Advances in Applied Mathematics 应用数学进展, 2021, 10(5), 1765-1775
Published Online May 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.105187

Hans汉斯

SCAD惩罚下基因－环境交互效应的识别方法研究

谢文玮[1]，李东喜[2]

[1]太原理工大学数学学院，山西 太原
[2]太原理工大学大数据学院，山西 太原
Email: wenweixie666@163.com, dxli0426@126.com

收稿日期：2021年4月25日；录用日期：2021年5月8日；发布日期：2021年5月28日

摘　要

对于许多复杂的癌症疾病，单一的基因效应或单一的环境效应不能进行有效的预测判断，识别与复杂疾病相关的基因－环境交互作用成为了高维数据下病理学和生物信息学研究的一大挑战。对于生存数据高维度、异质性、删失性等问题，我们提出了一种基于AFT模型的识别基因－环境交互作用的方法。该方法创新地通过采用LAD损失函数和SCAD惩罚函数相结合的目标函数减除数据不平衡带来的影响并选出服从主效应与交互效应间的强层次结构的交互项，并利用CCCP算法对目标函数进行优化求解。利用R进行了仿真研究和实证研究，从这两方面验证了该方法能稳健地选择出合适的基因效应和基因－环境交互效应，具有较好的预测性和稳定性，且该方法能有效压缩备选的变量，选出的模型简洁、有较好的解释性。

关键词

变量选择，基因－环境交互，SCAD惩罚项，加权LAD损失函数

Identification of Gene-Environment Interaction Using SCAD Penalty

Wenwei Xie[1], Dongxi Li[2]

[1]College of Mathematics, Taiyuan University of Technology, Taiyuan Shanxi
[2]College of Data Science, Taiyuan University of Technology, Taiyuan Shanxi
Email: wenweixie666@163.com, dxli0426@126.com

Received: Apr. 25th, 2021; accepted: May 8th, 2021; published: May 28th, 2021

文章引用: 谢文玮, 李东喜. SCAD 惩罚下基因－环境交互效应的识别方法研究[J]. 应用数学进展, 2021, 10(5): 1765-1775. DOI: 10.12677/aam.2021.105187

Abstract

For many complex cancer diseases, a single gene effect or a single environmental effect cannot account for the total variant of prediction results. Identifying the gene-environment interactions associated with complex diseases has become a major challenge for pathology and bioinformatics research under high-dimensional data. To solve the problems of high dimension, heterogeneity, and censored survival data, we proposed an AFT model-based method to identify gene-environment interactions. In this method, an objective function combining LAD loss function and SCAD penalty function is innovatively adopted to reduce the influence of unbalanced data and to select interaction terms that follow a strong hierarchical structure between main effects and interaction effects. The objective function is optimized and solved by CCCP algorithm. Simulation and empirical studies were carried out using R to verify that this method can select the appropriate gene effect and gene-environment interaction effect, and has good predictability and stability. Moreover, this method can effectively compress the alternative variables, and the selected model is simple and has good explanatory ability.

Keywords

Variable Selection, Gene-Environment Interaction, SCAD Penalty, Weighted LAD Loss Function

1. 引言

在病理学、生物信息学和生物医学的研究中，基因和环境的交互效应对许多复杂疾病的预测、治疗和药物开发有着重要影响。全基因组协会(Genome-Wide Association Studies, GWAS)的研究已经识别出了许多基因效应(如基因的表达、甲基化、单核苷酸多态性)，这些基因效应可以提供理解疾病和癌症的复杂机制的路径。也有研究表明环境和临床因素(如吸烟状况、空气污染、营养摄入等)同样与癌症有关。然而，对于许多复杂的疾病，单一的基因效应或单一的环境效应不能解释预测结果的全部变异。有许多癌症方面的例子可以体现基因－环境交互项的重要性。例如，携带 COX-2 基因单核苷酸多态性(基因效应)的人增加鲑鱼类鱼的摄入量(环境效应)会降低患前列腺癌的风险[1]。

基因－环境的交互作用被定义为单个或多个基因因素和环境因素的协同效应，这种协同效应不能通过上述基因因素或环境因素的边际效应来解释[2]。为了有效的识别出基因主效应及其与环境的交互效应，变量选择方法得到了广泛的发展。现有的交互项选择方法总体上分为两类。一种方法是，先采用边际分析选取所有的主效应，然后将对应的相关交互项加入到模型中。这些方法计算成本低，但效率也较低，因为它一次只分析少量的基因效应。另一类方法采用联合分析，通过层次约束、收缩方法、惩罚函数或不等式约束等方法，可以同时识别主效应和交互效应。这类方法不可避免地具有较高的计算成本[3]。

尽管许多统计方法已被应用其中，但在识别基因－环境交互作用的研究中仍存在一些挑战。首先，基因数据的高维度和大样本容量导致计算成本较高。此外，由于有不同的方式和工具来测量评估来自不同来源的效应，数据具有异质性的特点。例如，某些空气污染的粒子大小和化学组成成分在不同地区是不同的[4]。具有异质性的数据集在预测变量和响应变量中会出现异常值或者重尾性质等数据不平衡的特

性[5]。第三个挑战是数据污染和模型错误划分。通常模型会假设样本中没有异常值，但实际上数据中有异常值和杠杆点。而且大多数方法都会预设一个特定的参数或半参数模型。潜在的模型错误选择会导致估计偏差和分析结果不准确。这就要求模型具有稳健性，能够有效地处理数据污染和模型的误判。

Wu 等人使用 LASSO 惩罚建立了加速失效时间(Accelerate Failure Time, AFT)模型，该模型是一种稳健的服从层次结构的变量选择方法，可以处理脏数据[6]。Zhu 等人提出了一种分阶段回归方法和渐进式惩罚项选择基因－基因和基因－环境的交互作用。这种方法计算速度快，同时遵从了强层次结构[7]。Ma 等人采用加权最小二乘方法和一组稀疏惩罚来识别交互效应，这种方法降低了计算成本且有较好的预测性[8]。Ren 等人提出了一种基于 Spike-and-slab 先验的半参数贝叶斯模型，可以同时探索线性和非线性的相互作用[9]。Xu 等人利用截尾分位数偏相关方法(CQPCorr)建立分位数回归模型，对长尾数据和不平衡数据进行 G-E 交互效应分析[10]。

在本文中，我们建立了一种基于 AFT 模型并利用加权加权最小绝对偏差损失(Least Absolute Deviation, LAD)和平滑剪切绝对偏差(Smoothly Clipped Absolute Deviation, SCAD)惩罚函数进行基因－环境交互作用的变量选择方法。这种方法对于高维和删失的生存数据下的变量选择有很好的稳健性。本文采用的 LAD 损失函数可以适应数据中的污染和长尾误差，从而达到稳健性的目的。此外，SCAD 惩罚函数具有无偏性、稀疏性和连续性的特点，对应的解收敛于局部最小值[11]。三种误差分布下的数据仿真和乳腺癌数据案例研究表明，在基因－环境交互效应的识别效果方面，本文所提出方法的表现优于其他方法。

2. 基于 SCAD 惩罚函数的 AFT 模型

考虑一个包含 n 个独立同分布样本的数据集，其中有 p 个基因效应和 q 个环境/临床效应。对于第 i 个样本，设 Y_i 为感兴趣的响应变量，即生存时间的对数，$X_i=\left(X_{i1},\cdots,X_{ip}\right)$ 为 p 个基因效应，$Z_i=\left(Z_{i1},\cdots,Z_{iq}\right)$ 为 q 个环境效应。我们采用加速失效时间(AFT)模型来描述具有删失性的生存数据。与其他生存分析的模型相比，AFT 模型形式简单、计算代价较低，适合处理高维数据。而且在 AFT 模型下，被估计的回归系数会有更清晰的解释[8]。模型假设：

$$\begin{aligned} E\left(Y_i\right) &= \sum_{k=1}^{q}\alpha_k Z_{ik}+\sum_{j=1}^{p}\gamma_j X_{ij}+\sum_{j=1}^{p}\sum_{k=1}^{q}\beta_{jk}X_{ij}Z_{ik} \\ &= \sum_{k=1}^{q}\alpha_k Z_{ik}+\sum_{j=1}^{p}\left(\gamma_j X_{ij}+\sum_{k=1}^{q}\beta_{jk}X_{ij}Z_{ik}\right) \\ &= \sum_{k=1}^{q}\alpha_k Z_{ik}+\sum_{j=1}^{p}U_{ij}^{\mathrm{T}}b_j \end{aligned} \tag{1}$$

其中，$U_{ij}=\left(X_{ij},X_{ij}Z_{i1},\cdots,X_{ij}Z_{iq}\right)^{\mathrm{T}}$，$b_j=\left(\gamma_j,\beta_{j1},\cdots,\beta_{jq}\right)^{\mathrm{T}}=\left(b_{j0},b_{j1},\cdots,b_{jq}\right)^{\mathrm{T}}$。值得注意的是，这里省略了截距，因为我们假定数据已经经过标准化，这使得响应变量的均值为零。

记 $\alpha=\left(\alpha_1,\cdots,\alpha_q\right)^{\mathrm{T}}$，$b=\left(b_1^{\mathrm{T}},\cdots,b_p^{\mathrm{T}}\right)^{\mathrm{T}}$，$U_i=\left(U_{i1}^{\mathrm{T}},\cdots,U_{ip}^{\mathrm{T}}\right)^{\mathrm{T}}$。对于第 i 个观测样本，符号 b_j 和 U_{ij} 表示第 j 个基因效应对第 i 个观测的所有影响，包括主效应以及和环境/临床因素的交互效应。记 C 为截尾时间的对数。当生存时间为右删失时，令 $Y_{cen}=\min\left\{Y,C\right\}$ 为新的响应变量，$\delta=I\left\{Y\le C\right\}$ 为标记当前样本是否删失的指标。

此外，我们还需要考虑主效应与交互效应的层次结构，以防一个交互作用被确定为相关变量而对应的主效应却不被识别为相关变量的情况出现。现有如下两种层次结构。强层次结构假设只有当对应的主效应都被选择时才能认为其交互作用可能相关。相比之下，弱层次结构假设条件更加宽松，只要确定了至少两种主效应中的其中一种，就可以认为对应的交互作用是相关的。Liu 等认为在基因－环境交互作用

检测研究中，倾向于使用强层次结构，因为环境因素往往起重要作用，且维数低、解释性强，不需要被筛选压缩[8]。若考虑弱层次结构，需要考虑的交互项过多，模型复杂不利于解释。因此，我们采用强层次结构，用以下形式表示：

$$\beta_{jk} \neq 0 \to \gamma_j \neq 0 \text{ and } \alpha_k \neq 0, \text{ for } \forall j,k \tag{2}$$

2.1. 构建目标函数

2.1.1. 损失函数

最小二乘法是一种不稳健的损失函数，因此本文考虑一个具有更强稳健性的 L1 损失函数。我们采用加权 LAD 函数的损失函数，既能达到鲁棒性的目的，又能适应脏数据。本文使用的权重是由 Stute 等人提出的 Kaplan-Meier (K-M)权重[12]。设 $\hat{F}_n$ 为响应变量 Y_{cen} 的分布函数的估计，则 $\hat{F}_n$ 可以被表示为 $\hat{F}_n(y)=\sum_{i=1}^{n} w_{ni} I\{Y_{ceni} \le y\}$。其中，$w_{ni}$ 被定义为：

$$w_{n1} = \frac{\delta_{(1)}}{n}$$

$$w_{ni} = \frac{\delta_{(i)}}{n-i+1} \prod_{l=1}^{i-1} \left(\frac{n-l}{n-l+1} \right)^{\delta_{(i)}}, \ i = 2, \cdots, n \tag{3}$$

然后，计算加权LAD损失函数如下：

$$L(\alpha, b) = \sum_{i=1}^{n} w_{ni} \left| Y_{ceni} - \sum_{k=1}^{q} \alpha_k Z_{ik} - \sum_{j=1}^{p} U_{ij}{}^{\mathrm{T}} b_j \right| \tag{4}$$

2.1.2. 惩罚函数

Fan 和 Li 等人提出了平滑剪切绝对偏差惩罚函数(Smoothly Clipped Absolute Deviation, SCAD)，该惩罚在变量选择中具有良好的性质，例如：连续性、稀疏性、无偏性和高维情况下的渐近 oracle 性质。由于环境/临床因素的选择是基于广泛综合的先验知识，且数据的维度较低，所以不需要对环境/临床效应进行选择。因此，我们只对基因效应及其交互效应添加惩罚项。惩罚函数为：

$$P_\lambda(a, b; \lambda) = \sum_{j=1}^{p} \rho_\lambda \left(\left\| b_j \right\| \right) + \sum_{k=1}^{q} \rho_\lambda \left(\left| b_{jk} \right| \right) \tag{5}$$

其中，$\rho_\lambda(\cdot)$ 是带有调节参数 λ 的惩罚函数，可以通过以下函数计算：

$$\rho_\lambda(t) = \begin{cases} \lambda |t|, & \text{if } |t| < \lambda \\ -\dfrac{|t|^2 - 2a\lambda|t| + \lambda^2}{2(a-1)}, & \text{if } \lambda < |t| < a\lambda \\ \dfrac{(a-1)\lambda^2}{2} + \lambda^2, & \text{if } |t| > a\lambda \end{cases} \tag{6}$$

正如 Fan 和 Li 所讨论的，SCAD 惩罚项满足一个好的惩罚函数应有的三个理想特性：无偏性、稀疏性和连续性。也就是说，通过 SCAD 惩罚函数所得到的估计量将近似无偏，并自动同时将系数的极小值设为零，从而降低最终模型的复杂性，并具有连续性以保持预测的稳定性[12]。此外，Kim 等人还证明了在一定条件下 SCAD 惩罚函数的估计参数与 oracle 惩罚的解渐近等价，即估计参数具有渐近正态性[13]。

2.1.3. 目标函数

我们将上述损失函数和惩罚函数代入如下稳健目标函数：

$$\begin{aligned} Q(\alpha,b) &= L(\alpha,b)+P_\lambda(b) \\ &= \sum_{i=1}^{n} w_{ni}\left|Y_{cen\ i}-\sum_{k=1}^{q}\alpha_k Z_{ik}-\sum_{j=1}^{p}U_{ij}{}^{\mathrm{T}}b_j\right|+\sum_{j=1}^{p}\rho_\lambda\left(\|b_j\|\right)+\sum_{k=1}^{q}\rho_\lambda\left(|b_{jk}|\right) \end{aligned} \tag{7}$$

在这个目标函数中，b_j 为与所有第 j 个基因效应相关变量的系数向量，包括作为第一分量的基因主效应和相应的其它基因－环境交互效应。而 b_{jk} 只作为基因－环境交互效应的系数向量。对于第 j 个基因效应，我们采用群体水平和个体水平上的惩罚项来进行变量选择。在群体水平上，非零的 b_j 表示基因主效应或与环境的交互效应被识别到。而在个体水平上，非零的 b_{jk} 只能说明了交互效应的存在。因此，基因主效应仅在群体水平上受到惩罚约束，而基因与环境的交互效应在群体和个体水平上都受到惩罚。只要一种交互效应被认为是相关的，其对应主效应就会进入到模型中。

2.2. 模型求解

由于目标函数是非凸的，因此采用迭代方法更新基因主效应系数和基因－环境交互系数直到收敛。首先，我们初始化主效应的系数，即 α 和 γ。然后采用 CCCP (Convex-Concave procedure)算法求解 β。其次，同样地，我们固定交互项系数 β 的值，并应用 CCCP 算法更新 α 和 γ 的解。由于我们对惩罚函数施加了层次约束，使得主效应与交互效应之间存在强层次结构。然后重复这两个步骤，直到损失函数值足够小[14]。

An 和 Tao 提出了 CCCP 算法来解决非凸问题。该算法对初始值不敏感，且总是收敛于局部最小值，是一种鲁棒、稳定的算法。在高维情况下，SCAD 可以同时实现模型选择的一致性和最优预测，而 LASSO 则无法做到这一点。具体算法如下：

SCAD 罚函数可以被分解成凹函数和凸函数的和：

$$\rho_\lambda\left(|t|\right)=\tilde{\rho}_\lambda\left(|t|\right)+\mu|t| \tag{8}$$

其中，在 $\tilde{\rho}_\lambda\left(|t|\right)$ 是可微分的凹函数，$|t|$ 是凸函数。然后将目标函数改写为

$$\begin{aligned} Q(\alpha,b) &= L(\alpha,b)+\sum_{j=1}^{p}\left[\tilde{\rho}_\lambda\left(\|b_j\|\right)+\mu\|b_j\|\right]+\sum_{k=1}^{q}\left[\tilde{\rho}_\lambda\left(|b_{jk}|\right)+\mu|b_{jk}|\right] \\ &= L(\alpha,b)+\left[\sum_{j=1}^{p}\tilde{\rho}_\lambda\left(\|b_j\|\right)+\sum_{k=1}^{q}\tilde{\rho}_\lambda\left(|b_{jk}|\right)\right]+\left[\sum_{j=1}^{p}\mu\|b_j\|+\sum_{k=1}^{q}\mu|b_{jk}|\right] \end{aligned} \tag{9}$$

我们注意到 $Q(\alpha,b)$ 的第二部分是一些凹函数的总和，而 $Q(\alpha,b)$ 第三部分是凸函数的总和。设初始化的解为 b^c，然后严格凸上界可以通过下面的式子计算：

$$P(\alpha,b)=L(\alpha,b)+\left[\sum_{j=1}^{p}\nabla\tilde{\rho}_\lambda\left(\|b_j^c\|\right)+\sum_{k=1}^{q}\nabla\tilde{\rho}_\lambda\left(|b_{jk}^c|\right)\right]+\left[\sum_{j=1}^{p}\mu\|b_j\|+\sum_{k=1}^{q}\mu|b_{jk}|\right] \tag{10}$$

函数 $P(\alpha,b)$ 是一个分段二次函数，我们可以通过最小化 $P(\alpha,b)$ 来更新 b^c 当前的值直至收敛。

3. 模拟分析

在基因效应的测定中，考虑到基因表达数据和 SNP 数据同时存在，我们可使用连续变量和分类变量来模拟。具体来说，基因表达数据是使用自回归相关结构从多变量正态分布中生成的。在该结构中，第 i 和第 j 个基因效应在强层次结构下具有相关系数 $\rho_{ij}=0.8^{|i-j|}$。在弱层次下具有相关系数 $\rho_{ij}=0.2^{|i-j|}$。SNP

数据是分位数方法生成的 3 个水平的分类变量。我们根据第一个和第二个三分位数将数据分成三类。然后生成了具有 3 个水平的分类变量 SNP 数据。相似的模拟数据生成方法也被应用于其他的研究中，如 Wu 等人的研究[6]和 Shi 等人的研究[15]。对于环境/临床效应的测定，我们也采用自回归相关结构下的多元正态分布来生成数据。共生成 300 个基因表达因子和 SNP，环境因子 5 个，即主效应 305 个，交互效应 1500 个。生存时间的对数Y由以下数据生成模型计算：

$$Y=\sum_{k=1}^{5}\alpha_k Z_k+\sum_{j=1}^{15}\gamma_j X_j+\sum_{j=1}^{10}\sum_{k=2,3}\beta_{jk}X_j Z_k+\varepsilon \tag{11}$$

假定有 15 个基因主效应和 20 个基因－环境交互项与响应变量(生存时间的对数 Y)相关。式(11)中的系数由均匀分布 Uniform (0.6, 14)生成。删失率假定为 0.3。我们考虑如下三种不同的分布作为误差项的分布：第一种是标准正态分布 $N(0,1)$；第二种是标准正态分布与柯西分布的混合 $0.8*N(0,1)+0.2*Cauchy(0,1)$；第三种是 T 分布 $t(2)$。

除了记为 LAD-SCAD 的我们所提出的方法外，还增加了 5 种备选的方法来作比较，包括 LAD-LASSO、LAD-Group、LS-SCAD、LS-LASSO 和 LS-Group。在 LAD-LASSO 和 LAD-Group 中分别采用 LASSO 惩罚项和 Group LASSO 惩罚项代替 SCAD 惩罚项。二者的区别在于 LASSO 会在群体水平和个体水平上都产生稀疏性，而 Group LASSO 在组内不存在稀疏性、只在群体水平上产生稀疏性。对于 LS-SCAD，保留 SCAD 作为惩罚函数，用 LS (即最小二乘法)代替 LAD 作为损失函数。LS 是一个对异常值敏感的非鲁棒损失函数。以下是数据仿真的结果：

Table 1. Comparison of simulation results of six methods
表 1. 六种方法的模拟结果比较

Error	Method	TP	FP	TP1	FP1	RSSE	PMSE	N_VAR
N(0,1)	LAD-SCAD	18.92 (3.29)	30.54 (17.81)	10.92 (1.69)	14.2 (6.98)	13.23 (2.62)	0.96 (0.18)	49.46 (18.9)
	LAD-LASSO	14.98 (2.38)	29.68 (12.82)	6.98 (0.86)	12.18 (1.82)	10.91 (2.17)	0.77 (0.12)	44.66 (12.8)
	LAD-Group	18.38 (3.51)	67.98 (9.53)	10.38 (1.53)	25.92 (6.55)	15.46 (3.38)	0.82 (0.16)	86.36 (9.5)
	LS-SCAD	16.88 (3.92)	32.62 (11.69)	8.88 (1.09)	9.86 (4.78)	13.32 (2.97)	1.73 (0.26)	49.5 (12.0)
	LS-LASSO	23.16 (4.41)	104.32 (11.58)	15.16 (2.14)	45.4 (9.89)	13.27 (2.88)	1.72 (0.28)	127.48 (12.1)
	LS-Group	22.48 (5.12)	109.06 (13.91)	14.48 (2.15)	46.3 (11.12)	13.64 (3.05)	1.76 (0.24)	131.53 (14.2)
0.8 * N (0, 1) + 0.2 * Cauchy (0,1)	LAD-SCAD	15.3 (2.52)	25.8 (7.55)	7.3 (1.25)	10.28 (3.22)	12.31 (2.69)	0.86 (0.07)	41.1 (7.9)
	LAD-LASSO	14.84 (1.42)	23.72 (2.61)	6.84 (1.42)	11.66 (1.8)	10.29 (2.33)	1.12 (0.12)	38.6 (2.8)
	LAD-Group	14.5 (2.11)	32.22 (7.76)	6.5 (2.11)	14.34 (4.68)	13.62 (3.15)	0.92 (0.09)	46.72 (9.2)
	LS-SCAD	13.2 (1.22)	17.64 (3.3)	5.2 (0.22)	12.28 (3.18)	10.58 (2.56)	2.05 (0.24)	30.64 (3.3)
	LS-LASSO	18.28 (2.94)	59.32 (11.94)	10.28 (2.94)	26.48 (9.23)	10.55 (2.54)	1.99 (0.21)	77.6 (14.4)
	LS-Group	16.72 (2.91)	58.68 (9.87)	10.16 (2.91)	22.32 (8.57)	11.67 (2.82)	1.82 (0.19)	74.4 (12.1)
t(2)	LAD-SCAD	22.8 (3.26)	36.36 (15.55)	12.33 (3.12)	16.28 (6.17)	10.86 (2.57)	1.01 (0.12)	59.16 (18.2)
	LAD-LASSO	19.82 (2.08)	19.06 (4.63)	8.54 (2.78)	11.66 (3.28)	10.63 (2.25)	1.02 (0.09)	38.88 (5.3)
	LAD-Group	22.84 (3.81)	46.16 (11.44)	11.64 (2.99)	24.92 (4.33)	12.98 (2.95)	0.95 (0.10)	69 (12.6)
	LS-SCAD	19.5 (2.51)	21.3 (13.57)	9.38 (2.65)	18.76 (6.32)	10.12 (2.68)	2.14 (0.21)	40.8 (14.7)
	LS-LASSO	29.4 (3.81)	67.96 (11.21)	16.84 (3.26)	27.11 (3.28)	10.09 (2.58)	2.13 (0.16)	97.36 (12.4)
	LS-Group	29.16 (3.35)	65.04 (11.34)	16.27 (3.19)	26.6 (3.81)	11.48 (2.72)	2.13 (0.19)	94.2 (12.8)

仿真结果如表 1 所示。表中所有的值都是 100 次重复模拟的平均值，括号中是标准差的值。TP/FP 代表真/假阳性的总数。TP1/FP1 表示交互作用的真/假阳性的数量。由于本文的目的是检测 G-E 的相互作用效应，因此 TP 和 TP1 是模型选择的关键标准。其他模型性能评估的方法有 PMSE 和 RSSE。PMSE 表示预测均方误差。RSSE 代表平方误差总和的平方根,可以通过公式 $\sqrt{\left\|\hat{\theta}-\theta^0\right\|_2}$ 来计算，其中 $\hat{\theta}$ 和 θ^0 分布代表了系数 $\theta=\left(\alpha^{\mathrm{T}},\gamma^{\mathrm{T}},\beta^{\mathrm{T}}\right)^{\mathrm{T}}$ 的真实值和预测值。所选变量的总数也显示在表中，记为 N_VAR。

在三种误差分布下，比较六个不同模型的 PMSE 的值，我们可以发现带有最小二乘损失的模型，如：LS-SCAD、LS-LASSO 和 LS-Group，具有较大的 PMSE 值(大约为 2)和方差。其他三个带有 LAD 的模型的 PMSE 值(大约为 1)和方差较小。可以看出最小绝对偏差的方法(LAD)更加预测准确和稳健。

另外，我们注意到 LS-LASSO 和 LS-Group 是两个特殊的模型，其 TP 和 FP 值明显高于其他四个模型。虽然这两个模型可以识别出很多相关的效应，但模型中也包含了一些冗余效应，降低了模型的可解释性和简洁性。从表的第一部分我们可以看到，在误差服从标准正态分布的情况下，这两个模型将超过 120 个主效应和交互效应识别为重要的。因而有必要找到一些更有效的方法来压缩备选变量集中的变量。下面将分析其他四种方法。

在误差分布为标准正态分布 $N(0,1)$ 下，即数据未被污染的情况下，LAD-SCAD 和 LAD-Group 表现更好，TP 和 TP1 较高。但在两个模型中，LAD-SCAD 方法的 FP 和 FP1 值较小，它选择出的不必要的变量较少。我们再比较 LAD-LASSO 和 LAD-Group 的 FP 和 FP1 的值，LAD-Group 的值远大于 LAD-LASSO 的，这是因为 LASSO 会在群体水平和个体水平上都产生稀疏性，而 Group LASSO 只在群体水平上产生稀疏性，但在组内不存在稀疏性。Group LASSO 惩罚函数的压缩性较差。

在误差分布为标准正态分布和柯西分布的混合分布下，数据存在污染，LAD-SCAD、LAD-LASSO 和 LAD-Group 三种方法的 TP 和 TP1 值很相近。但 LAD-SCAD 的 PMSE 值(0.86)小于其他几种方法的 PMSE 值(1.12, 0.92, 2.05)，有较好的预测性。

在误差分布为 T 分布 $t(2)$下，LAD-SCAD 和 LAD-Group 的 TP 和 TP1 的值较其他两种误差分布下更高。可以看出，特别是在数据存在重尾分布的样本不平衡情况下，LAD-SCAD 和 LAD-Group 表现出了在基因－环境交互效应选择方面的优越性。但同时 FP 和 FP1 的值也高于其他两个模型(LAD-LASSO 和 LS-SCAD)。但为了更准确地识别交互作用，这仍然是值得的。我们还注意到，在三种误差分布下，LAD-SCAD 的 FP 和 FP1 值都小于 LAD-Group 的，这说明当两种方法都选择了相似数量的正确相关协变量时，LAD-SCAD 方法比 LAD-Group 方法更好。

此外，为了更直观地比较不同模型的变量选择性能，我们实现了数据可视化。

从图 1 可以看出，LAD-SCAD 是一种合适的选择变量的方法，这种方法选择了尽可能多的基因－环境交互效应，有效压缩了变量，预测误差较小。LS-LASSO 和 LS-Group 许多冗余的效应识别为相关的，这降低了模型的可解释性和简单性。LAD-Group 也是一种有效的方法。但是，真阳性数值相近的情况下，但 LAD-SCAD 的假阳性数值比 LAD-Group 少。当两种方法都选择了相似数量的正确相关协变量时，LAD-SCAD 模型选择的冗余变量更少。所以 LAD-SCAD 方法比 LAD-Group 方法更好。

4. 实证研究

本文利用来自北卡罗莱纳大学基因鉴定中心的乳腺癌(Breast Cancer, BRCA)基因表达序列数据集进行实证分析。该数据集共包含 17815 个基因效应和 597 个观测样本。为了便于比较，数据已经过标准化处理。我们根据突变概率从高到低对备选基因效应进行排序。由于突变较高的基因与癌症有较高的关联概率，因此选择前 300 个易突变的基因效应进行后续分析。环境/临床效应数据从癌症基因组图谱联盟下载，该数据由美国国家癌症研究所和美国国家人类基因组研究所共同提供。我们考虑了 5 种环境效应：

确诊年龄、性别、突变计数、基因组改变比例和放射治疗。即我们有 300 个基因效应，5 个环境效应和 1500 个交互效应。

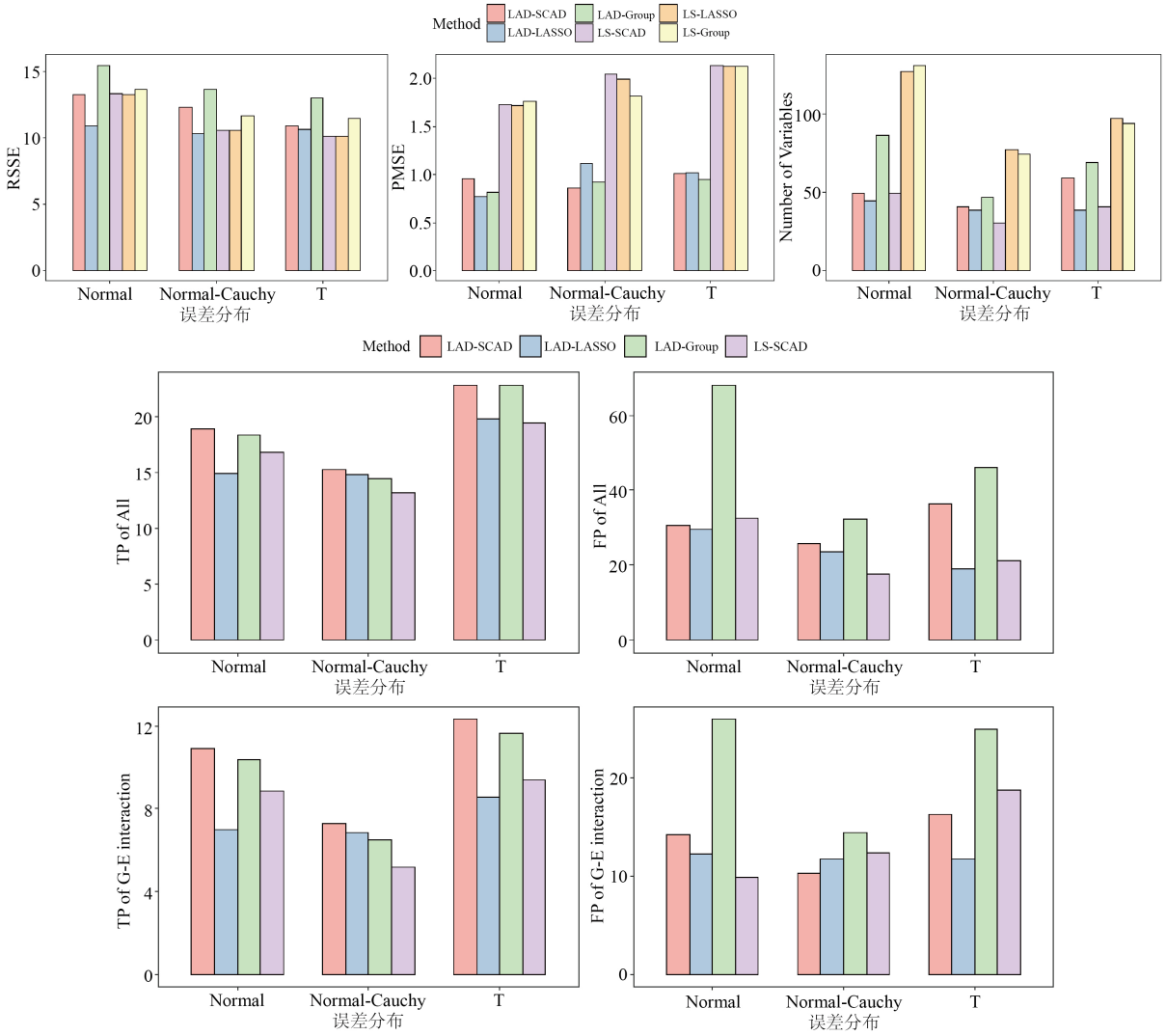

Figure 1. Histogram of simulation results of six methods
图 1. 六种方法的仿真结果直方图

在此，本文采用模拟研究中四种更有效的方法来识别与乳腺癌相关的基因主效应和基因－环境交互效应，并比较其分析结果。LS-LASSO 和 LS-Group 没有参与案例研究，因为这两种方法选择了很多冗余的变量，降低了模型的可解释性。

LAD-SCAD 方法的变量选择结果如下：

如表 2 所示，我们所提出的 LAD-SCAD 方法共检测到 35 种效应，包括 12 种基因主效应和 18 种基因－环境交互效应。环境效应的估计系数分别为−0.095 (确诊年龄)、−0.119 (基因组改变比例)、−0.235 (突变计数)、0.191 (放射治疗)和 0.047 (性别)。根据估计值的正负性质，这些结果是合理的。例如，较年轻的病人和基因组改变和突变较少的病人往往有较长的生存期。放射治疗对延长生存期有积极作用。与基因产生了最多交互项的环境变量是放射治疗，即放射治疗与多个基因效应共同作用对乳腺癌产生了影响。

Table 2. Results of variable selection analysis of BRCA data
表 2. 乳腺癌数据的变量选择分析结果

基因名称	主效应	与环境的交互效应				
		确诊年龄	基因组改变比例	突变计数	放射治疗	性别
COL11A1	0.334					
USP9X	0.346		−0.015		0.056	
UTRN	0.376				0.033	0.059
IREB2	−0.904		0.155			
ADAMTS20	0.017					0.008
TAF1	−0.13	0.007		0.052	−0.033	
MYH9	0.028	−0.15	−0.059		0.005	
RANBP2	0.384		−0.182		0.147	
PCLO	0.247					
WNK3	0.03					
USP34	0.034				0.131	
RUNX1	−0.092	0.03		0.315	−0.023	

其中有些基因效应已被证实与乳腺癌是相关的。例如，COL11A1 的基因表达与乳腺癌的发生和进展过程相关，它的 mRNA 水平用于乳腺癌诊断预后的灵敏度高达 81.31% [16]。USP9X 可以影响乳腺癌细胞的生长和侵袭调节中心体的复制，为寻找潜在的乳腺癌治疗干预的靶点提供参考[17]。UTRN 基因表达被抑制后，细胞生长增殖的速度加快，所以 UTRN 对乳腺癌的干预治疗可以提供切入点[18]。MYH9 基因编码一种非肌细胞肌球蛋白 II，影响乳腺癌细胞的生长、迁移、黏附、侵袭能力[19]。

除了 LAD-SCAD 方法，我们还使用其他三个模型分析了 BRCA 数据。如表 3 所示，LAD-LASSO、LAD-Group 和 LS-SCAD 识别出的变量数分别为 27 (包括 13 个交互作用)、30 (包括 14 个交互作用)和 38 (包括 24 个交互作用)。

Table 3. Result of BRCA data by four models
表 3. 四种模型对乳腺癌数据的分析结果

	RMSE	N_Var	N_Gene	N_Interaction
LAD-SCAD	1.189	35	12	18
LAD-LASSO	1.212	27	9	13
LAD-Group	1.214	44	11	34
LS-SCAD	1.368	38	9	24

从表 3 中可以看出，我们所提出的 LAD-SCAD 的预测误差是最小的，较为准确。LAD-LASSO 和 LAD-Group 两种方法的预测误差较为相近，但由于 LAD-LASSO 比 LAD-Group 多在个体水平上产生了稀疏性，LAD-LASSO 所选变量较少，模型更简洁。而 LAD-Group 选出了过多的与响应变量无关的冗余变量，降低了模型的可解释性。LS-SCAD 的预测变量最大，准确性不如其他三个模型高。

5. 结论

在本篇文章中，我们提出了一个满足“主效应－交互项”强层次结构的交互作用识别方法。与现有的交互项选择方法相比较，我们根据生存数据不平衡的删失特性采用了加权 LAD 损失函数，并对群体水平和个体水平都添加了 SCAD 惩罚函数使得选出的效应服从强层次结构。然后采用 CCCP 算法对目标函数进行优化，从而得到主效应和交互项的系数估计。我们进行了数据模拟，在三种不同的误差分布下比较了该模型与其他五个模型的多个指标，从结果可以看出，与其他模型相比，我们提出的方法预测更准确、得到的模型更简洁。我们还收集了乳腺癌的基因表达数据(BRCA)，利用所提出方法选出与乳腺癌相关的基因效应与其对应交互效应。仿真研究和实例研究都说明了该方法能准确、稳健地选择出合适的基因效应和基因－环境交互效应，且我们的方法能有效压缩备选的变量，选出的模型简洁、有较好的解释性。

基金项目

国家自然科学基金项目“高维数据变量间非线性交互作用的研究”(11571009)；山西省应用基础研究计划“复杂环境下肿瘤免疫系统的动力学建模及致病基因识别”(201901D111086)。

参考文献

[1] Hedelin, M., Chang, E.T. and Wiklund, F. (2010) Association of Frequent Consumption of Fatty Fish with Prostate Cancer risk Is Modified by COX-2 Polymorphism. *International Journal of Cancer*, **120**, 398-405. https://doi.org/10.1002/ijc.22319

[2] Thomas, D. (2010) Gene-Environment-Wide Association Studies: Emerging Approaches. *Nature Reviews Genetics*, **11**, 259-272. https://doi.org/10.1038/nrg2764

[3] Xu, Y., Wu, M., Ma, S. and Ejaz Ahmed, S. (2018) Robust Gene-Environment Interaction Analysis Using Penalized Trimmed Regression. *Journal of Statistical Computation and Simulation*, **88**, 3502-3528. https://doi.org/10.1080/00949655.2018.1523411

[4] Corella, D., Peloso, G. and Arnett, D.K. (2009) APOA2, Dietary Fat, and Body Mass Index: Replication of a Gene-Diet Interaction in 3 Independent Populations. *Archives of Internal Medicine*, **169**, 1897-1906. https://doi.org/10.1001/archinternmed.2009.343

[5] Wu, C. and Ma, S. (2015) A Selective Review of Robust Variable Selection with Applications in Bioinformatics. *Brief Bioinformatics*, **16**, 873-883. https://doi.org/10.1093/bib/bbu046

[6] Wu, C., Jiang, Y., Ren, J., Cui, Y. and Ma, S. (2018) Dissecting Gene-Environment Interactions: A Penalized Robust Approach Accounting for Hierarchical Structures. *Statistics in Medicine*, **37**, 437-456. https://doi.org/10.1002/sim.7518

[7] Zhu, R., Zhao, H. and Ma, S. (2014) Identifying Gene-Environment and Gene-Gene Interactions Using a Progressive Penalization Approach. *Genetic Epidemiology*, **38**, 353-368. https://doi.org/10.1002/gepi.21807

[8] Ma, S. (2012) Identification of Gene-Environment Interactions in Cancer Prognosis Studies Using Penalization. *Memorie Della Societa Astronomica Italiana*, **80**, 824.

[9] Ren, J., Zhou, F., Li, X., *et al*. (2019) Semiparametric Bayesian Variable Selection for Gene-Environment Interactions. *Statistics in Medicine*, **39**, 617-638. https://doi.org/10.1002/sim.8434

[10] Xu, Y., Wu, M., Zhang, Q. and Ma, S. (2018) Robust Identification of Gene-Environment Interactions for Prognosis Using a Quantile Partial Correlation Approach. *Genomics*, **111**, 1115-1123. https://doi.org/10.1016/j.ygeno.2018.07.006

[11] Fan, J. and Li, R. (2001) Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348-1360. https://doi.org/10.1198/016214501753382273

[12] Stute, W. (1996) Distributional Convergence under Random Censorship When Covariables Are Present. *Scandinavian Journal of Statistics*, **23**, 461-471.

[13] Kim, Y., Choi, H. and Oh, H. (2008) Smoothly Clipped Absolute Deviation on High Dimensions. *Journal of the American Statistical Association*, **103**, 1665-1673. https://doi.org/10.1198/016214508000001066

[14] Kim, G., Lai, C.-Q., Arnett, D.K., *et al*. (2018) Detection of Gene-Environment Interactions in a Family-Based Population Using SCAD. *Statistics in Medicine*, **37**, 506. https://doi.org/10.1002/sim.7537

[15] Shi, X., Liu, J. and Huang, J. (2014) A Penalized Robust Method for Identifying Gene-Environment Interactions. *Genetic Epidemiology*, **38**, 220-230. https://doi.org/10.1002/gepi.21795

[16] 王玉丽. 乳腺癌发生发展中 COL11A1 和 MMP3 基因表达的动态改变及其诊断和预后价值的研究[D]: [硕士学位论文]. 天津: 天津医科大学, 2007.

[17] 宋囡. 蛋白质去泛素化酶 USP9X 促进乳腺癌发生发展的分子机理研究[D]: [博士学位论文]. 天津: 天津医科大学, 2018.

[18] 高勇. RNAi 技术探讨 UTRN 基因在癌症发生过程中的作用[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2007.

[19] 吴珊珊. 非肌细胞肌球蛋白重链 IIA 亚型在乳腺癌转移中的作用[D]: [硕士学位论文]. 长春: 吉林大学, 2012.