2$ 的解.

值得注意的是, 候选模型集的压缩依赖于 ε , 选取合适的 ε 在执行 EAS 的过程中起着关键作用. 本文建议使用以下规则来确定 ε :

$$\varepsilon = \Lambda_M \hat{\sigma}_M^2 \left(\frac{n^{0.51}}{9} + |M| \frac{\log(p\pi)^{1.1}}{9} + 1 - p_0 \right)_+, \quad (2.11)$$

其中, $\Lambda_M := \|H_M X\|_{\text{Frobenius}}^2$, $H_M := X_M(X_M^T X_M)^{-1} X_M^T$, $\hat{\sigma}_M^2 := \frac{\text{RSS}_M}{n - |M|}$. 易见 EAS 的表现受阈值参数 p_0 的影响. 阈值参数 p_0 反映了试验者对真实模型 M_o 的稀疏程度的认知, 它可以直接被指定或通过数据驱动的方法来获得. 值得注意的是, 由 (2.11) 知, p_0 与 ε 成反比, 这从侧面反映出较小的 ε 会导致较大的 p_0 , 较大的 p_0 对概率描述准则的识别能力要求更高. 本文将在后面详细讨论如何给出阈值参数 p_0 的合理估计.

Williams 和 Hannig^[23] 将 θ_M 的 Fiducial 分布限制在容许的集合上, 推导出相应的 Fiducial 概率密度

$$r^\varepsilon(M | y) \propto \pi^{\frac{|M|}{2}} \Gamma\left(\frac{n - |M|}{2}\right) \text{RSS}_M^{-\frac{n - |M| - 1}{2}} E(h(\beta_M)), \quad M \in S, \quad (2.12)$$

其中期望是相对于 β_M 的 Fiducial 分布而言的. 具有最高 Fiducial 概率 $r^\varepsilon(M | y)$ 的模型作为最优模型, 此方法称为 EAS.

2.4 基于 Fiducial 边际似然的变量选择准则

受到 EAS 的启发, 本文也将参数 θ_M 的 Fiducial 密度 (2.8) 限制到容许的集合上, 即

$$r_{\text{new}}^\varepsilon(\theta_M | y) \triangleq r_{\text{new}}(\theta_M | y) \cdot h(\beta_M). \quad (2.13)$$

限制后的结果可以推导出对应的 $P_r^\varepsilon(M | y)$, 具体结果由命题 2.1 中给出.

命题 2.1 对于模型 M , 参数 β_M 的 Fiducial 分布满足

$$t_{n-|M|} \left(\hat{\beta}_M, \frac{\text{RSS}_M}{n-|M|} (X_M^T X_M)^{-1} \right), \quad (2.14)$$

所以基于 Fiducial 密度 (2.13) 的 Fiducial 边际似然函数具有下述的具体形式:

$$P_r^\varepsilon(y | M) \propto E\{h(\beta_M)[\|y - X_M \beta_M\|_2^2]^{-\frac{n}{2}}\},$$

其中期望是相对于参数 β_M 的 Fiducial 分布而言的. 进一步地, 假设所有模型的先验分布 $\pi(M)$ 都是均匀分布, 则有

$$P_r^\varepsilon(M | y) \propto E\{h(\beta_M)[\|y - X_M \beta_M\|_2^2]^{-\frac{n}{2}}\}. \quad (2.15)$$

综上所述, 基于 Fiducial 边际似然的变量选择准则最终被给出, 即

$$\hat{M} = \arg \max_{M \in S} P_r^\varepsilon(M | y). \quad (2.16)$$

该变量选择方法简称为 FMC. 通过 ε -容许集方法极大地压缩了候选模型集, 使得本文提出的变量选择方法 (2.16) 能够应用于高维甚至超高维设置. 观察 (2.13) 和 (2.15), 可以得知模型的概率描述是建立在所有候选模型集上的, 然而对于那些不被容许的集合, 被分配的模型概率可以忽略不计. 相应地, 真实模型的信号所反映的信号变得更强. 进一步地, 定理 2.1 表明在一定条件下, 即便是 $p \gg n$, 这些可忽略的概率也会随着样本量的增大而逐渐趋于 0, 即当 $n \rightarrow \infty$ 时, $P_r^\varepsilon(M_o | y)$ 渐近收敛于 1.

对于模型 (2.1), X_M 是一个 $n \times |M|$ 的矩阵, 其各列的长度为 1. M_o 指的是真实模型所具有的积极变量集, 且 $\beta_M^0 = (X_M^T X_M)^{-1} X_M^T X_{M_o} \beta_{M_o}^0$. 对于 $j \in \{1, \dots, p\}$, 定义下面的子集类 $\mathcal{M}_j := \{M \neq M_o : |M| = j\}$.

定理 2.1 假设附录 A 中的条件 A.1–A.4 均成立. 则真实模型 M_o 满足, 当 $n \rightarrow \infty$ 时,

$$\frac{P_r^\varepsilon(M_o | y)}{\sum_{j=1}^{n^\alpha} \sum_{M: |M|=j} P_r^\varepsilon(M | y)} \xrightarrow{P_y} 1.$$

定理 2.1 表明, 当 n 足够大时, 真实模型的 Fiducial 密度 $P_r^\varepsilon(M | y)$ 与满足 $|M| \leq n^\alpha$ 的所有其他协变量子集 M 的 Fiducial 密度总和之比在概率上收敛于 1. 在条件 A.1–A.4 下, Fiducial 密度 $P_r^\varepsilon(M | y)$ 集中于真实模型 M_o , 即本文提出的准则所选择的最优模型具有相合性.

Ghosh 和 Ghattas^[7] 认为协变量子集的联合概括统计量对共线性更稳健. 与基于正则化的变量选择方法相比, FMC 提供了协变量子集上 Fiducial 分布的估计以及足够的理论保证. 在实践中, 候选模型的这种保证是非常重要的. 另外, 本文所提出的准则包括两个组成部分: 拟合优度和模型复杂度惩罚, 并以乘法的形式存在. 相对于 (2.12), (2.15) 的表达式更简单, 且更容易计算.

3 FMC 的实现

3.1 算法

实现变量选择准则 FMC 的关键在于如何有效地从 Fiducial 密度 $P_r^\varepsilon(M | y)$ 中抽取 M . 由于无法获得概率质量函数的封闭形式, 本文采用基于分组独立 Metropolis-Hastings 算法^[16] 从伪边缘分布中

生成样本. 该算法在文献 [23] 中成功地解决了 (2.12) 的实现问题. 简单地讲, 首先通过 (2.14) 引入潜变量从而得到一个解析表达式. 然后抽取一个模型 M , 并从 M 对应的多元 t 分布中抽取 N_{MC} 样本来计算期望的 Monte Carlo 估计, 记为 $\mathcal{B}$. 经过上述过程可以得到一个联合的 Markov 链 $(M, \mathcal{B})$, 通过仅关注 M 部分, 便可从 M 的伪边际分布中获取样本. 通过计算模型在稳定后的 Markov 链中的后验概率, 选择后验概率最大的模型 M 作为最优模型即可实现变量选择准则 FMC. 算法 1 中给出了具体的实现步骤.

初始模型 M^0 通常用最小二乘估计来选择, 正如文献 [23] 中所说, 这样的选择可以减少计算 $h(\beta_M)$ 时的迭代次数. 算法 1 中的步骤 4 涉及的 $h(\beta_M^j)$ 将使用显示 L_0 最小化算法来计算.

算法 1 FMC 算法

输入:

y 、 X 、 N_{step} 、 N_{burnin} 和 N_{MC} , 一个初始的模型 M^0 , 一个提议概率函数 $q(\tilde{M} | M^0)$.

$S_{MCMC} = \{\emptyset\}$ 是一条 Markov 链以及默认的 ε .

输出:

最优模型 M_{best} .

对于 $i = 1, \dots, N_{\text{step}}$, 执行

步骤 1 自 M^0 开始通过随机游走的方式搜索下一个模型 $M \in S$, 记为 $\tilde{M}$,

$$\tilde{M} = \begin{cases} M^0 \cup \{\text{一个新的协变量}\}, & \omega = \frac{1}{3}, \\ M^0 \setminus \{\text{一个已存在的协变量}\}, & \omega = \frac{1}{3}, \\ M^0 \setminus \{\text{一个已存在的协变量}\} \cup \{\text{一个新的协变量}\}, & \omega = \frac{1}{3}. \end{cases}$$

步骤 2 通过 (2.6) 对 $(\beta_{M^0}, \sigma_{M^0})$ 和 $(\beta_{\tilde{M}}, \sigma_{\tilde{M}})$ 进行估计, 分别记为 $(\hat{\beta}_{M^0}, \hat{\sigma}_{M^0})$ 和 $(\hat{\beta}_{\tilde{M}}, \hat{\sigma}_{\tilde{M}})$.

步骤 3 分别从分布 (2.14) 中抽取样本 $\beta_{M^0}^i$ 和 $\beta_{\tilde{M}}^i$ ($i = 1, 2, \dots, N_{MC}$), $t_{n-|M|}(\hat{\beta}_M, \hat{\sigma}_M^2 (X_M^T X_M)^{-1})$, 其中 $M \in \{M^0, \tilde{M}\}$.

步骤 4 对于所有的 $j = 1, 2, \dots, N_{MC}$, 计算 $\mathcal{H}_M^j \triangleq h(\beta_M^j) [\|y - X_M \beta_M^j\|_2^2]^{-\frac{n}{2}}$; $M \in \{M^0, \tilde{M}\}$, 其中 $h(\beta_M^j)$ 可由显式 L_0 最小化算法计算.

步骤 5 计算 (2.15) 的 Monte Carlo 估计, $\widehat{P}_r^\varepsilon(M | y) \propto \frac{1}{N_{MC}} \sum_{j=1}^{N_{MC}} \mathcal{H}_M^j$; $M \in \{M^0, \tilde{M}\}$.

步骤 6 如果 $\min\{\frac{\widehat{P}_r^\varepsilon(\tilde{M}|y)q(M^0|\tilde{M})}{\widehat{P}_r^\varepsilon(M^0|y)q(\tilde{M}|M^0)}, 1\} > u^*$, $u^* \sim U(0, 1)$, 那么 $M^0 = \tilde{M}$ 并且 $S_{MCMC} = \{S_{MCMC}, \tilde{M}\}$.

否则 $M^0 = M^0$ 并且 $S_{MCMC} = \{S_{MCMC}, \emptyset\}$.

终止循环

步骤 7 舍弃 Markov 链 S_{MCMC} 上前 N_{burnin} 个样本. 计算所有模型在剩余 $N_{\text{step}} - N_{\text{burnin}}$ 个样本上的后验概率. 选择具有最高后验概率的模型作为最优模型 M_{best} .

输出 最优模型 M_{best} .

3.2 p_o 的选择问题

(2.11) 中的阈值参数 p_o 在平衡计算负担和选择精度方面起着关键作用. 这个量反映的是真实模型 M_o 的稀疏程度, 它代表了试验者对数据生成机制的理解. Williams 和 Hannig^[23] 建议采用 BIC 法或者 10 折交叉验证法来选择 p_o , 并认为后者的表现优于前者. 然而, 基于交叉验证的 EAS 方法计算成本较高. 考虑到在变量选择的初始阶段, 应当尽可能地节省计算资源, 为此, 本文借鉴了由 Zhu 等^[27]

提出的自适应最佳子集选择 (adaptive best subset selection, ABESS) 算法来获得真实模型 M_o 的稀疏度估计. ABESS 算法采用序列拼接的思想, 在模型的稀疏程度固定但未知的情形下, 能够在有限步内达到稳定解. 该算法已被证明在一定的条件下能够给出一致的估计, 并且可以通过 R 包 `abess` 来实现.

4 数值模拟

本节通过数值模拟试验来评估本文所提出方法的数值性能, 并将其与已得到充分研究的方法 (如 EAS、SCAD、MCP 和 EN) 在估计精度、模型复杂性和处理共线性的能力方面进行比较.

首先介绍本文所用到的评价指标. 真实模型中包含的变量称为积极效应, 否则称为消极效应. 设 M_o 和 M_o^c 为真实模型 M_o 中包含的积极效应和消极效应所构成的集合, M_i 为第 i 次试验中选择的最优模型, $\mathcal{N}$ 为重复试验次数. 通过稀疏度水平误差 (sparsity level error, SLE)、均方根误差 (root mean square error, RMSE)、积极效应识别率 (active effect identification error, AEIR) 和消极效应识别率 (inactive effect identification error, IEIR) 来评估性能, 各个指标的定义如下:

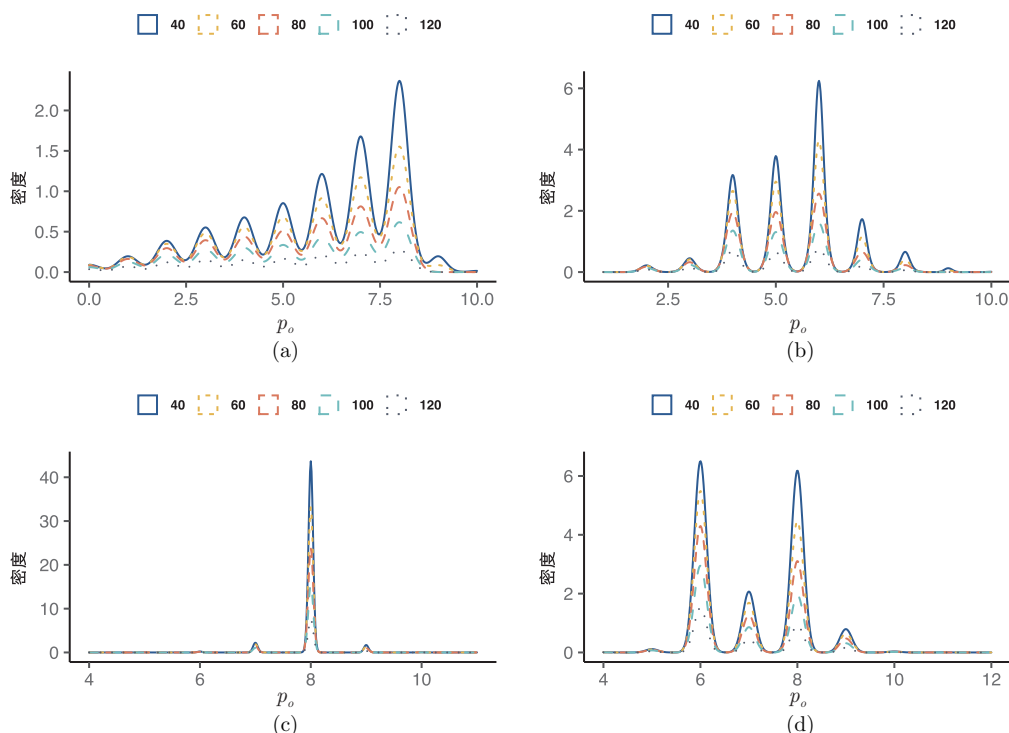
$$\begin{aligned} \text{SLE} &= \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \{|M_i| - |M_o|\}, \quad \text{RMSE} = \sqrt{\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_i - \hat{y}_i)^2}, \\ \text{AEIR} &= \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \frac{|M_i \cap M_o|}{|M_o|} \times 100\%, \quad \text{IEIR} = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \frac{|M_i \cap M_o^c|}{|M_o^c|} \times 100\%, \end{aligned}$$

其中 RMSE 是在大小为 $n_{\text{test}} = 30$ 的测试样本上获得的. 当 SLE 接近 0 且 RMSE 很小时, 所选模型更接近真实模型 (可能比真实模型更简单), 但仍能提供足够准确的预测. 当 SLE 和 IEIR 都接近 0 且 AEIR 接近 1 时, 所选模型接近真实模型, 说明所选模型具有较好的可解释性.

对于每个数据集抽取 $N_{\text{step}} = 10,000$ 个 MCMC (Markov chain Monte Carlo) 样本来实现本文所提出的方法, 并舍弃前 $N_{\text{burnin}} = 5,000$ 个抽样结果. 采用弹性网 (使用 `scikit-learn`) 对系数的平方进行估计, 其结果加上 n^{-2} 作为 MCMC 中协变量的建议权重. 算法中每一步用于估计 (2.15) 的重要性抽样数设为 $N_{\text{MC}} = 100$. 以上两个设置与文献 [23] 中的设置相同. 在给定或通过数据驱动方法获取 p_o 之后, ε 由 (2.11) 给出. 将上述设置应用于本文所提出的方法和 EAS. 另外 MCP、SCAD 和 EN-LASSO 由现有的 R 包 `ncvreg` 和 `glmnet` 来实现, 其中所需的设置均采用 R 包中的默认设置.

例 4.1 [23] 从模型 (2.1) 中生成数据, 其中设计矩阵 $X_{n \times p}$ 服从多元正态分布 $N_p(\mathbf{0}_p, \Sigma)$. 真实模型为 $\beta_{M_o} = (-1.5, -1, -0.8, -0.6, 0.6, 0.8, 1, 1.5, \mathbf{0}_{p-8})$, 其中 $p \in \{40, 60, 80, 100, 120\}$. $\Sigma (= (\sigma_{ij})_{p \times p})$ 被设定为对角矩阵或者 AR (auto-regressive) 相关矩阵, 即对于 $i, j \in \{1, \dots, p\}$, $\sigma_{ij} = I\{i = j\}$ 或 $\sigma_{ij} = 0.5^{|i-j|}$. 样本量 n 设定为 50 或者 100. 误差项的方差 σ^2 固定为 1. 在这个例子中, 通过比较 SLE、RMSE、AEIR 和 IEIR 4 个方面来评价 FMC、EAS、SCAD、MCP 和 EN 的数值表现.

借助例 4.1, 首先来验证 ABESS 算法的有效性和快速性. 在生成 20,000 组数据集后, 对每个数据集应用 ABESS 算法来获得 p_o 的一个估计值. 对于不同的参数设定, 绘制出 p_o 估计值的密度曲线, 结果展示在图 1, 其中, (a) $\sigma_{ij} = I(i = j)$, $n = 50$; (b) $\sigma_{ij} = 0.5^{|i-j|}$, $n = 50$; (c) $\sigma_{ij} = I(i = j)$, $n = 100$; (d) $\sigma_{ij} = 0.5^{|i-j|}$, $n = 100$. 可以看出, 当协变量之间相互独立时, ABESS 算法能以最大概率获得真实模型稀疏度的一致估计; 当协变量之间呈现 AR 相关时, ABESS 也可以提供一个合理的估计值. 另外, 对于每个数据集, ABESS 的平均计算时间仅为 1.19×10^{-2} 秒. 而 10 折交叉验证的平均运行时间远远超过了它. 由此可见, 与交叉验证相比, ABESS 的效率是非常有吸引力的. 因此在接下来的模拟研究

图 1 (网络版彩图) ABESS 对 p_o 的估计密度曲线图

中, 均先采用 ABESS 算法得到 p_o 的合理估计.

例 4.1 的数值模拟结果见图 2. 图 2 的第一行为稀疏度估计误差, 第二行为均方根误差, 第三行为积极效应识别率, 第四行为消极效应识别率. 左侧两列代表 $n = 50$; 右侧两列代表 $n = 100$; 列 (a) 和 (c) 代表协变量之间相互独立; 列 (b) 和 (d) 表示协变量之间 AR 相关两种情形. 根据数值模拟结果可知, 基于正则化的方法都存在不同程度的过拟合, 其中表现最好的是 MCP. 然而正则化变量选择方法和参数估计均依赖于系数真实大小的量级, 因此当样本量较小时, 它们的性能会受到一定的影响. 由图 2 可知, 在样本量较小时, 正则化方法的 RMSE 会随着维数的增加而增大, 远高于 FMC 和 EAS 的 RMSE; 随着样本量的增加, 正则化方法的 RMSE 变化趋于平稳, 与 FMC 和 EAS 的 RMSE 相当. 而 FMC 和 EAS 的表现差别不大. 此外, FMC 和 EAS 非常有效地控制了模型的长度, 在较小的模型下也能获得足够准确的预测结果. 特别是 FMC, 当其 SLE 接近于 0 时, 其 RMSE 在大多数情形下都是最小. 此外, 从被选模型的识别效果或可解释性方面来看, FMC 和 EAS 的表现优于竞争对手. 具体而言, 当两种方法的 AEIR 基本相同时, FMC 和 EAS 都能对消极效应进行更精确的压缩, 且在各种情形下, FMC 的 IEIR 值几乎最小. 综合而言, 本文提出的 FMC 在预测和识别方面都更胜一筹.

例 4.2 本例主要关注小样本下 FMC 和 EAS 方法的差异性. 采用与文献 [23] 中模拟设定 2 相同的数据生成方式, 其中 $n = 30$. 数据 Y 服从正态分布, 即 $Y \sim N_n(1 \cdot x^{(1)} + 1 \cdot x^{(2)} + \cdots + 1 \cdot x^{(9)}, I_n)$, 其中 $x^{(1)}, x^{(2)}, x^{(3)} \sim_{\text{i.i.d.}} N_n(0, I_n)$, $x^{(4)} \sim N_n(0.25 \cdot x^{(1)}, 0.1^2 I_n)$, $x^{(5)} \sim N_n(0.50 \cdot x^{(2)}, 0.1^2 I_n)$, $x^{(6)} \sim N_n(-0.75 \cdot x^{(3)}, 0.1^2 I_n)$, $x^{(7)} \sim N_n(1 \cdot x^{(1)} + 1 \cdot x^{(3)}, 0.1^2 I_n)$, $x^{(8)} \sim N_n(1 \cdot x^{(2)} - 1 \cdot x^{(3)}, 0.1^2 I_n)$, $x^{(9)} \sim N_n(1 \cdot x^{(1)} + 1 \cdot x^{(2)} + 1 \cdot x^{(3)}, 0.1^2 I_n)$. 由于 $x^{(4)}, \dots, x^{(9)}$ 均可以表示成 $x^{(1)}, x^{(2)}$ 和 $x^{(3)}$ 的线性组合, 因此例 4.2 的数据生成模型是缺乏唯一性的, 仅需要少数的协变量就可以到达足够精确的预测效果. 另外, 为了探索 FMC 和 EAS 两种方法对于阈值参数 p_o 的敏感程度, 设定 $p_o \in \{2, 3, \dots, 9\}$, 并

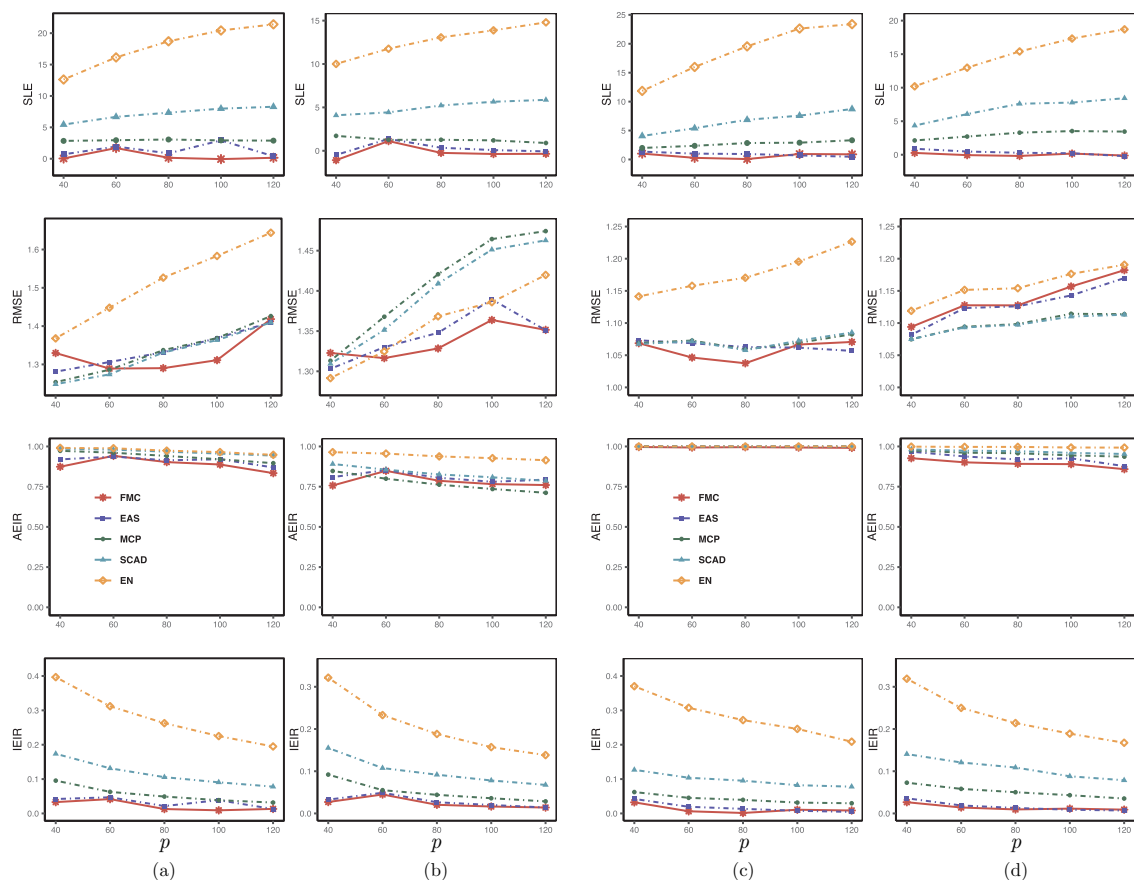


图 2 (网络版彩图) 例 4.1 的数值模拟结果

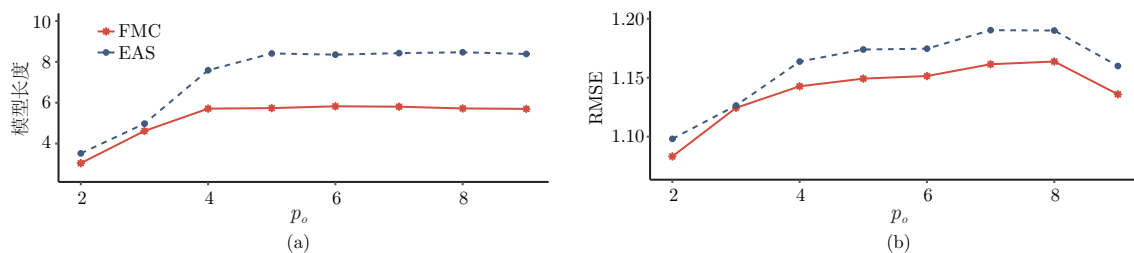


图 3 (网络版彩图) 例 4.2 的数值模拟结果

且对于每一种 p_o 的取值, 均随机生成 $N = 1,000$ 组数据集. 通过模型长度和 RMSE 两项指标来比较 FMC 和 EAS 的数值表现.

数值模拟结果见图 3, 其中 (a) 是模型长度相对于 p_o 的变化, (b) 是 RMSE 相对于 p_o 的变化. 可以看出, 当真实模型 M_o 的稀疏度被低估时, 即 p_o 较小时, FMC 和 EAS 都可以有效地压缩共线性的变量; 而当真实模型 M_o 的稀疏度被较好地估计或者高估 (即 $p_o \geq 4$) 时 FMC 的表现要比 EAS 好得多. 在各个情形下, FMC 都可以稳定地压缩共线性的变量, 并取得可观的预测效果. 这表明 FMC 对 p_o 的选择具有更强的鲁棒性. 此外, FMC 可以用一个更简单的模型获得足够准确的预测效果, 这在高维回归问题中通常是被鼓励的.

5 真实数据案例

本节分析影响核黄素 (维生素 B₂) 生产率的潜在重要基因. 核黄素产量的真实数据集来自于文献 [1], 数据集包含 71 个观测值和 4,088 个基因表达, 其中响应变量为枯草芽孢杆菌的核黄素生产率的对数. Bühlmann 等^[1] 以及 Javanmard 和 Montanari^[13] 均采用线性回归模型来探索响应变量与协变量之间的关系, 并确定了 YXLD_at 为影响核黄素 (维生素 B₂) 产量的重要基因表达, 因此本文在模型中直接保留 YXLD_at 基因, 分别用 FMC、EAS、MCP、SCAD 和 EN 来识别影响核黄素产量的重要基因表达.

首先用 ABESS 算法得到 p_o 的估计值 $\hat{p}_o = 6$, 然后用 BESS (best subset selection) 算法进行初筛并保留 $p = 71$ 个基因表达进行分析. 需要注意的是, BESS 和 ABESS 算法都源自文献 [27], 区别在于前者需要指定稀疏度 (在这里稀疏度指定为 $p = 71$), 而后者是依靠数据驱动的自适应版本. 在表 1 中,

表 1 5 种方法所选的重要基因及其显著性水平

指标	基因名称	FMC	EAS	MCP	SCAD	EN
1	YXLD_at	✓***	✓×	✓***	✓***	✓×
2	CARA_at	✓***	✓***	✓***	✓***	✓***
3	XHLA_at	✓***	✓***	✓***	✓***	✓***
4	YHDS_r_at	✓*	✓*	✓**	✓**	✓*
5	YOAB_at	✓***	✓***	✓***	✓***	✓***
6	YHDZ_at	✓***		✓***	✓***	✓×
7	LACA_at	✓°		✓×	✓×	✓×
8	YFHB_at		✓***			
9	YXLG_at		✓×			✓×
10	SPOVAA_at			✓***	✓***	✓***
11	USD_i_at			✓**	✓*	
12	CYSE_at			✓×	✓°	✓×
13	YEBC_at			✓***	✓**	✓*
14	YTSA_at			✓*	✓×	✓*
15	YSHB_at				✓×	✓×
16	YDDJ_r_at				✓×	✓×
17	MUTL_at					✓×
18	SPOIID_at					✓°
19	YACN_at					✓×
20	YCLB_at					✓×
21	YDGL_at					✓×
22	YJZB_at					✓×
23	YQHL_at					✓×
模型长度		7	7	12	14	21
调整后的 R^2		0.9981	0.9977	0.9989	0.9989	0.9989
RMSE		0.2950	0.3310	0.2190	0.2290	0.2130

***、**、* 和 ° 分别代表显著性水平小于 10^{-5} 、0.001、0.01 和 0.05, × 表示系数显著为 0.

列出了各个方法所选择的重要基因, 以及所选模型对应的模型长度. 表 1 中还标注了各系数的显著性检验结果、调整后的 R^2 和 RMSE.

根据表 1 可知, EN 选择的模型最大, 其次是 SCAD 和 MCP, FMC 和 EAS 最小, 且 FMC 和 EAS 选择的模型长度远小于其他. 虽然 FMC 和 EAS 选择的模型长度相同, 与 EAS 选择模型中有 YXLD_at、CARA_at、XHLA_at、YHDS_r_at 和 YOAB_at 等 5 个基因重叠, 但 EAS 选择的 YFHB_at 是其他方法没有识别的, 另外 EAS 还选择了只有 EN 选择的基因 YXLG_at. 此外, FMC 选择的模型的 RMSE 略小于基于 EAS 选择的模型的 RMSE. 用这些方法得到的模型调整后的 R^2 都高于 0.99, 即所有的被选模型都几乎完美地解释了响应变量与解释变量之间的变化趋势.

接下来讨论 FMC 所选模型的优越性. 根据表 1 可以得出, 其他方法选择的模型至少包含一个在统计学意义上不显著的基因表达, 如 EAS 方法选择的 YXLG_at, MCP 方法选择的 LACA_at 和 CYSE_at, 以及 SCAD 和 EN 方法选择的 LACA_at、YTSA_at、YSHB_at 和 YDDJ_r_at. 进一步地, 为了研究某些基因表达之间是否存在共线性问题, 图 4 绘制了这些“显著”变量并集的相关系数图. 通过观察图 4 可以立即识别出协变量 17 至 23 之间的相关性大于其余变量. 这一发现表明, EN 并没有很好地区分协变量之间的共线性问题. 对于 1 至 16 之间的非显著协变量, 采用线性回归模型将这些非显著协变量与其他“显著非零”协变量进行回归. 作为说明, 表 2 给出了 4 种回归的结果. 由表 2 可以看出, 4 个线性模型的 F 检验和 t 检验均显著, 调整后的 R^2 也接近于 1. 这些结果均表明, 协变量 8 至 16 之间或协变量 8 至 16 与协变量 1 至 7 之间存在较强的共线性问题. 有趣的是, FMC 所选的模型巧妙地避免了这些共线性问题, 并且所选模型仍然具有可观的预测精度. 由于选择了更高质量的协变量集, 因此所选模型的可解释性也非常令人印象深刻.

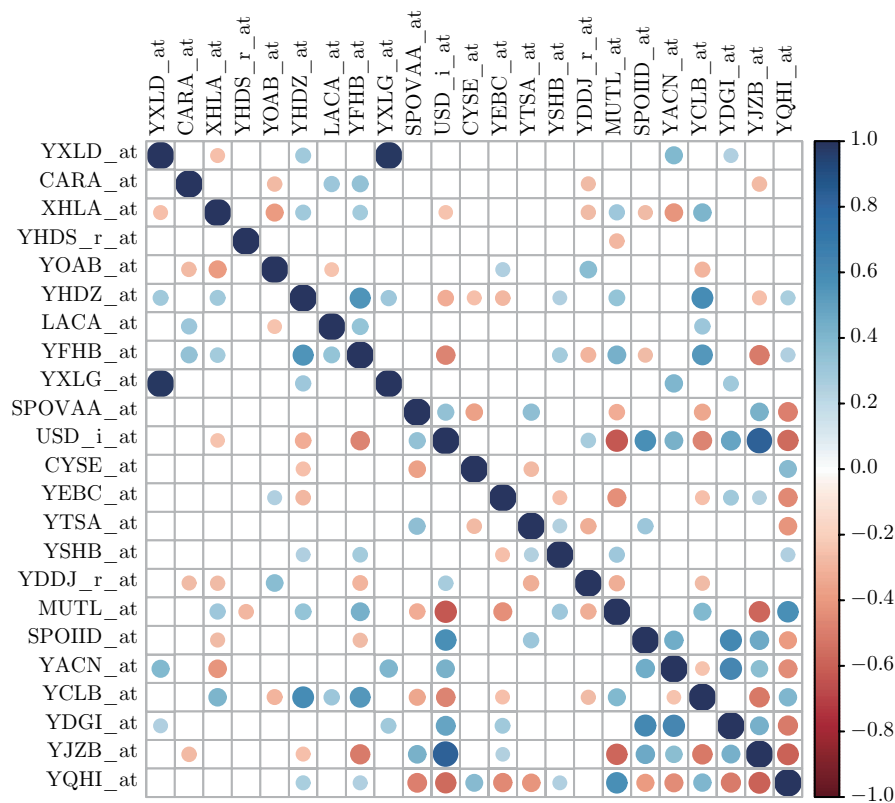


图 4 (网络版彩图) 显著变量之间的相关系数图

表 2 4 种线性回归模型的拟合结果汇总. 图例的说明与表 1 保持一致

(M1): $YFHB_at \sim XHLA_at + CARA_at + YOAB_at + YHDZ_at + LACA_at - 1$				
调整后的 R^2 : 0.9997 F- 统计量: $5.332E+04$, 自由度为 (5, 66), p - 值: $< 2.2E-16$				
基因名称	估计值	标准误	t- 值	$Pr(> t)$
XHLA_at	0.08175	0.02194	3.726	0.000405 ***
CARA_at	0.08418	0.02104	4.001	0.000162 ***
YOAB_at	0.32868	0.04661	7.052	1.31E-09 ***
YHDZ_at	0.47233	0.06609	7.147	8.87E-10 ***
LACA_at	0.06921	0.02469	2.804	0.006629 **
(M2): $YXLG_at \sim YXLD_at - 1$				
调整后的 R^2 : 0.9998 F- 统计量: $2.944E+05$, 自由度为 (1, 70), p - 值: $< 2.2E-16$				
基因名称	估计值	标准误	t- 值	$Pr(> t)$
YXLD_at	0.944858	0.001741	542.6	$< 2E-16$ ***
(M3): $YTSA_at \sim SPOVAA_at + YDDJ_r_at - 1$				
调整后的 R^2 : 0.9988 F- 统计量: $2.955E+04$, 自由度为 (2, 69), p - 值: $< 2.2E-16$				
基因名称	估计值	标准误	t- 值	$Pr(> t)$
SPOVAA_at	0.9795	0.0949	10.322	1.24E-15 ***
YDDJ_r_at	0.3294	0.0811	4.062	0.0001***
(M4): $YDDJ_at \sim USD_i_at + CYSE_at - 1$				
调整后的 R^2 : 0.9983 F- 统计量: $2.054E+04$, 自由度为 (2, 69), p - 值: $< 2.2E-16$				
基因名称	估计值	标准误	t- 值	$Pr(> t)$
USD_i_at	0.21959	0.04842	4.535	2.36E-05 ***
CYSE_at	0.60363	0.03675	16.426	$< 2E-16$ ***

6 讨论

本文提出了一种基于 FML 的变量选择方法. 该方法有效地避免 EAS 方法潜在的计算复杂的问题, 并建议使用一种更有效的算法来选择用于生成 ε -容许集的阈值参数. 本文所提出的方法 FMC 在继承了 EAS 方法良好的大样本性质的同时, 还具有计算方便的特点和良好的有限样本性能, 并且进一步增强了高维数据分析所需的冗余变量压缩能力.

借鉴部分线性模型和部分线性可加模型的主要工作, 本文工作可通过以下两个步骤扩展到高维半参数模型的范畴: (1) 采用样条逼近非参数或半参数函数, 从而得到一个近似线性模型; (2) 将所提出的方法应用于近似线性模型, 建立参数的 Fiducial 密度, 得到相应的 FML, 然后推导出每个候选模型的 FMC. 虽然这些扩展在框架和实现上都很简单, 但是理论上的论证可能会带来很大的困难. 对这些问题进行细致研究很有意义, 但超出了本文的范围.

另外, 正如文献 [27] 在总结部分所讨论的, 当模型不满足稀疏性假设, 即存在大量有较小系数的候选模型时, ABESS 的效果还有待提高. 稀疏性假设便于能够在数学的框架下研究高维问题, 本文是在稀疏性假设下来研究高维回归下的 Fiducial 变量选择问题. 对于非稀疏情形下的 p_0 的估计问题, 还有待进一步研究.

致谢 作者感谢审稿人所提出的宝贵建议.

参考文献

- 1 Bühlmann P, Kalisch M, Meier L. High-dimensional statistics with a view toward applications in biology. *Annu Rev Stat Appl*, 2014, 1: 255–278
- 2 Chan J C, Jeliazkov I. Efficient simulation and integrated likelihood estimation in state space models. *Int J Math Model Numer Optim*, 2009, 1: 101–120
- 3 Chib S. Marginal likelihood from the Gibbs output. *J Amer Statist Assoc*, 1995, 90: 1313–1321
- 4 Chib S, Jeliazkov I. Marginal likelihood from the Metropolis-Hastings output. *J Amer Statist Assoc*, 2001, 96: 270–281
- 5 Chib S, Jeliazkov I. Accept-reject Metropolis-Hastings sampling and marginal likelihood estimation. *Stat Neerl*, 2005, 59: 30–44
- 6 Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Amer Statist Assoc*, 2001, 96: 1348–1360
- 7 Ghosh J, Ghattas A E. Bayesian variable selection under collinearity. *Amer Statist*, 2015, 69: 165–173
- 8 Hannig J. On generalized fiducial inference. *Statist Sinica*, 2009, 19: 491–544
- 9 Hannig J, Iyer H, Lai R C S, et al. Generalized fiducial inference: A review and new results. *J Amer Statist Assoc*, 2016, 111: 1346–1361
- 10 Hannig J, Lee T C M. Generalized fiducial inference for wavelet regression. *Biometrika*, 2009, 96: 847–860
- 11 Hazimeh H, Mazumder R. Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *Oper Res*, 2020, 68: 1517–1537
- 12 Jameson G J O. Inequalities for Gamma function ratios. *Amer Math Monthly*, 2013, 120: 936–940
- 13 Javanmard A, Montanari A. Confidence intervals and hypothesis testing for high-dimensional regression. *J Mach Learn Res*, 2014, 15: 2869–2909
- 14 Johnson V E, Rossell D. Bayesian model selection in high-dimensional settings. *J Amer Statist Assoc*, 2012, 107: 649–660
- 15 Kass R E, Raftery A E. Bayes factors. *J Amer Statist Assoc*, 1995, 90: 773–795
- 16 Koner S, Williams J P. The EAS approach to variable selection for multivariate response data in high-dimensional settings. *arXiv:2107.04873*, 2021
- 17 Lai R C S, Hannig J, Lee T C M. Generalized fiducial inference for ultrahigh-dimensional regression. *J Amer Statist Assoc*, 2015, 110: 760–772
- 18 Maity A K, Basu S, Ghosh S. Bayesian criterion-based variable selection. *J R Stat Soc Ser C Appl Stat*, 2021, 70: 835–857
- 19 Ročková V, George E I. The spike-and-slab LASSO. *J Amer Statist Assoc*, 2018, 113: 431–444
- 20 Schwarz G. Estimating the dimension of a model. *Ann Statist*, 1978, 6: 461–464
- 21 Spiegelhalter D J, Best N G, Carlin B P, et al. Bayesian measures of model complexity and fit. *J R Stat Soc Ser B Stat Methodol*, 2002, 64: 583–639
- 22 Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Ser B Stat Methodol*, 1996, 58: 267–288
- 23 Williams J P, Hannig J. Nonpenalized variable selection in high-dimensional linear model settings via generalized fiducial inference. *Ann Statist*, 2019, 47: 1723–1753
- 24 Williams J P, Ommen D M, Hannig J. Generalized fiducial factor: An alternative to the Bayes factor for forensic identification of source problems. *Ann Appl Stat*, 2023, 17: 378–402
- 25 Williams J P, Xie Y Y, Hannig J. The EAS approach for graphical selection consistency in vector autoregression models. *Canad J Statist*, 2023, 51: 674–703
- 26 Zhang C H. Nearly unbiased variable selection under minimax concave penalty. *Ann Statist*, 2010, 38: 894–942
- 27 Zhu J, Wen C, Zhu J, et al. A polynomial algorithm for best-subset selection problem. *Proc Natl Acad Sci USA*, 2020, 117: 33117–33123
- 28 Zou H. The adaptive Lasso and its oracle properties. *J Amer Statist Assoc*, 2006, 101: 1418–1429
- 29 Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B Stat Methodol*, 2005, 67: 301–320

附录 A

本附录陈述了主要结果所需的必要条件、初步引理以及主要结果的证明.

附录 A.1 正则性条件

本小节列出了定理 2.1 所需的正则性条件:

条件 A.1 对于较大的 n 和 p , 真实模型 M_o 满足 $|M_o| \leq \log(n)^\gamma$,

$$\lim_{\substack{n \rightarrow \infty \\ p \rightarrow \infty}} \min \left\{ \frac{\Delta_M}{|M_o| \log(p)} : M \neq M_o, |M| \leq |M_o| \right\} = \infty, \quad \liminf_{\substack{n \rightarrow \infty \\ p \rightarrow \infty}} \frac{n^{1-\alpha}}{\log(p)} > 2, \quad \log(p) < \frac{n}{4 \log(n)^\gamma},$$

其中 $\Delta_M := \|X_{M_o} \beta_{M_o}^0 - H_M X_{M_o} \beta_{M_o}^0\|_2^2$ [17].

值得注意的是, 对于条件 A.1 中的参数 γ , 它表示即使真实模型的增长速度快于 $\log(n)$, 但只要不快于 $\log(n)$ 的幂次, 渐近结果仍然成立. 而常数 $\alpha \in (0, 1)$ 是 2^p 个子集的样本空间满足最终渐近性质所需要的唯一明确限制. 在稀疏性假设下, 忽略一些候选模型是合理的. 对于较小的 α , 备选模型收敛到真实模型 M_o 的速度更快, 因为更少的模型需要被考虑; 然而过于小的 α 可能会导致真实模型 M_o 也被排除在外.

条件 A.2 下列等式成立:

$$\lim_{\substack{n \rightarrow \infty \\ p \rightarrow \infty}} \min_{|M| \leq n^\alpha} \frac{\frac{\varepsilon}{18 \Lambda_M \hat{\sigma}_M^2} - \varphi(M, n, p)}{\log(n)} = \infty,$$

其中 $\varphi(M, n, p) := 4e^2 n^\alpha + (1 + 4e^2) |M| \log(p) + \log\left(\frac{|M|}{1 - \frac{1}{n - |M|}}\right)$. 另外, 对于所有满足 $|M| \leq n^\alpha$ 的模型 M , 都成立 $\frac{\varepsilon}{9 \Lambda_M \hat{\sigma}_M^2} < \frac{n - |M|}{2}$.

条件 A.3 与文献 [23, 条件 3.1] 一致.

条件 A.4 与文献 [23, 条件 3.3] 一致.

简单而言, 条件 A.1 和 A.3 确保了真实模型 M_o 是可识别的. 条件 A.2 和 A.4 表示为实现定理 2.1 的一致结果, ε 所需要满足的条件. 上述条件与 Williams 和 Hannig [23] 建立其主要结论时所施加的正则性条件相似但有所区别: 对于条件 A.1, Williams 和 Hannig [23] 对 $\log(p)$ 施加的边界与本文不同; 条件 A.2 关于 ε 的收敛要求也比文献 [23] 中的更精简.

附录 A.2 初步引理

为证明定理 2.1, 首先需要一些初步的引理. 引理 A.2 在证明定理 2.1 时起到了关键作用. 不失一般性, 固定真实模型的误差项的未知标准差为 $\sigma_{M_o} = 1$.

引理 A.1 情形 1 当 $|M| \leq |M_o|$ 时, 有

$$P_y \left(\bigcap_{j=1}^{|M_o|} \bigcap_{\mathcal{M}_j} \left\{ Y : \left(\frac{\text{RSS}_{M_o}}{\text{RSS}_M} \right)^{\frac{\alpha}{2}} \leq e^{-2 \log(n)^\gamma \log(p)} \right\} \right) \geq 1 - V_1,$$

其中 $\mathcal{M}_j := \{M \neq M_o : |M| = j\}$, P_y 表示与 Y 的抽样分布相关的概率测度, 并且

$$V_1 := \max_{\substack{M \neq M_o \\ |M| \leq |M_o|}} \left\{ \frac{12 |M_o| \exp\left[-\frac{\Delta_M}{72} + |M_o| \log(p)\right]}{\sqrt{2\pi} \Delta_M} + |M_o| \exp \left[-\frac{\Delta_M}{12} + \frac{|M_o|}{2} + |M_o| \log(p) \right] \right. \\ \left. + |M_o| \exp \left[-\frac{\xi_n}{48} \frac{n \Delta_M}{\log(n)^\gamma \log(p)} + \frac{n - |M_o|}{2} + |M_o| \log(p) \right] \right\},$$

其中 $\xi_n = 1 - \frac{2 \log(n)^\gamma \log(p)}{n/2} \rightarrow 1$.

情形 2 当 $|M_o| + 1 \leq |M| \leq n^\alpha$ 时, 有

$$P_y \left(\bigcap_{j=|M_o|+1}^{n^\alpha} \bigcap_{M:|M|=j} \left\{ Y : \left(\frac{\text{RSS}_{M_o}}{\text{RSS}_M} \right)^{\frac{n}{2}} \leq \exp[4e^2(n^\alpha + j \log(p))] \right\} \right) \geq 1 - V_2,$$

其中

$$V_2 := n^\alpha \exp \left\{ -2 \frac{n - n^\alpha - |M_o|}{n} n^\alpha + \left(1 - 2 \frac{n - n^\alpha - |M_o|}{n} \right) (|M_o| + 1) \log(p) + \frac{n^\alpha + |M_o|}{2} \right\} \\ + \frac{n^\alpha \exp \left\{ -\frac{n - n^\alpha - |M_o|}{2} + n^\alpha \log(p) \right\}}{\sqrt{\pi(n - n^\alpha - |M_o|)}}.$$

引理 A.1 与 Williams 和 Hannig^[23] 提出的引理 A.1 相似, 不同之处在于本文关注的是 $(\frac{\text{RSS}_{M_o}}{\text{RSS}_M})^{n/2}$, 而他们关注的是 $(\frac{\text{RSS}_{M_o}}{\text{RSS}_M})^{(n-j-1)/2}$, 这也导致了收敛概率 V_1 和 V_2 的差异. 考虑到证明方法与文献 [23, 引理 A.1] 的证明过程类似, 因此这里仅陈述证明的不同之处, 详细证明可参见文献 [23].

引理 A.1 的证明 情形 1 $M \in \mathcal{M}_j$, 其中 $j \leq |M_o|$. 不同于文献 [23], 取 $\xi_n = 1 - \frac{2 \log(n)^\gamma \log(p)}{n/2}$, 当 n 和 p 较大时, 由条件 A.1 可知, ξ_n 介于 0 和 1 之间. 易得文献 [23, 引理 A.1] 的证明过程将 $\xi_{n,j}$ 替换成 ξ_n 依旧成立, 因此有

$$P_y \left(\bigcup_{j=1}^{|M_o|} \bigcup_{M_j} \left\{ \left(\frac{\text{RSS}_{M_o}}{\text{RSS}_M} \right)^{\frac{n}{2}} > \xi_n^{\frac{n}{2}} \right\} \right) \\ \leq \sum_{j=1}^{|M_o|} \max_{M_j} \left\{ \frac{12 \exp[-\frac{\Delta_M}{72} + j \log(p)]}{\sqrt{2\pi\Delta_M}} + \exp \left[-\frac{\Delta_M}{12} + \frac{|M_o|}{2} + j \log(p) \right] \right. \\ \left. + \exp \left[-\frac{\xi_n}{48} \frac{n\Delta_M}{\log(n)^\gamma \log(p)} + \frac{n - |M_o|}{2} + j \log(p) \right] \right\}.$$

根据条件 A.1, 当 $\xi_n \rightarrow 1$ 时, 该界限趋近于 0. 因此对于所有满足 $|M| \leq |M_o|$ 的模型 M , 均有

$$\left(\frac{\text{RSS}_{M_o}}{\text{RSS}_M} \right)^{\frac{n}{2}} \leq \left(1 - \frac{2 \log(n)^\gamma \log(p)}{n/2} \right)^{\frac{n}{2}} \leq \exp\{-2 \log(n)^\gamma \log(p)\}.$$

情形 2 同样地, 用 $a_n \triangleq \xi_n$ 代替 $\xi_{n,j}$, 则文献 [23, 引理 A.1] 的情形 2 的证明过程依旧成立, 因此有

$$P_y \left(\frac{\text{RSS}_{M_o}}{\text{RSS}_M} > a_n \right) \leq \exp \left\{ -\left(\frac{n-j-l}{e^2} \right) \frac{a_n - 1}{4} + \frac{j+l}{2} \right\} + \frac{\exp(-\frac{n-j-l}{2})}{\sqrt{\pi(n-j-l)}}.$$

不同于文献 [23], 取

$$a_n = 1 + \frac{4e^2(n^\alpha + j \log(p))}{n/2}.$$

则对于满足 $|M_o| < |M| \leq n^\alpha$ 的模型 M , 有

$$P_y \left(\bigcup_{j=|M_o|+1}^{n^\alpha} \bigcup_{M:|M|=j} \left\{ \left(\frac{\text{RSS}_{M_o}}{\text{RSS}_M} \right)^{\frac{n}{2}} > a_n^{\frac{n}{2}} \right\} \right) \\ \leq \sum_{j=|M_o|+1}^{n^\alpha} \exp \left\{ -2 \frac{n-j-l}{n} n^\alpha + \left(1 - 2 \frac{n-j-l}{n} \right) j \log(p) + \frac{j+l}{2} \right\}$$

$$+ \frac{\exp\{-\frac{n-j-l}{2} + j \log(p)\}}{\sqrt{\pi(n-j-l)}} \leq V_2.$$

注意到当条件 A.1 满足时, 上述界限趋近于 0. 因此, 对于所有满足 $|M_o| < |M| \leq n^\alpha$ 的 M 都有

$$\left(\frac{\text{RSS}_{M_o}}{\text{RSS}_M}\right)^{\frac{n}{2}} \leq \left(1 + \frac{4e^2(n^\alpha + j \log(p))}{n/2}\right)^{\frac{n}{2}} \leq \exp\{4e^2(n^\alpha + j \log(p))\}.$$

引理得证. $\square$

引理 A.2 假定模型 M_o 和 β_{M_o} 的 Fiducial 分布满足 (2.14), 则当 $n \rightarrow \infty$ 时, 有

$$\mathbb{E}\left(h(\beta_{M_o})\left[1 + \frac{\|X_{M_o}(\beta_{M_o} - \hat{\beta}_{M_o})\|_2^2}{\text{RSS}_{M_o}}\right]^{-\frac{n}{2}}\right) \rightarrow \mathbb{E}(h(\beta_{M_o})),$$

其中期望部分是相对于参数 β_M 的 Fiducial 分布而言的.

证明 若无特别说明, 本证明所涉及的所有依概率收敛都是基于 β_M 的 Fiducial 分布. 首先证明

$$\frac{\|X_{M_o}(\beta_{M_o} - \hat{\beta}_{M_o})\|_2^2}{\text{RSS}_{M_o}} \cdot \frac{n}{2} \xrightarrow{p} 0. \quad (\text{A.1})$$

根据 β_{M_o} 的 Fiducial 分布可知,

$$T \triangleq \sqrt{\frac{n - |M_o|}{\text{RSS}_{M_o}}}(X_{M_o}^T X_{M_o})^{\frac{1}{2}}(\beta_{M_o} - \hat{\beta}_{M_o}) \sim t_{n-|M_o|}(0, I_{|M_o|}).$$

记 $A = X_{M_o}(X_{M_o}^T X_{M_o})^{-\frac{1}{2}}$, 它的奇异值分解结果为 $A = QDW^T$, 其中 Q 和 W 是正交矩阵. 于是 $\|X_{M_o}(\beta_{M_o} - \hat{\beta}_{M_o})\|_2^2 / \text{RSS}_{M_o} = \|AT\|_2^2 / (n - |M_o|) = \|QDW^T T\|_2^2 / (n - |M_o|)$, 因此有

$$T = \sqrt{\frac{n - |M_o|}{V}} \cdot Z \Rightarrow W^T T = \sqrt{\frac{n - |M_o|}{V}} W^T Z \triangleq \sqrt{\frac{n - |M_o|}{V}} \tilde{Z} \triangleq \tilde{T},$$

其中 $V \sim \chi_{n-|M_o|}^2$ 以及 $Z, \tilde{Z} \sim N_{|M_o|}(0, I_{|M_o|})$. 所以有

$$\|AT\|_2^2 = \tilde{T}^T D^T D \tilde{T} = \sum_{i=1}^{|M_o|} \tilde{T}_i^2 \lambda_i \leq \max_{i=1}^{|M_o|} \tilde{T}_j^2 \sum_{i=1}^{|M_o|} \lambda_i = |M_o| \cdot \max \tilde{T}_j^2,$$

其中 $\sum_{i=1}^{|M_o|} \lambda_i = \text{tr}(A^T A) = |M_o|$ 并且 λ_i 是 $A^T A$ 的第 i 个特征值. 由此可得

$$\begin{aligned} \mathbb{P}\left(\frac{\|X_{M_o}(\beta_{M_o} - \hat{\beta}_{M_o})\|_2^2}{\text{RSS}_{M_o}} \geq \frac{2\varepsilon}{n}\right) &= \mathbb{P}\left(\|AT\|_2^2 \geq 2\varepsilon \cdot \frac{n - |M_o|}{n}\right) \\ &\leq \mathbb{P}\left(\max \tilde{T}_j^2 \geq 2\varepsilon \cdot \frac{n - |M_o|}{n|M_o|}\right) \\ &= \mathbb{P}\left(\max |Z_j| \geq \sqrt{\frac{2\varepsilon v}{n|M_o|}}\right). \end{aligned}$$

又因为 $V \sim \chi_{n-|M_o|}^2$, 故

$$\mathbb{P}\left(\frac{\|X_{M_o}(\beta_{M_o} - \hat{\beta}_{M_o})\|_2^2}{\text{RSS}_{M_o}} \geq \frac{2\varepsilon}{n}\right) \leq \int_0^\infty \mathbb{P}\left(\max |Z_j| \geq \sqrt{\frac{2\varepsilon v}{n|M_o|}}\right) \cdot f_{\chi_{n-|M_o|}^2} dv$$

$$\begin{aligned}
&\leq \int_0^\infty \sum_{j=1}^{|M_o|} 2 \cdot P\left(Z_j \leq -\sqrt{\frac{2\varepsilon v}{n|M_o|}}\right) \cdot f_{\chi^2_{n-|M_o|}} dv \\
&\leq \int_0^\infty \frac{|M_o|^{3/2} \sqrt{n}}{\sqrt{\varepsilon\pi}} v^{-\frac{1}{2}} \cdot \exp\left(-\frac{\varepsilon v}{n|M_o|}\right) \cdot f_{\chi^2_{n-|M_o|}} dv \\
&\leq \frac{|M_o|^{3/2}}{\sqrt{\varepsilon\pi}} \left(\frac{n|M_o|}{n|M_o|+2\varepsilon}\right)^{\frac{n-|M_o|-1}{2}} \frac{\sqrt{n(n-|M_o|)}}{n-|M_o|-1},
\end{aligned}$$

其中第 3 个不等式成立是因为对于 $x > 0$, 有 $\Phi(-x) \leq \frac{1}{x\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ 成立, 其中 Φ 是标准正态分布; 最后一个不等式成立是因为 Jameson^[12] 建立的不等式:

$$\frac{\Gamma(n-|M_o|-1)}{\Gamma(n-|M_o|)} \leq \frac{\sqrt{2(n-|M_o|)}}{n-|M_o|-1}.$$

对于任意 $\varepsilon > 0$, 均有 $\frac{n|M_o|}{n|M_o|+2\varepsilon} < 1$ 成立. 综上所述, (A.1) 得证.

记 $\mathcal{F} \triangleq \frac{\|X_{M_o}(\beta_{M_o} - \hat{\beta}_{M_o})\|_2^2}{\text{RSS}_{M_o}}$, 则 $\mathcal{F} \xrightarrow{p} 0$, $\mathcal{F} \cdot \frac{n}{2} \xrightarrow{p} 0$, 从而有

$$h(\beta_{M_o}) \left[1 + \frac{\|X_{M_o}(\beta_{M_o} - \hat{\beta}_{M_o})\|_2^2}{\text{RSS}_{M_o}}\right]^{-n/2} \xrightarrow{p} h(\beta_{M_o}). \quad (\text{A.2})$$

又因为 (A.2) 左侧可被常数 1 控制住, 由 Lebesgue 控制收敛定理可知, 对 (A.2) 两侧同时取期望, 引理得证. $\square$

附录 A.3 命题 2.1 的证明

易得 $\|y - X_M \beta_M\|_2^2 = \|X_M(\beta_M - \hat{\beta}_M)\|_2^2 + \text{RSS}_M$. 由 (2.8) 可知, β_M 的 Fiducial 分布为

$$\begin{aligned}
r_{\text{new}}(\beta_M | y) &= \int_0^\infty \frac{\pi^{-\frac{|M|}{2}}}{2^{\frac{n}{2}-1} \Gamma(\frac{n-|M|}{2})} \text{RSS}_M^{\frac{n-|M|}{2}} |X_M^T X_M|^{\frac{1}{2}} \sigma_M^{-(n+1)} \exp\left\{-\frac{\|y - X_M \beta_M\|_2^2}{2\sigma_M^2}\right\} d\sigma_M \\
&= \frac{\Gamma(\frac{n}{2}) \text{RSS}_M^{\frac{n-|M|}{2}}}{\pi^{\frac{|M|}{2}} \Gamma(\frac{n-|M|}{2})} |X_M^T X_M|^{\frac{1}{2}} [\|y - X_M \beta_M\|_2^2]^{-\frac{n}{2}} \\
&= \frac{\Gamma(\frac{n}{2}) \text{RSS}_M^{\frac{n-|M|}{2}}}{\pi^{\frac{|M|}{2}} \Gamma(\frac{n-|M|}{2})} |X_M^T X_M|^{\frac{1}{2}} \\
&\quad \times \left[1 + \frac{1}{n-|M|} (\beta_M - \hat{\beta}_M)^T \left(\frac{n-|M|}{\text{RSS}_M} (X_M^T X_M)^{-1}\right) (\beta_M - \hat{\beta}_M)\right]^{-\frac{n}{2}}.
\end{aligned}$$

所以 β_M 服从多元 t 分布 $t_{n-|M|}(\hat{\beta}_M, \frac{\text{RSS}_M}{n-|M|} (X_M^T X_M)^{-1})$. 因此直接计算可得

$$\begin{aligned}
P_r^\varepsilon(y | M) &= \int_{R^{|M|}} \int_0^\infty f(y | \theta, M) \cdot r_{\text{new}}(\theta_M | y) h(\beta_M) d\sigma_M d\beta_M \\
&= \int_{R^{|M|}} \int_0^\infty h(\beta_M) \frac{\text{RSS}_M^{\frac{n-|M|}{2}}}{2^{n-1} \pi^{\frac{n+|M|}{2}} \Gamma(\frac{n-|M|}{2})} |X_M^T X_M|^{\frac{1}{2}} \sigma_M^{-2n-1} \exp\left\{-\frac{\|y - X_M \beta_M\|_2^2}{\sigma_M^2}\right\} d\sigma_M d\beta_M \\
&= \int_{R^{|M|}} h(\beta_M) \cdot \frac{\text{RSS}_M^{\frac{n-|M|}{2}} \Gamma(n)}{2^n \pi^{\frac{n+|M|}{2}} \Gamma(\frac{n-|M|}{2})} |X_M^T X_M|^{\frac{1}{2}} (\|y - X_M \beta_M\|_2^2)^{-n} d\beta_M
\end{aligned}$$

$$\begin{aligned}
&= \frac{\Gamma(n)}{\Gamma(\frac{n}{2})} 2^{-n} \pi^{-\frac{n}{2}} \int_{R^{|M|}} h(\beta_M) [\|y - X_M \beta_M\|_2^2]^{-\frac{n}{2}} r_{\text{new}}(\beta_M | y) d\beta_M \\
&\propto E\{h(\beta_M) [\|y - X_M \beta_M\|_2^2]^{-\frac{n}{2}}\},
\end{aligned}$$

其中期望部分是相对于 β_M 的 Fiducial 分布而言的.

附录 A.4 定理 2.1 的证明

当模型的先验信息取为均匀分布时, $P_r^\varepsilon(M | y) \propto P_r^\varepsilon(y | M)$. 所以只需要关注 $P_r^\varepsilon(y | M)$. 首先需要证明的是, 对于任意协变量子集 M , 若 $M \neq M_o$, 则当 n 趋近于无穷时, $\frac{P_r^\varepsilon(y | M)}{P_r^\varepsilon(y | M_o)}$ 趋近于 0.

对于任意 $M \in \mathcal{M}_j$, 有

$$\begin{aligned}
\frac{P_r^\varepsilon(y | M)}{P_r^\varepsilon(y | M_o)} &\propto \frac{E\{h(\beta_M) [\|y - X_M \beta_M\|_2^2]^{-\frac{n}{2}}\}}{E\{h(\beta_{M_o}) [\|y - X_{M_o} \beta_{M_o}\|_2^2]^{-\frac{n}{2}}\}} \\
&\leq \frac{\text{RSS}_M^{-\frac{n}{2}} E\{h(\beta_M)\}}{\text{RSS}_{M_o}^{-\frac{n}{2}} E\{h(\beta_{M_o}) [1 + \frac{\|X_{M_o}(\beta_{M_o} - \hat{\beta}_{M_o})\|_2^2}{\text{RSS}_{M_o}}]^{-\frac{n}{2}}\}} \\
&\rightarrow \left(\frac{\text{RSS}_{M_o}}{\text{RSS}_M}\right)^{\frac{n}{2}} \frac{E\{h(\beta_M)\}}{E\{h(\beta_{M_o})\}} \\
&\triangleq \mathcal{H},
\end{aligned} \tag{A.3}$$

其中的收敛是由引理 A.2 保证的.

文献 [23, 定理 3.8] 保证以超过 $1 - V_3$ 概率使得下式成立:

$$\mathcal{H} \leq \left(\frac{\text{RSS}_{M_o}}{\text{RSS}_M}\right)^{\frac{n}{2}} \frac{E\{h(\beta_M)\}}{1 - g_1(M_o, n, p)},$$

其中

$$g_1(M, n, p) := \frac{2^{3/2} 3j \hat{\sigma}_M \sqrt{\Lambda_M}}{\sqrt{\pi \varepsilon} (1 - \frac{1}{n-j})} \exp\left\{-\frac{\varepsilon}{18 \Lambda_M \hat{\sigma}_M^2}\right\}.$$

根据条件 A.2, 对于一个给定的 $A_1 \in (0, 1)$, 可以选择一个充分大的 n 使得 $g_1(M_o, n, p) < A_1$ 成立. 当 $j \leq |M_o|$ 时, 应用引理 A.1 给出的界控制残差平方和比的部分, 并且期望部分放大至 1. 从而有超过 $1 - V_1 - V_3$ 的概率保证, $\mathcal{H}$ 可由下述的界控制住:

$$\mathcal{H} \leq \left(\frac{\text{RSS}_{M_o}}{\text{RSS}_M}\right)^{\frac{n}{2}} \frac{E(h(\beta_M))}{1 - A_1} \leq \left(\frac{\text{RSS}_{M_o}}{\text{RSS}_M}\right)^{\frac{n}{2}} \frac{1}{1 - A_1} \leq \frac{\exp\{-2 \log(n)^\gamma \log(p)\}}{1 - A_1}.$$

当 $|M_o| < j \leq n^\alpha$ 时, 应用文献 [23, 定理 3.7] 给出的界来放大期望部分, 并且应用引理 A.1 给出的界来控制残差平方和比的部分. 则有超过 $1 - V_2 - V_3 - V_4$ 的概率保证 $\mathcal{H}$ 可由下述的界控制住:

$$\mathcal{H} \leq \frac{\exp\{4e^2(n^\alpha + j \log(p))\}}{1 - A_1} g_1(M, n, p) = \frac{2^{3/2} 3j \hat{\sigma}_M \sqrt{\Lambda_M} \exp\{-\frac{\varepsilon}{18 \Lambda_M \hat{\sigma}_M^2} + 4e^2(n^\alpha + j \log(p))\}}{\sqrt{\pi \varepsilon} (1 - \frac{1}{n-j}) (1 - A_1)},$$

其中 $V_4 = \sum_{j=1}^{n^\alpha} \sum_{M \in \mathcal{M}_j} \frac{3|M| \sigma_M \sqrt{\Lambda_M}}{\sqrt{\varepsilon \pi}} \exp(-\frac{\varepsilon}{9 \sigma_M^2 \Lambda_M})$. 根据条件 A.2, 上面的结果可以由

$$\sum_{j=1}^{n^\alpha} \max_{\mathcal{M}_j} \frac{3|M| \sigma_M \sqrt{\Lambda_M}}{\sqrt{\varepsilon \pi} \exp\{\frac{\varepsilon}{9 \sigma_M^2 \Lambda_M} + j \log(p)\}} \rightarrow 0$$

进一步控制. 这里应用了文献 [23, 定理 3.7], 但 $c_1 n$ 被替换成了 n^α . 这样的替换并不影响结论的正确性, 因为对于任意 α 和 $c_1 \in (0, 1)$, n^α 的增长速度都要慢于 $c_1 n$.

接下来证明, 当 $n \rightarrow \infty$ 时, 有 $\sum_{j=1}^{n^\alpha} \sum_{M \in \mathcal{M}_j} \frac{P_r^\varepsilon(y|M)}{P_r^\varepsilon(y|M_o)} \rightarrow_p 0$. 将求和项分解为两部分, 即

$$\sum_{j=1}^{n^\alpha} \sum_{M \in \mathcal{M}_j} \frac{P_r^\varepsilon(y|M)}{P_r^\varepsilon(y|M_o)} \leq \sum_{j=1}^{|M_o|} \binom{p}{j} \max_{M \in \mathcal{M}_j} \frac{P_r^\varepsilon(y|M)}{P_r^\varepsilon(y|M_o)} + \sum_{j=|M_o|+1}^{n^\alpha} \binom{p}{j} \max_{|M|=j} \frac{P_r^\varepsilon(y|M)}{P_r^\varepsilon(y|M_o)} =: S_1 + S_2.$$

有超过 $1 - V_1 - V_3$ 的概率保证下列不等式成立:

$$S_1 \leq \frac{1}{1 - A_1} \sum_{j=1}^{|M_o|} \exp\{-2 \log(n)^\gamma \log(p) + j \log(p)\} \leq \frac{|M_o| \exp\{-(2 \log(n)^\gamma - |M_o|) \log(p)\}}{1 - A_1},$$

上式用到了二项式系数部分被放大. 类似地, 对于任意常数 $A_2 > 0$, 有超过 $1 - V_1 - V_2 - V_3 - V_4$ 的概率保证下列不等式成立:

$$\begin{aligned} S_2 &\leq A_2 \sum_{j=|M_o|+1}^{n^\alpha} \frac{\frac{j}{1-\frac{1}{n-j}} \max_{|M|=j} (\frac{\hat{\sigma}_M^2 \Lambda_M}{\varepsilon})^{\frac{1}{2}}}{\exp\{\min_{|M|=j} \{\frac{\varepsilon}{18 \Lambda_M \hat{\sigma}_M^2} - 4e^2(n^\alpha + j \log(p)) - j \log(p)\}\}} \\ &\leq \frac{A_2}{n^{1-\alpha}} \max_{|M_o| < |M| \leq n^\alpha} \left(\frac{\hat{\sigma}_M^2 \Lambda_M}{\varepsilon} \right)^{\frac{1}{2}}, \end{aligned}$$

最后一个不等式成立是因为条件 A.2.

综上所述, 有超过 $1 - V_1 - V_2 - V_3 - V_4$ 的概率保证, 下述不等式成立:

$$\sum_{j=1}^{n^\alpha} \sum_{M \in \mathcal{M}_j} \frac{P_r^\varepsilon(y|M)}{P_r^\varepsilon(y|M_o)} \leq \frac{|M_o| \exp\{-(2 \log(n)^\gamma - |M_o|) \log(p)\}}{1 - A_1} + \frac{\max_{|M_o| < |M| \leq n^\alpha} (\hat{\sigma}_M^2 \Lambda_M / \varepsilon)^{1/2}}{A_2^{-1} n^{1-\alpha}}.$$

在条件 A.1–A.4 成立的前提下, 当 $n \rightarrow \infty$ 时, 上式右侧部分趋近于 0. 由此可知

$$P_r^\varepsilon(M_o | y) \propto P_r^\varepsilon(y | M_o) = \frac{P_r^\varepsilon(y | M_o)}{\sum_{j=1}^{n^\alpha} \sum_{M: |M|=j} P_r^\varepsilon(y | M)} = \frac{1}{1 + \sum_{j=1}^{n^\alpha} \sum_{\mathcal{M}_j} \frac{P_r^\varepsilon(y | M)}{P_r^\varepsilon(y | M_o)}}.$$

定理得证.

Fiducial variable selection for the high-dimensional regression model

Yongchao Zhao, Hua Liang & Xinmin Li

Abstract In this study, we propose a new model uncertainty measure based on the fiducial inference, and a fiducial marginal likelihood-based variable selection procedure to improve the EAS (ε -admissible set) method. The optimal model selected is shown to be consistent under the conventional sparsity assumption. Furthermore, we suggest a fast algorithm to select threshold parameters. We conduct simulation experiments to demonstrate the superiority and the expedience of the proposed procedure and analyze a real data example as an illustration.

Keywords variable selection, ε -admissible set, fiducial marginal likelihood function, model uncertainty measure, grouped independence Metropolis-Hastings algorithm

MSC(2020) 62J99, 62F30

doi: 10.1360/SCM-2022-0245