基于基因筛选的结肠癌预测

一、引言

全基因组关联分析(Genome-wide association studies, GWAS)是通过剖析生命体全基因组单核苷酸多态性检测(Single Nucleotide Polymorphism, SNP)位点的基因型信息和相关疾病表型信息来揭露复杂疾病致病基因的有效途径。

传统 GWAS 采用的主要模式是疾病与单个 SNP 位点相关统计分析的方法，但是人类的复杂疾病往往是多基因交互作用(Gene-gene interactions)的结果。大量研究结果表明，乳腺癌、结肠癌、糖尿病和冠心病等人类常见疾病多与基因交互作用密切相关，而基于单个 SNP 位点的统计方法可能无法探测到所有的基因交互作用。探测基因交互作用有助于基因功能的识别，对于发现隐藏的药物靶标和人类复杂疾病的遗传机制尤为突出。

随着基因技术的快速发展，常见疾病表型信息及相关个体的全基因组基因型信息呈爆发性增长，为了发现具有统计显著性的与疾病相关的基因交互项，所需要进行检测的基因及基因交互变量数量更是呈指数增长，采用简单的穷尽策略从大规模全基因组数据中搜索出隐藏的多基因交互作用显然是低效且几乎不可能实现的。与此同时，有关癌症基因的筛选面临的另外一个问题是可供研究的样本数量极少，因此癌症相关基因的筛选通常都是超高维数据问题。如何在超高维度下有效地筛选出复杂疾病的致病基因是一个值得研究的问题，我们此次研究的课题主要考虑的是包含基因-基因交互项的致病基因筛选。

本文首先介绍了高维及超高维下包含交互项的基因筛选问题的相关定义，并介绍了几种不同的筛选方法的原理和理论支撑，对上述方法通过调包或代码复现的方式进行结果对比，选出其中表现最优的方法在超高维度的结肠癌数据下进行了结肠癌的基因筛选及预测。

二、背景知识

2.1 层次结构

从$(\mathbf{X}, \mathrm{Y})$中独立同分布地抽取数据集$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$，其中$\mathbf{X} = (\mathrm{X}_1, \mathrm{X}_2, \cdots, \mathrm{X}_\mathrm{p})^{\mathrm{T}}$是 p 维向量，Y是响应变量，则标准的线性回归模型可以写作

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X}_1 + \cdots \beta_\mathrm{p} \mathrm{X}_\mathrm{p} + \varepsilon, \tag{1}$$

而考虑交互项后的模型可以写作

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X}_1 + \cdots \beta_\mathrm{p} \mathrm{X}_\mathrm{p} + \gamma_{12} \mathrm{X}_1 \mathrm{X}_2 + \cdots + \gamma_{(\mathrm{p}-1)\mathrm{p}} \mathrm{X}_{\mathrm{p}-1} \mathrm{X}_\mathrm{p} + \varepsilon, \tag{2}$$

其中ε是随机误差项，$\beta_0, \boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T, \boldsymbol{\gamma} = (\gamma_{11}, \gamma_{12}, \cdots, \gamma_{pp})^T$是回归系数。在模型(2)中$X_1, \cdots, X_p$被称作主效应，而$X_j X_k (1 \le j < k \le p)$被称作交互项，其中交互项$X_j X_k$是主效应$X_j$和$X_k$的子系，$X_j$和$X_k$是$X_j X_k$的父系。层次结构的含义是，如果子系出现在了模型中，则其父系也应该出现在模型中。

Nelder(1977[1])和 Peixoto(1990[2])的研究表明，在包含基因-基因交互项的模型中进行重要变量识别时，有必要保持这种层次结构，即只有在主效应进入模型的情况下，才需要考虑交互项是否重要。他们认为，一个合理的模型在经过简单的变换后应该保持不变性。例如考虑模型$f(x_1, x_2) = \beta_0 + \gamma_{12} x_1 x_2$，如果将$x_1$变换为$\tilde{x}_1 = x_1 - 1$，则模型会变为$f(\tilde{x}_1, x_2) = \beta_0 + \gamma_{12} x_2 + \gamma_{12} \tilde{x}_1 x_2$，可见，仅考虑交互项和常数项的模型是不合理的，只要对变量进行简单的变换，主效应就会出现在模型中。直观上，由于主效应可以被视为与全局平均值的偏差，而交互作用是与主效应的偏差，因此没有主效应的交互作用几乎没有意义。

对模型(2)而言，交互项的两个父系均作为主效应被纳入模型，或交互项的其中一个父系作为主效应被纳入模型，均可被视作保持了模型的层次结构，通常将前者称作强层次结构，将后者称作弱层次结构。可以将这两个层次结构的定义写作：

强层次结构：$\gamma_{jk} \neq 0 \Rightarrow \beta_j \neq 0$ *且* $\beta_k \neq 0$；

弱层次结构：$\gamma_{jk} \neq 0 \Rightarrow \beta_j \neq 0$ *或* $\beta_k \neq 0$。

本文考虑的几种算法，均为在强层次结构下对遗传基因进行筛选。

2.2 衡量变量重要性

在维度 p 较高的情况下，稀疏性假定通常是较为合理的假定，即仅有少数解释变量会对响应变量产生影响。对于标准线性回归模型(1)，变量X_j重要等价于

$\beta_j \neq 0$，因此系数向量$\boldsymbol{\beta}$的支撑集可以记作$S(\boldsymbol{\beta}) = \{j: \beta_j \neq 0, j =, \cdots, p\}$。很容易可以证明，$S(\boldsymbol{\beta})$的定义是满足不变性的。

但是对于模型(2)，由于需要考虑不变性，$\beta_j = 0$不再意味着主效应X_j不重要。考虑如下例子的三种等价表达形式：

$$Y = X_1 X_2 + X_3 + \epsilon,$$

$$Y = \widetilde{X_1} X_2 + X_2 + X_3 + \epsilon,$$

$$Y = \widehat{X_1} X_2 - X_2 + X_3 + \epsilon,$$

其中 $\widetilde{X_1} = X_1 - 1$， $\widehat{X_1} = X_1 + 1$，可见对第一个主效应做了三种简单的平移变换后，第二个主效应对应的系数分别为 0、1、-1，并会产生对第二个主效应的三种不同的解释方式，即第二个主效应与响应变量分别成不相关、正相关和负相关关系。会出现这样的情况，是因为交互项$X_1 X_2$的存在。一般而言，只要$\gamma_{jk} \neq 0$，就会存在可以影响β_j和β_k的符号的线性变换。

在模型(2)的情况下，Hao 和 Zhang(2017[3])提出了衡量变量重要性的新的定义：

$$\text{主效应} X_j \text{是重要的} \quad \Leftrightarrow \quad \beta_j^2 + \sum_{k=1}^{p} \gamma_{jk}^2 > 0,$$

$$\text{交互项} X_j X_k \text{是重要的} \quad \Leftrightarrow \quad \gamma_{jk} \neq 0。$$

很容易证明，在模型(2)下对主效应和交互项重要性的定义在简单的线性变换下是满足不变性的，因此这种定义方式是更为合理的定义。

对X_j进行简单线性变换，令$\widetilde{X_j} = a_j(X_j - c_j)(a_j > 0)$，则模型（2）可以写作

$$Y = \left(\beta_0 + \sum_{j=1}^{p} \beta_j c_j + \sum_{1 \leq j < k \leq p} \gamma_{jk} c_j c_k\right) + \sum_{j=1}^{p} a_j^{-1} \left(\beta_j + \sum_{k=1}^{p} \gamma_{jk} c_k\right) \widetilde{x_j} + \sum_{1 \leq j < k \leq p} \gamma_{jk} a_j^{-1} a_k^{-1} \widetilde{x_j}\, \widetilde{x_k},$$

则

$$\widetilde{\beta_0} = \beta_0 + \sum_{j=1}^{p} \beta_j c_j + \sum_{1 \leq j < k \leq p} \gamma_{jk} c_j c_k,$$

$$\widetilde{\beta_j} = \beta_j + \sum_{k=1}^{p} \gamma_{jk} c_k ,$$

$$\widetilde{\gamma_{jk}} = \gamma_{jk} a_j^{-1} a_k^{-1},$$

可以看出，

(1) $\beta_j^2 + \sum_{k=1}^{p} \gamma_{jk}^2 = 0 \Leftrightarrow \beta_j = 0, \gamma_{jk} = 0, \forall j,k \Leftrightarrow \widetilde{\beta_j} = 0, \widetilde{\gamma_{jk}} = 0, \forall j,k$

$$\Leftrightarrow \widetilde{\beta_j^2} + \sum_{k=1}^{p} \widetilde{\gamma_{jk}^2} = 0,$$

(2) $sign(\widetilde{\gamma_{jk}}) = sign(\gamma_{jk})$,

三、几种不同的用于高维或超高维数据下基因筛选的算法

3.1 直接筛选

3.1.1 Lasso for hierarchical interactions[4]

$$\underset{\beta_0 \in R, \boldsymbol{\beta} \in R^p, \boldsymbol{\gamma} \in R^{p \times p}}{\text{Min}} q(\beta_0, \boldsymbol{\beta}, \boldsymbol{\gamma}) + \lambda \|\boldsymbol{\beta}\|_1 + \frac{\lambda}{2} \|\boldsymbol{\gamma}\|_1 \tag{3}$$

$$s.t.\ \boldsymbol{\gamma} = \boldsymbol{\gamma}^T, \left\|\boldsymbol{\gamma_j}\right\|_1 \le |\beta_j|, j = 1, \cdots, p$$

其中，$q(\cdot)$为损失函数，通常取$\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \beta_0 - x_i^T\boldsymbol{\beta} - \frac{1}{2}x_i^T\boldsymbol{\gamma}x_i\right)^2$；$\|\cdot\|_1$为$L_1$范数，$\|\boldsymbol{\gamma}\|_1 = \sum_{j \ne k}|\gamma_{jk}|$；$\boldsymbol{\gamma_j}$为$\boldsymbol{\gamma}$的第 j 行或第 j 列。这一算法比一般的 lasso 多出了一个限制条件，而这个限制条件可以保证强层次结构得到满足。如果$\gamma_{jk} \ne 0$，则$\left\|\boldsymbol{\gamma_j}\right\|_1 > 0$且$\|\boldsymbol{\gamma_k}\|_1 > 0$，由于$\left\|\boldsymbol{\gamma_j}\right\|_1 \le |\beta_j| (j = 1, \cdots, p)$，可得$\beta_j \ne 0$且$\beta_k \ne 0$，这刚好满足强层次结构的定义。

3.1.2 Hierarchical group-lasso regularization[5]

为保持强层次结构，可以使用 group-lasso 的方法，将交互项及其两个父系主效应看作一个组，对这个组共同施加惩罚和进行系数估计。由于某个主效应可以是很多交互项的父系，因此不同组之间可能会出现主效应的重叠，这种方法也被

称作 overlapped-group lasso。为了更清楚地说明这一问题，我们考虑仅有两个主效应$\mathbf{Z_1}$和$\mathbf{Z_2}$的情况。令$\mathbf{Z_1}$单独成一组，同时与$\mathbf{Z_2}, \mathbf{Z_1} * \mathbf{Z_2}$合并成一组，不同组的$\mathbf{Z_1}$系数不同。

$$\min_{\mu,\tilde{\mu},\alpha,\tilde{\alpha}} \frac{1}{2}\left\| Y - \mu 1^T - Z_1\alpha_1 - Z_2\alpha_2 - [1, Z_1, Z_2, Z_1 * Z_2]\begin{bmatrix}\tilde{\mu}\\ \tilde{\alpha}_1\\ \tilde{\alpha}_2\\ \alpha_{1:2}\end{bmatrix}\right\|_2^2$$
$$+ \lambda\left(|\alpha_1| + |\alpha_2| + \sqrt{\tilde{\mu}^2 + \tilde{\alpha}_1^2 + \tilde{\alpha}_2^2 + \alpha_{1:2}^2}\right)$$

主效应的系数分别为$\widehat{\boldsymbol{\theta}}_{\mathbf{1}} = \widehat{\boldsymbol{\alpha}}_{\mathbf{1}} + \widehat{\tilde{\boldsymbol{\alpha}}}_{\mathbf{1}}$，$\widehat{\boldsymbol{\theta}}_{\mathbf{2}} = \widehat{\boldsymbol{\alpha}}_{\mathbf{2}} + \widehat{\tilde{\boldsymbol{\alpha}}}_{\mathbf{2}}$，交互效应的系数为$\widehat{\boldsymbol{\theta}}_{\mathbf{1:2}} = \widehat{\boldsymbol{\alpha}}_{\mathbf{1:2}}$。若$\sqrt{\tilde{\boldsymbol{\mu}}^{\mathbf{2}} + \tilde{\boldsymbol{\alpha}}_{\mathbf{1}}^{\mathbf{2}} + \tilde{\boldsymbol{\alpha}}_{\mathbf{2}}^{\mathbf{2}} + \boldsymbol{\alpha}_{\mathbf{1:2}}^{\mathbf{2}}}} \neq 0$,则$\widehat{\tilde{\boldsymbol{\alpha}}}_{\mathbf{1}} \neq 0$，$\widehat{\tilde{\boldsymbol{\alpha}}}_{\mathbf{2}} \neq 0$，$\widehat{\boldsymbol{\theta}}_{\mathbf{1}} \neq 0$，$\widehat{\boldsymbol{\theta}}_{\mathbf{2}} \neq 0$。因此可以看出，group-lasso 是满足强层次结构的。

3.1.3、Threshold Gradient Descent Regularization(TGDR)[6]

TGDR 需要输入被解释变量$Y_i\left(i = 1, 2, \cdots n\right), Y_i = 0$或$Y_i = 1$，解释变量$X_{n\times p}$，$l\left(\alpha, \Theta\right)$是关于参数$\alpha$和$\Theta$的对数似然函数，输出结果为$\alpha$，$\Theta$。以下为具体算法：

第一步：初始化$\alpha^0 = 0_{p\times 1}$，$\Theta^0 = 0_{p\times p}$,迭代次数M；

第二步：在第t步迭代中，计算$\frac{\partial l}{\partial \alpha_i}^{(t)}\left(i = 1, 2, \cdots p\right)$，$\frac{\partial^2 l}{\partial \theta_i \theta_j}^{(t)}\left(i, j = 1, 2, \cdots p, i \neq j\right)$；

第三步：对每个参数α_i和θ_{ij}，考虑是否更新：

$$g\left(\alpha_i\right)^{(t)} = I\left(\left|\frac{\partial l}{\partial \alpha_i}^{(t)}\right| > \tau \max\left|\frac{\partial l}{\partial \alpha_i}^{(t)}\right|\right),$$

$$g\left(\theta_{ij}\right)^{(t)} = I\left(\left|\frac{\partial^2 l}{\partial \theta_i \theta_j}^{(t)}\right| > \tau \max\left|\frac{\partial^2 l}{\partial \theta_i \theta_j}^{(t)}\right|\right)$$其中τ为更新阈值(threshold)，

记 $g\left(\alpha^{(t)}\right)=\left(g\left(\alpha_1^{(t)}\right),g\left(\alpha_2^{(t)}\right),\cdots g\left(\alpha_p^{(t)}\right)\right)^T, g\left(\Theta^{(t)}\right)=\begin{pmatrix} 0 & \cdots & g\left(\theta_{1p}^{(t)}\right) \\ \vdots & \ddots & \vdots \\ g\left(\theta_{p1}^{(t)}\right) & \cdots & 0 \end{pmatrix}$；

第四步：满足强层次性(strong hierarchy)：若 $g\left(\Theta^{(t)}\right)_{ij}=1$，则 $g\left(\alpha\right)_i=1$ 且 $g\left(\alpha\right)_j=1$；

第五步：更新参数：$\alpha^{(t+1)}=\alpha^{(t)}-\delta g\left(\alpha^{(t)}\right)\odot\alpha^{(t)}$，$\Theta^{(t+1)}=\Theta^{(t)}-\delta g\left(\Theta^{(t)}\right)\odot\Theta^{(t)}$，其中 δ 为迭代步长，$\odot$ 为哈达玛积(Hadamard product),表示矩阵（或向量）对应元素相乘。

3.2 两阶段筛选方法

为保持模型的层次结构，上述筛选方法会添加一些特殊的惩罚约束。在超高维数据下，需要解决大规模复杂的优化问题，计算成本会非常高。现实中，通常会采用两阶段筛选方法进行基因筛选，第一步仅考虑对主效应进行回归，第二步对第一步筛选出的重要主效应及其交互项采用以上直接筛选的方法进行基因筛选。两阶段的方法可以在第一步将维度大大降低，使得计算变得高效。另外，Hao 和 Zhang(2017[3])证明了两阶段方法在某些条件下可以保证估计的一致性。

令$\tau(\boldsymbol{\beta},\boldsymbol{\gamma})=\{j:\beta_j^2+\sum_{k=1}^{p}\gamma_{jk}^2>0,j=1,\cdots,p\}$表示模型（2）情况下的重要主效应。

（1）证明$\mathcal{S}(\boldsymbol{\beta})=\tau(\boldsymbol{\beta},\boldsymbol{\gamma})$：

在强层次结构的约束下，$\beta_j^2+\sum_{k=1}^{p}\gamma_{jk}^2>0\Leftrightarrow\beta_j\neq0$，即$\mathcal{S}(\boldsymbol{\beta})=\tau(\boldsymbol{\beta},\boldsymbol{\gamma})$。

（2）证明$\mathcal{S}(\boldsymbol{\beta})=\mathcal{S}(\widetilde{\boldsymbol{\beta}})$：

首先，不失一般性，假设$E(Y)=0,E\left(X_j\right)=0,j=1,2,\cdots,p$，$Z_{jk}=X_jX_k-E(X_jX_k)$，则模型（2）等价于$Y=\beta_1X_1+\cdots+\beta_pX_p+\gamma_{11}Z_{11}+\cdots+\gamma_{pp}Z_{pp}+\epsilon$，

则$\left(X_1,\cdots,X_p,Z_{11},\cdots,Z_{pp}\right)$的协方差矩阵可表示为$\Sigma=\begin{pmatrix}\Sigma^{(1)} & 0\\ 0 & \Sigma^{(2)}\end{pmatrix}$，其中$\Sigma^{(1)}$和$\Sigma^{(2)}$分别为$\left(X_1,\cdots,X_p\right)$和$\left(Z_{11},\cdots,Z_{pp}\right)$的协方差矩阵。

定义$\omega = \gamma_{11}Z_{11} + \cdots + \gamma_{pp}Z_{pp} + \epsilon$，则$cov(\omega, X_j) = 0$。假设$\boldsymbol{\beta}^*$为真实参数，则：

$$\widehat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} E\left(Y - \sum_{j=1}^{p} X_j\beta_j\right)^2$$

$$= \underset{\beta}{\operatorname{argmin}} E\left(\sum_{j=1}^{p} X_j\beta_j^* + \omega - \sum_{j=1}^{p} X_j\beta_j\right)^2$$

$$= \underset{\beta}{\operatorname{argmin}} E\left[\left(\sum_{j=1}^{p} X_j\beta_j^* - \sum_{j=1}^{p} X_j\beta_j\right)^2 + \omega^2\right] = \boldsymbol{\beta}^*。$$

有了对两阶段方法的理论支撑，我们可以将 3.1 节中的三种直接筛选方法引入两阶段方法中。即第一阶段采用向前回归的方法对主效应进行筛选，第二阶段分别采用 lasso、group lasso、TGDR 算法进行模型拟合。

尽管以上两阶段方法可以在很大程度上降低计算成本，但在第二阶段时，由于引入了交互项，变量维度仍然很高，计算成本仍然很大。

iFORT 算法[7]对以上问题进一步进行了改进。iFORT 算法本质上也是一种两阶段方法，第一阶段同样使用向前回归的算法对主效应进行筛选。其优势在于，第二阶段对纳入交互效应的模型进行拟合时，没有选择施加复杂的惩罚项的方式，而是使用了较为简单的方法来保证模型满足强层次结构。接下来，简单介绍 iFORT 算法。

第一步：令$\mathcal{C} = \{1,2,\cdots,p\}$，对下标在$\mathcal{C}$内的主效应使用向前回归，记录解路径为$\{\mathcal{S}_t^{(1)}, t = 1,2,\ldots\}$，将筛选出的主效应的下标记为$\widehat{\mathcal{M}} = \{j_1,\ldots,j_{t_1}\}$；

第二步：更新$\mathcal{C} = \widehat{\mathcal{M}} \cup \{(k,l): k \in \widehat{\mathcal{M}}, l \in \widehat{\mathcal{M}}\}$，对下标在$\mathcal{C}$内的主效应和交互项使用向前回归。在这一步向前回归中强制将下标在$\widehat{\mathcal{M}}$中的主效应纳入模型中，记录解路径为$\{\mathcal{S}_{t_1+t}^{(2)}, t = 1,2,\ldots\}$。

3.3 无缝路径算法

两阶段方法虽然计算效率高，但是也存在很大的局限性。例如，第一阶段仅

考虑主效应而不考虑交互效应，则交互效应就被视作错误设定的模型的噪声，这个噪声水平可能相当高，会影响对较弱的主效应的识别，从而导致部分主效应在第一阶段无法被正确识别出来。有两种无缝路径算法，可以在筛选的过程中，既避免过高维度的计算，又同时考虑受层次结构影响的主效应和交互项的识别。

3.3.1 iFORM[7]

首先介绍必要的记号，在第 t 步中，记$\mathcal{S}_t$为解路径，$\mathcal{M}_t$为当前所选主效应，$\mathcal{C}_t$包含全部主效应及所选主效应形成的全部交互项。接下来具体介绍算法，

第一步：初始化：令$\mathcal{S}_0 = \emptyset, \mathcal{M}_0 = \emptyset, \mathcal{C}_0 = \{1,2,\cdots,p\}$；

第二步：对于$t = 1,2,\cdots,D$，重复以下步骤：

在第 t 步中，给定$\mathcal{S}_{t-1}, \mathcal{M}_{t-1}, \mathcal{C}_{t-1}$，用向前回归从下标在$\mathcal{C}_{t-1}\backslash\mathcal{S}_{t-1}$内的主效应及交互项中选择一个变量进入模型，记该变量下标为a。，将这个变量加入，更新$\mathcal{S}_t = \mathcal{S}_{t-1}\cup\{a\}$。如果新选择出的变量为主效应，将这个主效应加入更新$\mathcal{M}_t = \mathcal{M}_{t-1}\cup\{a\}$，$\mathcal{C}_t = \{1,2,\cdots,p\}\cup\{(k,l): k,l \in \mathcal{M}_t\}$；否则$\mathcal{M}_t = \mathcal{M}_{t-1}$，$\mathcal{C}_t = \mathcal{C}_{t-1}$；

第三步：记录第二步中的解路径$\{\mathrm{S_t}: \mathrm{t} = 1,2,\ldots,\mathrm{D}\}$。

3.3.2 RAMP[8]

边际准则下的正则化算法(Regularization algorithm under marginality principle, RAMP)是一种通过坐标下降法计算解路径，同时在解路径上保持模型层次结构的算法。

首先介绍必要的记号。在算法的第 k-1 步中，将当前所有活跃的主效应的集合记为$\mathcal{M}_{k-1}$，将当前所有活跃的交互效应的集合记为$\mathcal{I}_{k-1}$，将$\mathcal{I}_{k-1}$中所有交互项的父系主效应的集合记为$\mathcal{H}_{k-1}$，令$\mathcal{H}_{k-1}^C = \mathcal{M} - \mathcal{H}_{k-1}$。令$\lambda_{max} = n^{-1}\max|X^T y|$，令$\lambda_{min} = \xi\lambda_{max}$，其中$\xi$是一个足够小的正数，生成一个递减的序列$\lambda_{max} = \lambda_1 > \lambda_2 > \cdots > \lambda_K = \lambda_{min}$。

第一步：对参数进行初始化，取$\mathcal{M}_0 = \emptyset, \mathcal{I}_0 = \emptyset$。然后对于$k = 1,\cdots,K$，重复以下步骤：

第二步：给定$\mathcal{M}_{k-1}$，$\mathcal{I}_{k-1}$，$\mathcal{H}_{k-1}$，将$\mathcal{M}_{k-1}$中主效应所可能产生的所有交互项加入模型，并关于$\left(\beta_0, \beta_{\mathcal{M}}^T, \beta_{\mathcal{M}_{k-1}^{\circ 2}}\right)^T$最小化以下损失函数：

$$\frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \beta_0 - x_i^T\beta_{\mathcal{M}} - (x_i^T)_{\mathcal{M}_{k-1}}^{\circ 2}\beta_{\mathcal{M}_{k-1}^{\circ 2}}\right)^2 + \lambda_k\left\|\beta_{\mathcal{H}_{k-1}^C}\right\|_1 + \lambda_k\left\|\beta_{\mathcal{M}_{k-1}^{\circ 2}}\right\|_1,$$

其中$\mathcal{M}_{k-1}^{\circ 2} = \mathcal{M}_{k-1} \circ \mathcal{M}_{k-1} = \{(j,k)\colon\ j \le k;\ j,k \in \mathcal{M}\}$，$X^{\circ 2} = X \circ X$为一个由$X$中所有成对列向量乘积构成的$n \times \frac{p(p-1)}{2}$的矩阵。根据模型拟合结果记录第 k 步的解路径$\mathcal{M}_k$，$\mathcal{I}_k$，$\mathcal{H}_k$，为保证强层次结构，还需要将$\mathcal{I}_k$中所有交互效应的父系主效应（即$\mathcal{H}_k$中的元素）加入$\mathcal{M}_k$。与两阶段筛选方法不同的是，RAMP 的每一次迭代都允许交互项加入模型。

四、方法对比与结果展示

我们采用了几种不同的筛选方法，包括 FS-Lasso、FS-Group Lasso、FS-TGDR、iFORT、iFORM、RAMP 方法。其中两阶段筛选方法的第一阶段均采用向前回归的方法，第二阶段分别采用 lasso for hierarchical interactions、hierarchical group-lasso regularization 和 tgdr 算法。为对比不同方法的效果优劣，我们首先通过数值模拟的方式对不同方法的表现结果进行展示。

首先介绍模型设定。取$n = 200, p = 1000$，假设$\mathrm{X} \sim \mathrm{N_p}(0,\Sigma)$，其中$\Sigma = \mathrm{cov}(\mathrm{X_j},\mathrm{X_k}) = 0.5^{|\mathrm{j}-\mathrm{k}|}\ (1 \le \mathrm{j,k} \le \mathrm{p})$。假设$\mathrm{P(Y=1)} = \frac{\mathrm{e}^{\eta}}{1+\mathrm{e}^{\eta}}$，其中$\eta = \beta_0 + \beta_1 \mathrm{X}_1 + \cdots \beta_\mathrm{p} \mathrm{X}_\mathrm{p} + \gamma_{12}\mathrm{X}_1\mathrm{X}_2 + \cdots + \gamma_{(\mathrm{p}-1)\mathrm{p}}\mathrm{X}_{\mathrm{p}-1}\mathrm{X}_\mathrm{p} + \varepsilon$，$\beta_1 = \beta_3 = \beta_5 = \beta_7 = \beta_9 = 2$，$\gamma_{13} = 1.5$，$\gamma_{17} = 1.7$，$\gamma_{57} = 1.9$，$\gamma_{79} = 2.1$，其余的系数均为，即$\eta = 2\mathrm{X}_1 + 2\mathrm{X}_3 + 2\mathrm{X}_5 + 2\mathrm{X}_7 + 2\mathrm{X}_9 + 1.5\mathrm{X}_1\mathrm{X}_3 + 1.7\mathrm{X}_1\mathrm{X}_7 + 1.9\mathrm{X}_5\mathrm{X}_7 + 2.1\mathrm{X}_7\mathrm{X}_9 + \varepsilon$。

表 1 数值模拟结果展示

方法	错判率
FS-Lasso	38.12%
FS-Group Lasso	38.12%
FS-TGDR	55%
iFORT	27%
iFORM	13%
RAMP	22.5%

可以看出，两阶段方法的表现普遍不如无缝路径的方法，其中 iFORM 的表现最优，其筛选结果为：X_1，X_3，X_5，X_7，X_9，X_1X_3，X_1X_7，X_3X_7，X_5X_7，X_7X_9。

接下来，在实际数据集上使用 iFORM 方法进行筛选。选取结肠癌数据集，其中 $n = 62,\ p = 2000$，Y为是否得结肠癌，X为基因表达量。最后使用混淆矩阵，ROC 曲线和平均错判率、AUC 值来对模型进行评估。用 iFORM 进行模型拟合，错判率为 18.09%，输出的 AUC 值为 0.8505。总体来看，模型表现较好。

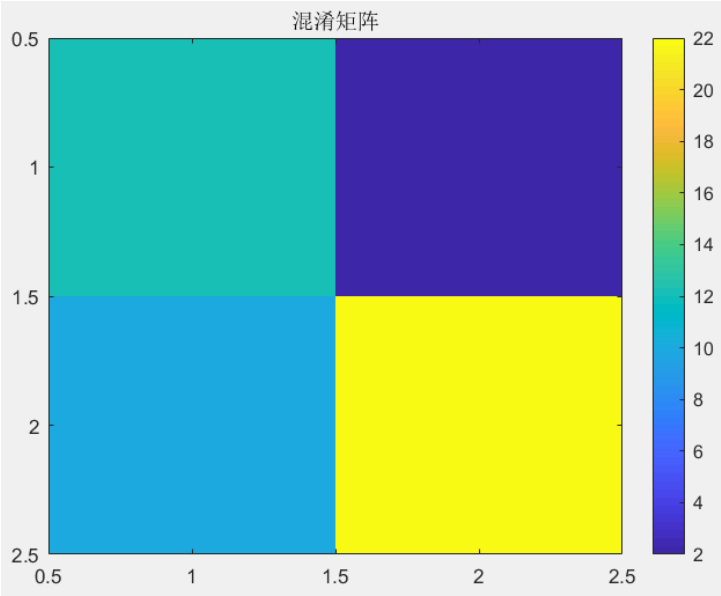

图 1 混淆矩阵

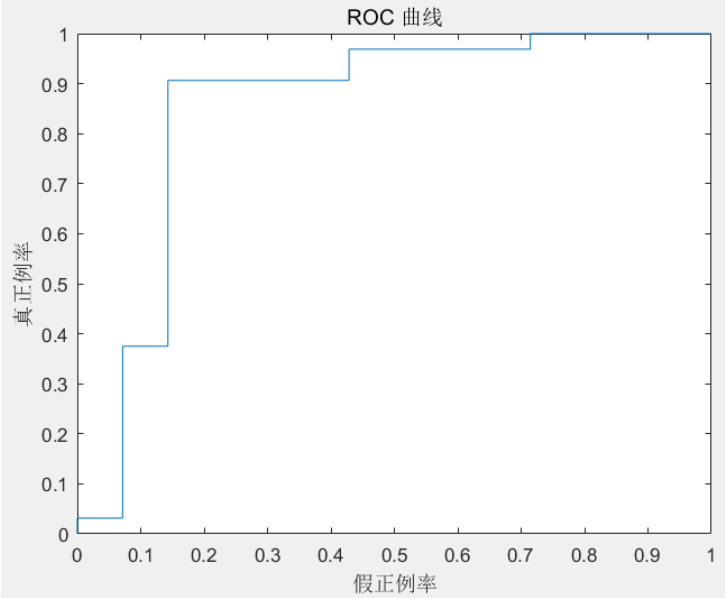

图 2 ROC 曲线

参考文献

[1] Nelder, J A. A Reformulation of Linear Models. Journal of the Royal Statistical Society, 1977, 140, 48–77.

[2] Peixoto, J L. A Property of Well-Formulated Polynomial Regression Models. The American Statistician, 1990, 44, 26–30.

[3] Hao N, Zhang H H. A Note on High-Dimensional Linear Regression With Interactions. The American Statistician, 2017, 71[4], 291-297.

[4] Bien J, Taylor J, Tibshirani R. A Lasso for Hierarchical Interactions. The Annals of Statistics: An Official Journal of the Institute of Mathematical Statistics, 2013, 41[3], 1111-1141.

[5] Lim M, Hastie T J. Learning Interactions via Hierarchical Group-Lasso Regularization. Journal of computational & graphical statistics, 2015, 24[3].

[6] Li H, Gui J. Gradient Directed Regularization for Sparse Gaussian Concentration Graphs, with Applications to Inference of Genetic Networks. Biostatistics, 2006, 7[2], 302-317.

[7] Hao N, Zhang H H. Interaction Screening for Ultra-high Dimensional Data. Journal of the American Statistical Association, 2014, 109, 1285–1301.

[8] Hao N, Feng Y, Zhang H H. Model Selection for High Dimensional Quadratic Regressions via Regularization. Journal of the American Statistician, 2014.