高维基因报告

史长浩　2022103697

2023 年 6 月 27 日

1　引言

在许多复杂疾病的病理、预后和治疗效应中，除了主要的基因（G）和环境（E）作用外，基因-环境（G-E）的交互效应也有着同等重要甚至更为重要的影响。粗略地说，当前关于 G-E 交互的研究可以分为两大类：边际分析和联合分析。前者每次分析一个或少量的 G 变量，后者则在一次分析中纳入所有的或大量的 G 变量。

边际分析处理的维度相对较低，容易实施计算，但在生物学上却是有问题的，因为复杂疾病的结局或表型可能同时与多个 G 和 E 效应及其交互效应有关，只纳入少数 G 变量可能会导致完全错误的识别。联合分析则避免了这一问题，然而，高维的 G 效应和 G-E 交互效应带来了计算上的挑战，此外，从生物学意义上，还需要考虑“主效应-交互效应”的层次结构，即若交互效应显著，则对应的主效应都显著（强层次结构），或至少有一个对应的主效应显著（弱层次结构），这为分析带来了额外的困难。本文仅关注联合分析方法。

事实上，对于 G-E 交互的联合分析已经有了一些经典的基于惩罚的方法，如 [Choi et al., 2010] 提出基于系数分解的方法可以在满足强层次结构限制的条件下选出重要的主效应和交互效应，[Bien et al., 2013] 提出基于 Lasso 的方法通过将一个非凸优化问题放松为凸问题实现了强层次结构限制下的变量选择，[Lim and Hastie, 2015] 进一步推广到利用 group-Lasso 实现层次结构限制条件下的变量选择。

近几年，随着数据收集能力的提升和统计技术的进步，这些方法又有了进一步发展，本文将在第二部分回顾近 5 年有代表性的一些文献，在第三部分呈现基于两种方法对结肠癌数据的分析报告。

2　文献综述

近 5 年来，涌现了大量关于“G-E 交互效应”的研究，对于联合分析方法，前文已经提到，“维度太高”是面临的主要挑战，而该问题从另一个角度观察即是“信息的缺乏”。事实上，最简单的引入更多信息的办法是提高样本量，但这在大多数时候是不切实际的，因此我们需要考虑利用 G 和 E 之外的，但与之紧密相关的变量。很多杰出的学者都从这一方向进

行突破，提出了各种精彩的引入“额外信息”的策略，例如：[Wang et al., 2019] 考虑了过往文献中已被证实的关于 G-E 交互效应的先验信息（prior information），[Wu et al., 2020] 考虑了 G 的结构信息（the structures of G measurements，如单核苷酸多态性的邻接结构和基因表达的网络结构信息），[Du et al., 2021] 考虑了多组学数据（Multi-Omics Data），[Xu et al., 2022] 考虑了多维分子数据（multidimensional molecular data），[Fang et al., 2023b] 考虑了病理影像学数据（pathological imaging）等。

与此同时，还有一个重要的研究方向是在不引入额外信息的条件下，提出新的方法将原有问题做得更好，或对过往方法做更深入的理论上的讨论，例如：[Wu et al., 2023] 提出了深度学习和惩罚相结合的方法，[Fang et al., 2023a] 在 cox 模型下为 G-E 交互联合分析建立了严格的渐近理论等。

下面将从这两个方向分别展开讨论。

2.1 引入额外信息

2.1.1 文献先验信息

随着实验和数据的长期积累，对于许多感兴趣的生物医学问题，已有的研究可以为 G-E 交互效应或主效应的识别提供有价值的信息。[Wang et al., 2019] 开发了一种基于准似然的方法去整合从现有文献中挖掘到的信息，并采用惩罚方法进行识别和选择，这一切是在考虑“主效应-交互效应”的层次结构限制和联合建模的框架下实现的。

考虑到先验信息并不总是“有信息的”，或者说，并不总是正确的，作者分两步去建立最终的估计量，以实现“在先验信息正确时，对估计结果有显著提升；在先验信息不准确时，不会对估计结果造成严重影响”的目标。第一步，考虑先验信息完全可信的场景，构造优化目标为：

$$L\left(\alpha_0, \mathbf{a}, \mathbf{b}; \mathbf{Y}\right) + P_p\left(\mathbf{b}; \xi, \kappa\right), \tag{1}$$

其中

$$P_p(\mathbf{b}; \xi, \kappa) = \sum_{j \notin S_c} \rho\left(||\mathbf{b}_j||; \sqrt{(q+1)}\kappa_1, \xi\right) + \sum_{(j,k) \notin S_{d-E}} \rho\left(|b_{jk}|; \kappa_2, \xi\right). \tag{2}$$

这是修正的 sgMCP 惩罚，它确保了文献中先验信息被自动识别，并且只与那些以前没有识别的效应进行选择。第二步，在先验信息和样本信息之间进行权衡，将式 (1) 的优化结果记为

$$\hat{Y}_i^{\text{prior}} = g^{-1}\left(\hat{a}_0^{\text{prior}} + \sum_{k=1}^{q} X_{ik}\hat{a}_k^{\text{prior}} + \sum_{j=1}^{p} \hat{b}_j^{\text{priorT}} W_{ij}\right). \tag{3}$$

引入调节参数 $0 \leq \tau \leq 1$ 描述先验信息和样本信息之间的权衡，得到最终的优化目标：

$$(1-\tau)L(\alpha_0, \boldsymbol{a}, \boldsymbol{b}; \boldsymbol{Y}) + \tau L\left(\alpha_0, \boldsymbol{a}, \boldsymbol{b}; \hat{\boldsymbol{Y}}^{\text{prior}}\right) + P(\boldsymbol{b}; \xi, \lambda). \tag{4}$$

数值模拟中，对先验信息的可信比例、G 变量的协方差结构、样本量、G 变量维度都进行了不同组合的探索，结果表明，当先验信息的可信比例较高时，本文方法的估计结果有显著提升；当先验信息的可信程度较差时，本文方法的估计也没有比候选方法更差，体现了估计的稳健性。在实证分析中，通过对皮肤黑色素瘤和多形性胶质母细胞瘤的癌症基因组图谱 (TCGA) 数据的分析，证明了所提方法的实用性，并得出了有生物学意义的结论。

2.1.2 G 的结构信息

在 G-E 交互效应的研究中，惩罚方法已经得到了充分地发展，但这些研究很少考虑 G 的结构信息，例如单核苷酸多态性的邻接结构和基因表达的网络结构。[Wu et al., 2020] 提出了结构化的 G-E 交互效应分析方法，其中这种结构是通过对 G 效应和交互效应的惩罚来调节的。

具体来说，作者构造了如下目标函数：

$$\begin{aligned} Q_n(\theta) = & \frac{1}{2n}\left\|V - Z\alpha - X\beta - \textstyle\sum_{k=1}^{q} W^{(k)}\left(\beta \odot \gamma_k\right)\right\|_2^2 \\ & + \sum_{j=1}^{p} \rho(|\beta_j|; \lambda_1, r) + \sum_{j=1}^{p}\sum_{k=1}^{q} \rho(|\gamma_{kj}|; \lambda_1, r) \\ & + \frac{1}{2}\lambda_2 \beta' J \beta + \frac{1}{2}\lambda_2 \sum_{k=1}^{q} \gamma_{k'} J \gamma_k, \end{aligned} \tag{5}$$

其中最值得关注的是矩阵 J，它反映了基于生物学知识对 G 结构的约束，如当 G 是 SNP 数据时，由于 SNP 紧密排列在染色体上，邻近的 SNP 往往具有相似的行为（值），如果被测量的 SNP 已经根据其物理位置进行了排序，则可以应用样条惩罚 $\sum_{j=2}^{p-1}[(\beta_{j+1}-\beta_j)-(\beta_j-\beta_{j-1})]^2$ 和 $\sum_{j=2}^{p-1}[(\gamma_{k(j+1)}-\gamma_{kj})-(\gamma_{kj}-\gamma_{k(j-1)})]^2$。因此，$J$ 的构造需要特定的领域知识，并且可能因数据类型而异。

理论上，作者证明了在高维设定下，该估计具有相合性。数值模拟中，作者对不同的样本容量、G 变量维度、G 变量的协方差结构、G 的结构信息（以 SNP 为例）等进行了不同组合的探索，对比了 MA、HierMCP、SMCP 三种惩罚方法，表明了本文方法的优越性。实证分析中，通过对具有 SNP 记录的 GENEVA 糖尿病数据和具有基因表达记录的 TCGA 黑色素瘤数据的模拟和分析，展示了本文方法的有效性。

2.1.3 多组学数据

大量研究表明，跨多平台分析组学数据（omics data）不仅具有生物学意义，而且可以提高识别和预测性能。[Du et al., 2021] 提出的整合模型可以通过稀疏降维有效地确定基因表达的重要调节因子，并通过容纳一个稀疏的双层结构将疾病结果与整合 G-E 交互研究中的多种效应联系起来。

主要关注如下模型：

$$Y = \sum_{k=1}^{q} \alpha_k E_k + \sum_{j=1}^{p_k} \left(\beta_j G_j + \sum_{k=1}^{q} \eta_{jk} G_j E_k \right) + \sum_{t=1}^{p_r} \gamma_t R_t + \epsilon, \tag{6}$$

与以往的不同之处是 $R_{n\times p_r} = (R_1, \cdots, R_{p_r})$ 表示 p_r 个调控因子（regulator），可能包括 DNA 甲基化（DNA methylation (DM)）和拷贝数变异（copy number alterations (CNA)）等，这是纳入模型的额外信息。作者开发了一个两阶段模型，在第一阶段，通过惩罚确定稀疏的调控关系，采用线性调控模型（LRM），以识别影响 GEs 集合的调控因子集合，以及 LRM 无法捕获的基因表达残差和调控因子残差。第二阶段，在 G-E 交互模型中，将 LRMs 和两类残差作为对癌症结果的直接效应，并进行惩罚，以识别重要的主效应和交互效应。

在数值模拟中，对样本量、G 变量的维度、G 协方差的结构、调控因子的维度、信号强度等做了不同组合的探索，并与 IGE、S-LASSO、J-LASSO、ColReg 等惩罚方法进行对比，展示了所提方法的优越性。在实证分析中，分别基于肺腺癌数据（LUAD）和肺鳞状细胞癌据（LUSC）进行分析，得出了具有生物学意义的结果。

2.1.4 多维分子数据

最近的大量研究表明，对多种类型的分子测量（遗传学、基因组学、表观遗传学等）进行联合分析不仅具有生物学意义，而且可以得到更好的估计和预测，[Xu et al., 2022] 进行了 M-E 交互效应分析，其中 M 代表多维分子测量，E 代表环境风险因子，以容纳多种类型的分子测量，并充分考虑它们之间的重叠和独立信息。

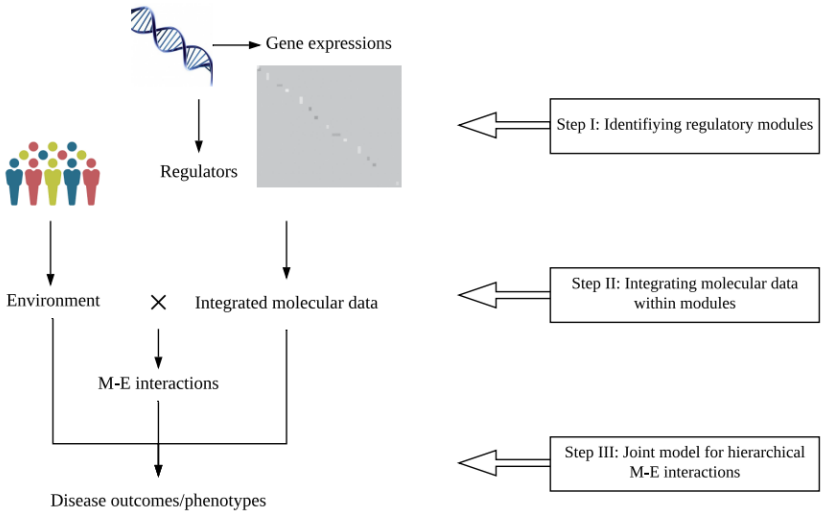

图 1: M-E 相互作用分析流程图。

图1展示了作者的分析思路：第一步通过惩罚回归来估计基因表达-调控因子关系，然后

依次进行双聚类以识别调控模块；第二步通过 PCA 对每个调控模块内部的信息进行整合；第三步考虑“主效应-交互效应”的层次结构限制进行优化。

在数值分析中，对于样本容量、M 变量的维度、M 变量的协方差结构、调控模式进行了不同组合的探索，结果表明所提方法在调控模式识别正确的条件下对估计效果有显著提升。在实证分析中，对肺腺癌和皮肤黑色素瘤的 TCGA 数据进行分析，得到了一些稳定的生物学发现，并实现了稳定的预测。

2.1.5 病理影像学数据

为了辅助癌症 G-E 交互效应的分析，[Fang et al., 2023b] 采取了不同于现有文献的策略，从病理影像学数据（pathological imaging）中提取信息，这些数据是活检的“副产品”，具有广泛的可用性和较低的成本，并且在最近的研究中被证明对于预测预后和其他癌症结果（或表型）具有积极意义。

作者考虑 cox 模型，对“基因 Z -环境 E”和“病理影像 X -环境 E”分别考虑两个 cox 模型，延续现有文献的方法，用线性模型对 $X \sim Z$ 进行建模

$$\mathbf{X}_{i,\cdot} = \mathbf{Z}_{i,\cdot}\mathbf{H}^* + \mathbf{U}_{i,\cdot}, \tag{7}$$

作者的直觉在于，从统计学上，式 (7) 中的 Z 与 X 具有同等地位，但在生物学上，Z 在我们心中的地位更高，因此，我们希望从 X 中“恢复”出那些有价值的 Z 的信息，于是认为 $\mathbf{Z}_{i,\cdot} \approx \mathbf{X}_{i,\cdot}\mathbf{F}^*$，其中

$$\mathbf{F}^* = \arg\min_{\mathbf{F}} E\big(\|\mathbf{Z}_{i,} - \mathbf{X}_{i,}\mathbf{F}\|_F^2\big). \tag{8}$$

然后将该式代入“基因 Z -环境 E”的 cox 模型中，最后将两个 cox 模型的损失函数相加，施加适当的惩罚（group-Lasso 等），可以在满足“主效应-交互效应”的层次结构约束下进行优化。

在数值分析中，对不同的样本量、G 变量维度、G 变量的协方差结构、病理影像信息维度等进行不同组合的探索，并与四种候选方法进行对比，所提方法可以在维持低 FP 的情况下取得高 TP，且具有低的 SSE 和高的 C-index，证实了本方法的优越性。在实证分析中，作者分析了肺腺癌的 TCGA 数据，感兴趣的结果是总生存期，G 变量为基因表达，在病理成像数据的辅助下，本文的 G-E 交互效应分析发现了不同的生物学结果，具有好的稳定性和预测性能。

2.2 不引入额外信息

2.2.1 深度学习

在包括生物医学和组学在内的众多领域中，深度学习以其特有的灵活性和良好的预测性能得到越来越多的认可，然而，在面向 G-E 交互分析的深度学习方面一直缺乏进展，[Wu

et al., 2023] 填补了这一重要的知识空白，发展了一种新的基于深度神经网络结合惩罚的分析方法。

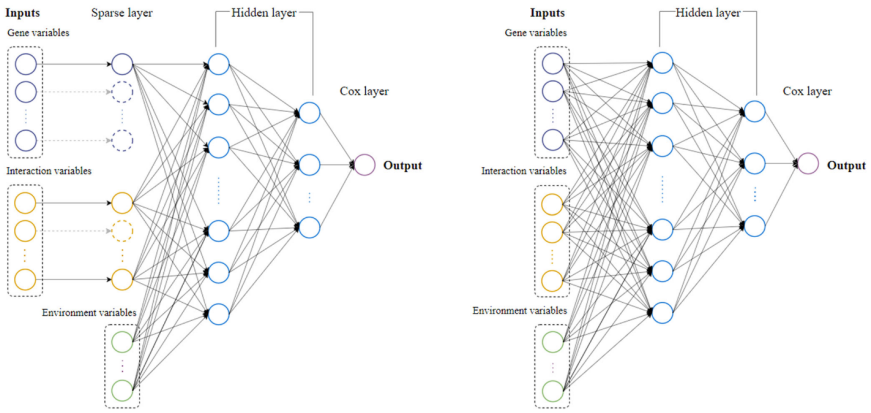

图 2: 左：本文提出的神经网络架构；右：作为对比，全连接神经网络架构。

作者所提神经网络的整体架构如图 (2) 左所示，与不考虑变量选择的全连接神经网络相比，本文的神经网络纳入了特殊的“稀疏层”，考虑损失函数：

$$\begin{aligned} L(\theta) = l(\theta) + \sum_{j=1}^{p} \rho(\|\boldsymbol{b}_j\|; \lambda_1, s) \\ + \sum_{j=1}^{p} \sum_{k=1}^{q} \rho(|\beta_{kj}|; \lambda_2, s) + \lambda \left(\sum_{k=1}^{K} \|\omega_k\|_F^2 \right), \end{aligned} \tag{9}$$

其中，前两项惩罚利用 MCP 实现了稀疏组惩罚，且保证了在变量选择时自动满足“主效应-交互效应”的层次结构限制，最后一项惩罚则基于岭惩罚对网络的复杂度进行控制，避免隐藏层参数的膨胀。

在数值模拟中，对样本容量、G 变量维数、G 变量协方差结构的不同组合进行了探索，并与其他网络结构和惩罚函数的组合（DeepGE-Lasso，FNN，MA，CoxMCP，CoxLasso）进行了对比，结果表明在不同设置下，本文方法都有更高的 C-index，对于主效应和交互效应的识别上，相对于其他方法有显著低的 FP 和有竞争力的 TP。在对肺腺癌和皮肤黑色素瘤总体生存数据的分析中，本文方法识别出 53 个主 G 效应和 41 个交互效应，其中大部分都得到了已有文献的支持，进一步印证了本文方法的实用性。

2.2.2 Cox 渐近理论

[Fang et al., 2023a] 考虑 Cox 模型下的联合 G-E 交互效应分析，利用“稀疏组惩罚”选择重要的主效应和交互效应，且考虑“主效应-交互效应”层次结构，并在高维情形下建立了参数估计的相合性和渐近理论，并发展了一种有效的计算算法。

本文的主要结果是在 7 个正则条件下推导出了参数的相合性（包括选择一致性）：

$$(a) \quad \left\|\hat{\phi}_1 - \phi_1^*\right\|_2 = O_p\left(\sqrt{s/n}\right), \quad (b) \quad \hat{\phi}_2 = \mathbf{0}, \tag{10}$$

以及渐近正态性：

$$(c) \quad \sqrt{n}\nu_n^\top \Sigma(\phi_1^*)^{1/2}\left(\hat{\phi}_1 - \phi_1^*\right) \to_d \mathcal{N}(0,1). \tag{11}$$

关于正则条件，前 5 个是通常带惩罚的 Cox 模型下都会假定的条件；条件 6 是最小信号假定，由于“稀疏组惩罚”的存在，不必对重要主效应的最小信号施加限制，而只需要假设重要组和交互效应的最小信号的条件；条件 7 是关于折叠凹函数的惩罚，在 SCAD 的文献中出现过。

数值模拟中，对样本量、G 变量维度、交互效应真实数量、G 变量协方差结构等不同组合进行了探索，对比了 MCP 惩罚和边缘方法等，结果表明，在不同设置下，在识别交互效应和主效应方面，本文方法总能在保持更低 FP 的条件下，实现比边缘方法更高的 TP，相对于其他惩罚项没有特别明显的优势。在实证分析中，对胃腺癌的 TCGA 数据做分析，本文方法识别了 11 个主要的 G 效应和 12 个交互效应，这些结果都能在现有文献中找到相应的证据，进一步证实了所提方法的实用性。

3 数据分析

考虑从http://genomics-pubs.princeton.edu/oncology/affydata/index.html 获取的结肠癌数据，样本容量为 62，基因个数为 2000，为计算可行性计，仅考虑前 50 个基因，并假设第 46 到第 50 个基因为环境因素。首先将基因表达数据做对数变换，然后做标准化，使得每个基因表达对应样本的均值为 0，方差为 1；然后将指示组织是否癌变的标签 0-1 化。接下来，基于 [Lim and Hastie, 2015] 的 hierarchical group-lasso regularization 方法和 [Wang et al., 2019] 的考虑文献先验的方法进行分析。代码及数据归入附件中。

3.1 方法一：hierarchical group-lasso regularization

[Lim and Hastie, 2015] 通过对系数施加 group-Lasso 惩罚，自动实现了“主效应-交互效应”的层次结构约束，该方法只含有一个超参数，通过 10 折交叉验证选出的最佳超参数为 0.0093，经变量选择出的 5 个 G 主效应、2 个 E 主效应和 2 个 G-E 交互作用及其对应系数如表 (3.1) 所示，其中只有第 14 个 G 主效应和第 4 个 E 主效应是负的，其余效应均为正。

表 1: 方法一系数估计结果。

Z_{12}	Z_{14}	Z_{26}	Z_{27}	Z_{43}	Z_{47}	Z_{49}	$Z_{12} \times Z_{47}$	$Z_{27} \times Z_{49}$
0.0324	-0.9027	0.1632	0.0030	0.8696	0.0968	-0.0545	0.1366	0.0098

为了评价本方法的预测性能，绘制 ROC 曲线如图 (3) 所示，ROC 曲线下面积为 0.960，该方法在预测组织是否癌变方面有着良好的效果。

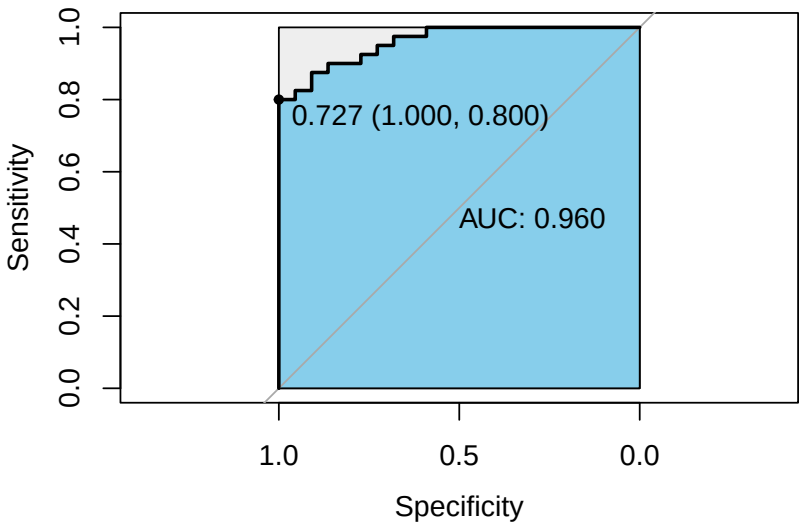

图 3: 方法一 ROC 曲线。

3.2 方法二：incorporating prior information

[Wang et al., 2019] 提出的方法可以利用已有文献的先验信息，其本质上是利用那些一定程度上已被证实的具有 G 主效应和 G-E 交互效应的信息，我们没有现成的文献可供参考，但是我们有方法一给出的结果可以利用，这里我们将方法一中的结果作为方法二的先验，即将表 (3.1) 指示的潜在可能的 G 主效应和 G-E 交互效应考虑进来（这里不考虑 E 主效应是因为方法二专注于识别 G 主效应以及与这些基因相关的交互效应，总假设环境效应是显著的）。

本方法存在两个超参数，做 5 折交叉验证后选择出最佳的参数组合为 $(2.0, 1.6)$，代入最优超参数进行变量选择，得到变量选择结果如图 (4) 所示，得到 5 个显著的 G 主效应和 2 个显著的 G-E 交互效应，其中 G27-E4 交互效应与方法一一致，但 G27-E1 的交互效应与方法一指示的不同，对照表 (3.1)。此外识别出的 5 个 G 主效应分别为 G12,G14,G26,G27,G43，这与方法一是完全相同的。从这一结果中也不难看出方法二在先验信息与样本信息之间所做的权衡。

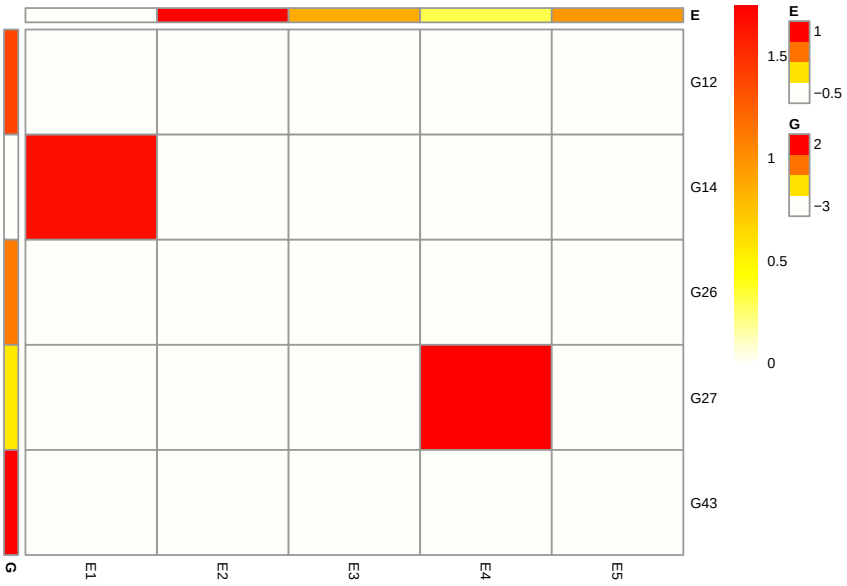

图 4: 方法二变量选择结果（G 主效应和 G-E 交互效应）。

为了进一步观察方法二的预测效果，并于方法一进行对比，我们继续绘制反映其分类性能的 ROC 曲线，如图 (5) 所示，ROC 曲线下面积为 0.975，相对于方法一的 0.960 有一定提升，该方法呈现出了预期中的更好的结果。

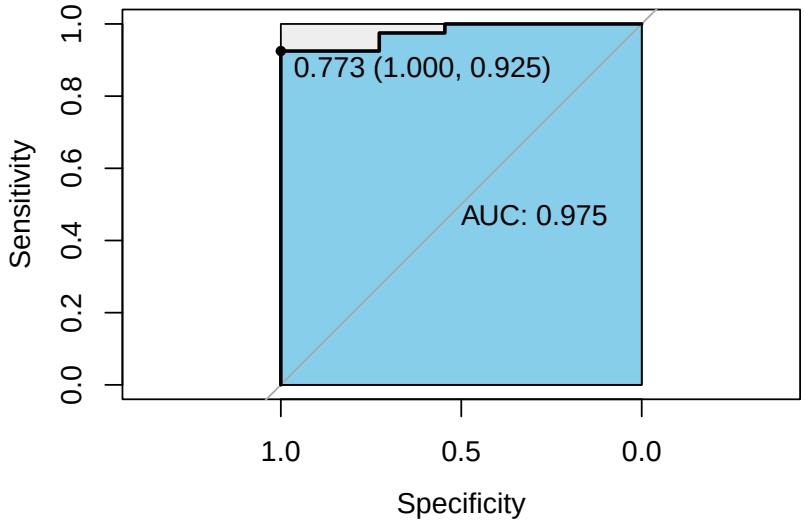

图 5: 方法二 ROC 曲线。

参考文献

J. Bien, J. Taylor, and R. Tibshirani. A lasso for hierarchical interactions. *Annals of statistics*, 41(3):1111, 2013.

N. H. Choi, W. Li, and J. Zhu. Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105(489):354–364, 2010.

Y. Du, K. Fan, X. Lu, and C. Wu. Integrating multi–omics data for gene-environment interactions. *BioTech*, 10(1):3, 2021.

K. Fang, J. Li, Y. Xu, S. Ma, and Q. Zhang. Gene–environment interaction analysis under the cox model. *Annals of the Institute of Statistical Mathematics*, pages 1–18, 2023a.

K. Fang, J. Li, Q. Zhang, Y. Xu, and S. Ma. Pathological imaging-assisted cancer gene-environment interaction analysis. *Biometrics*, 2023b.

M. Lim and T. Hastie. Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654, 2015.

X. Wang, Y. Xu, and S. Ma. Identifying gene-environment interactions incorporating prior information. *Statistics in medicine*, 38(9):1620–1633, 2019.

M. Wu, Q. Zhang, and S. Ma. Structured gene-environment interaction analysis. *Biometrics*, 76(1):23–35, 2020.

S. Wu, Y. Xu, Q. Zhang, and S. Ma. Gene–environment interaction analysis via deep learning. *Genetic Epidemiology*, 47(3):261–286, 2023.

Y. Xu, M. Wu, and S. Ma. Multidimensional molecular measurements–environment interaction analysis for disease outcomes. *Biometrics*, 78(4):1542–1554, 2022.