arXiv:2304.08605v1 [stat.ME] 17 Apr 2023

基于 Gini距离相关的超高维分类分组特征筛选

Yongli Sang[a] 和 Xin Dang[b*]

[a]数学系，路易斯安那大学拉斐特分校，美国路易斯安那州拉斐特，邮编 70504[b]数学系，密西西比大学，美国密西西比州大学城，邮编 38677

2023年4月19日

摘要

近期提出了基尼距离相关（GDC），用于衡量分类变量 Y 与数值随机向量 $\boldsymbol{X}$ 之间的依赖关系。它可以相互刻画 $\boldsymbol{X}$ 与 Y 之间的独立性。本文利用 GDC 为响应变量为分类变量的超高维判别分析建立特征筛选方法。该方法既可筛选单个特征，也可筛选分组特征。所提出的程序具有若干优良性质：它是无模型的，无需设定模型；它具有确定性独立筛选性质和排序一致性性质。所提出的筛选方法还可处理响应变量类别数发散的情形。我们开展了若干蒙特卡罗模拟研究，以考察所提出筛选程序的有限样本性能，并通过两个真实数据集展示真实数据分析。

关键词：判别分析；Gini距离相关；分组筛选；特征筛选；排序一致性性质；确定性独立筛选；超高维。

*2020*年数学主题分类：62H30

1 引言

过去十年中，独立性特征筛选方法被认为适用于超高维数据。确定性独立筛选（SIS） [6] 根据单个变量与响应变量的皮尔逊相关来衡量其重要性并进行排序。在线性模型设定下，已有研究证明该程序具有确定性筛选性质，即所有重要变量都被选中的概率趋近于 1。自这篇开创性论文以来，人们提出了各种筛选方法 [8, 7, 9, 35]。Hall 和 Miller [11] 利用广义相关进行筛选。广义相关定义为特征变量的变换与响应变量之间的最大相关。如果在

*联系人：Xin Dang；电子邮箱：xdang@olemiss.edu

特征变量的变换以及响应变量的变换中进行最大化，就能得到最大相关，并将其用于筛选，见 [14]。Kendall 的 τ 相关 [15, 16] 用于 [18] ，以增强稳健性。距离相关被应用于 [19]，以实现无模型性质；与之密切相关的鞅差相关则用于 [28]。上述所有程序都针对数值型响应变量。

当响应变量为分类变量时，已有若干筛选方法被提出 [3, 21, 2, 17]。Cui、Li和Zhong [3] 利用边际分布函数与条件分布函数之差构造特征筛选指标。在二分类中，该指标可简化为 Cramér-von Mises 检验统计量，其基于秩的表示已在 [4]中得到研究。Cheng 等人 [2]提出了一种基于预测变量秩的条件期望的基于秩的筛选程序。He、Ma和Xu [12] 改进了 [3] 中的程序，并提出了 Anderson-Darling 型指标，该指标对尾部观测值赋予更大的权重。在 [21], 中，采用熵来衡量各预测变量与分类响应之间的依赖关系。所有这些特征筛选程序都基于单个预测变量，无法纳入分组变量这类结构化预测变量，因为这类变量中的所有协变量都高度相关。在许多超高维数据情形下，预测变量天然地分成若干组。

关于超高维数据的分组特征筛选研究成果较为有限。Niu $et\ al$[22] 通过对每个分组预测变量的边际估计量的绝对值进行排序，提出了线性模型中的一种边际分组确定性筛选方法。Qiu 和 Ahn [25]提出了三种新的筛选方法：gSIS（分组确定性独立筛选）、gHOLP（分组高维普通最小二乘投影器）和 gAR2（分组调整 R 平方筛选）。Song 和 Xie [29] 在已知分组信息的情况下，通过扩展原始确定性独立筛选程序，基于 F-检验统计量提出了一种分组筛选程序。上述方法主要集中于线性模型。对于分类问题，近期提出了三种无模型分组筛选方法。在 [13] 和 [32], 中，使用熵和互信息来衡量每个分组协变量的重要性。在 [33], 中，则使用基尼不纯度这一类别指标作为筛选度量。然而，这些程序都是针对分类协变量设计的。当协变量为数值型时，必须依赖分箱将数值协变量转换为类别，从而损失相当一部分信息。此外，如何进行分箱也是应用这些方法时面临的一大问题。

本文提出了一种用于超高维数据判别分析的新型无模型分组筛选方法。尽管该方法能够处理分类协变量，但它尤其适用于筛选数值协变量，这是分类中最常见的情形。该新程序基于协变量组与响应变量之间的基尼距离相关（GDC）。GDC 是一种近期提出的分类变量 Y 与 p 维随机向量 $\boldsymbol{X}$ [5] 之间的依赖性度量。已有研究表明， $\boldsymbol{X}$ 与Y 之间的 GDC 等于零，当且仅当 $\boldsymbol{X}$ 与 Y 相互独立。这种独立性意味着，在分类变量Y的不同类别下，预测变量 $\boldsymbol{X}$ 具有相同的统计分布。这使我们能够为超高维数据构建一种新的特征筛选指标。所提出的筛选程序根据组协变量与响应变量之间的 GDC，对组预测变量的重要性进行排序。在 GDC 中，协变量的维数可以任意，因此所提出的程序既适用于个体特征筛选，也适用于分组特征筛选。

所提出的筛选程序不仅具有确定性筛选性质和排序一致性性质，还具备以下优点。

1. 该方法是无模型的，也就是说，它不需要对响应变量或预测变量进行模型设定。2. 这一新程序既可用于筛选单个预测变量，也可用于筛选分组预测变量。3. 此外，该方法还允许响应变量的类别数发散。

本文其余部分安排如下。在第2节中，我们首先介绍 GDC，然后建立确定性独立筛选程序。本文主要介绍分组筛选，并将个体筛选视为特例。在第3节中，我们通过模拟研究评估所提出筛选程序的性能。第4 节展示了两个数据集的真实数据分析，以便将所提出的程序与现有方法进行比较。我们在第5节总结全文，并讨论该方法与一种现有方法的联系。所有技术性证明均见附录。

2 分组特征筛选

令 Y为取值为 L_1，...， L_K的分类响应变量， $K \geq 2$且 $\boldsymbol{X}$ =（X_1，...， X_p）为预测变量向量。假设根据某些领域知识，协变量 $\boldsymbol{X}$ 被划分为 r 个预测变量组（$\boldsymbol{X}_{(1)}$，$\boldsymbol{X}_{(2)}$，...，$\boldsymbol{X}_{(r)}$），其中第 l^{th} 组为 $\boldsymbol{X}_{(l)}$ =（X_{l1}，$...X_{lq_l}$），其维数 q_l 满足 $q_1 + q_2 + ... + q_r = p$。我们假设 q_l、 $l = 1$，...， r，均为有限值。在超高维情形下，组数 r （因而 p）远大于样本量 n。因此，自然可以假设只有少数预测变量组与 Y相关。筛选程序的目标是选取一个规模适中的约化模型，使其几乎完整地包含相关预测变量组。在介绍筛选程序的细节之前，我们先回顾 $\boldsymbol{X}_{(l)}$ 与 Y之间的 Gini距离相关。

2.1 Gini距离相关

设 Y的分布为 $P(Y = L_k) = p_k > 0$，其中 $k = 1$，2，...， K。假设$\boldsymbol{X}_{(l)}$ 的分布为 F_l ，且给定 $Y = L_k$ 时 $\boldsymbol{X}_{(l)}$ 的条件分布为 $F_{l|k}$。分别将 $F_{l|k}$ 和 F_l的特征函数记为 $\psi_{l|k}$ 和 ψ_l 。则$\boldsymbol{X}_{(l)}$ 与 Y 之间的Gini距离协方差定义为

$$\mathrm{gCov}(\boldsymbol{X}_{(l)}, Y) = c(q_l) \sum_{k=1}^{K} p_k \int_{\mathbb{R}^{q_l}} \frac{|\psi_{l|k}(\boldsymbol{t}) - \psi_l(\boldsymbol{t})|^2}{\|\boldsymbol{t}\|^{q_l+1}} d\boldsymbol{t}, \tag{1}$$

其中 $c(\cdot)$是依赖于维度 q_l的已知常数， $\|\boldsymbol{t}\|$ 是 $\boldsymbol{t} \in \mathbb{R}^{q_l}$ 的欧几里得范数，$|\psi|^2 = \psi\bar{\psi}$ 是复函数 ψ。分类Gini协方差基于边际特征函数与条件特征函数之间的加权（平方）距离来衡量依赖关系。我们看到，$\mathrm{gCov}(\boldsymbol{X}_{(l)}, Y) \geq 0$ 与$\mathrm{gCov}(\boldsymbol{X}_{(l)}, Y) = 0$ 相互蕴含 $\boldsymbol{X}_{(l)}$ 与 Y独立。

当 $q_l = 1$时， Y与单变量 $X_{(l)}$ 之间在（1）中定义的基尼距离协方差可简化为

$$\mathrm{gCov}(X_{(l)}, Y) = \sum_{k=1}^{K} p_k \int_{\mathbb{R}} \left(F_{l|k}(x) - F_l(x)\right)^2 dx.$$

它是边际分布与条件分布之间的加权平方距离，与 [3] 中使用的均值方差（MV）指标密切相关。也就是说，

$$\mathrm{MV}(X_{(l)},Y)=\sum_{k=1}^{K}p_k\int_{\mathbb{R}}\big(F_{l|k}(x)-F_l(x)\big)^2dF_l(x),$$

它是边际分布与条件分布之间的加权 Cramér-von Mises 距离。MV 仅适用于一维情形，且没有直接的多元推广，而吉尼协方差则适用于单变量和多变量情形。

与 Y 在（1）中的吉尼协方差可表示为多元基尼均值差（GMD）[5]。即：

$$\mathrm{gCov}(\boldsymbol{X}_{(l)},Y)=\Delta_l-\sum_{k=1}^{K}p_k\Delta_{l|k},\tag{2}$$

其中，$\Delta_l=\mathbb{E}\|\boldsymbol{X}_{(l)1}-\boldsymbol{X}_{(l)2}\|$ 是 F_l 的 GMD，$\Delta_{l|k}=\mathbb{E}\|\boldsymbol{X}_{(l|k)1}-\boldsymbol{X}_{(l|k)2}\|$ 表示 $F_{l|k}$的 *GMD*，$k=1,.....,K$，其中（$\boldsymbol{X}_{(l1)}$，$\boldsymbol{X}_{(l2)}$）和（$\boldsymbol{X}_{(l|k)1}$，$\boldsymbol{X}_{(l|k)2}$）是相互独立的成对变量，分别独立于 F_l 和 $F_{l|k}$。与（1）相比，这种表示（2）提供了一种更直接的吉尼协方差估计方法。此外，相应的 Gini距离相关（GDC）可简单定义为：

$$\rho_l=\frac{\Delta_l-\sum_{k=1}^{K}p_k\Delta_{l|k}}{\Delta_l}.\tag{3}$$

Gini距离相关（3）可以很好地解释为组间变异与总变异之比。其取值范围为 $[0,1]$。当且仅当数值变量 $\boldsymbol{X}_{(l)}$ 与分类变量 Y 相互独立时，该值为零。这些显著的性质促使我们在分类中使用 GDC 对 $\boldsymbol{X}_{(l)}$ 的重要性进行排序。

假设$\mathcal{D}=\{(\boldsymbol{X}_1,Y_1),(\boldsymbol{X}_2,Y_2),....,(\boldsymbol{X}_n,Y_n)\}$是来自 $\boldsymbol{X}$与 Y联合分布的随机样本。令 $\mathcal{D}_l=\{(\boldsymbol{X}_{(l)i},Y_i)\}_{i=1}^{n}$为仅包含 l^{th}组预测变量的样本数据。记作 $\mathcal{D}_l=\mathcal{D}_{l|1}\cup\mathcal{D}_{l|2}...\cup\mathcal{D}_{l|K}$，其中 $\mathcal{D}_{l|k}=\{\boldsymbol{X}_{(l|k)1},\boldsymbol{X}_{(l|k)2},...,\boldsymbol{X}_{(l|k)n_k}\}$是 $\mathcal{D}_l$ 的子样本，其中 $Y=L_k$ 且 n_k 是 k^{th} 类别中的样本点数。则式（3）中的 Gini距离相关可估计为

$$\hat{\rho}_l=\frac{\binom{n}{2}^{-1}\sum_{1\le i<j\le n}\|\boldsymbol{X}_{(l)i}-\boldsymbol{X}_{(l)j}\|-\sum_{k=1}^{K}\hat{p}_k\binom{n_k}{2}^{-1}\sum_{1\le i<j\le n_k}\|\boldsymbol{X}_{(l|k)i}-\boldsymbol{X}_{(l|k)j}\|}{\binom{n}{2}^{-1}\sum_{1\le i<j\le n}\|\boldsymbol{X}_{(l)i}-\boldsymbol{X}_{(l)j}\|},\tag{4}$$

其中 $\hat{}p_k=\frac{n_k}{n}=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(Y=L_k)$ $k=1,...,K$，$\mathbb{1}(\cdot)$ 为指示函数。

计算4的过程很简单，通常计算复杂度为 O（n^2）。然而，对于维度为 $X_{(l)}$ 的单变量 $q_l=1$，计算4的简单快速算法只需O（$n\log n$）。因此，即使 n很大，边际特征筛选仍颇具吸引力。Dang *et al.* [5]以及 Sang 和 Dang [26] 分别探究了 GDC 在低维和高维情形下的性质。与常用的距离相关相比，GDC 更便于进行统计推断，并且在处理不平衡数据时更稳健。我们预期，GDC 的这些优良性质也将由面向超高维数据的特征筛选方法继承。

2.2 基于 GDC 的分组特征筛选

本节提出一种基于 GDC 的独立性筛选程序。记 $G(Y = L_k|\boldsymbol{X})$ 为给定 $\boldsymbol{X}$ 时 Y 的条件分布。无需指定分类模型，我们分别将活跃组和非活跃组的指标集合定义为

$$\mathcal{A} = \{l : G(Y = L_k|\boldsymbol{X}) \text{ 对于某个 } k\} \text{ 在函数上依赖于 } X_l\,,\ \mathcal{I} = \{l : G(Y = L_k|\boldsymbol{X}) \text{ 对于所有 } k\}. \text{ 均不在函数上依赖于 } X_l$$

定义 $\boldsymbol{X}_{\mathcal{A}} = \{\boldsymbol{X}_{(l)} : l \in \mathcal{A}\}$ 和 $\boldsymbol{X}_{\mathcal{I}} = \{\boldsymbol{X}_{(l)} : l \in \mathcal{I}\}$，并将 $\boldsymbol{X}_{\mathcal{A}}$ 称为活跃预测变量向量，其补集 $\boldsymbol{X}_{\mathcal{I}}$ 称为非活跃预测变量向量。所有活跃预测变量的索引子集 $\mathcal{A}$，或等价地，所有非活跃预测变量的索引子集 $\mathcal{I}$，是我们主要关注的目标。$\mathcal{A}$ 和 $\mathcal{I}$ 的定义意味着，在给定 $\boldsymbol{X}_{\mathcal{A}}$ 的条件下，$\boldsymbol{X}_{\mathcal{I}}$ 与 Y 相互独立。为识别集合 $\mathcal{A}$，我们使用 $\boldsymbol{X}_{(l)}$ 与 Y 之间的基尼相关。

我们通过对 $\boldsymbol{X}_{(l)}$ 的 GDC 排序，来评估分组预测变量 $\boldsymbol{X}_{(l)}$ 的重要性，其中 Y，ρ_l，$l = 1, \ldots r$。注意，$\rho_l = 0$ 当且仅当 $\boldsymbol{X}_{(l)}$ 与 Y 相互独立时成立。ρ_l 是区分活跃组与非活跃组的有效指标。例如，在部分正交条件[8] 下，其中 $\{\boldsymbol{X}_{(l)}:\ l \in \mathcal{A}\}$ 与 $\{\boldsymbol{X}_{(l)}:\ l \in \mathcal{I}\}$ 相互独立，则当 $\rho_l > 0$，且 $l \in \mathcal{A}$ 对 $\rho_l = 0$ 成立时，$l \in \mathcal{I}$。此外，GDC 不需要模型假设，并允许预测变量具有任意维数。因此，它可用于筛选单个协变量或分组协变量。

ˆ 我们选取一组 ˆρ_j 较大的重要组。也就是说，我们用 $\mathcal{A}$ 估计 $\mathcal{A}$，即

$$\hat{\mathcal{A}} = \{l : \hat{\rho}_l \geq cn^{-\kappa},\ \text{for } 1 \leq l \leq r\},$$

其中，$c > 0$ 和 $\kappa \in [0, 1/2)$ 是条件 **C3** 中定义的预设阈值。在实践中，对于给定的大小 $d < n$，我们可以选取一个约化模型:

$$\hat{\mathcal{A}}^* = \{l : \hat{\rho}_l \text{ is among the top } d \text{ largest of all}\}.$$

在对预测变量的活跃组的最小基尼相关作出一些假设（例如，$\min_{l\in\mathcal{A}} \rho_l \geq 2cn^{-\kappa}$）后，我们希望证明确定性筛选性质；这可以通过推导统一的指数收敛界来建立。

为便于技术性证明，我们提出以下条件，尽管它们可能并非最弱条件。

C1 存在两个正常数 c_1 和 c_2，使得

$$c_1/K_n \leq \min_{1\leq k\leq K_n} p_k \leq \max_{1\leq k\leq K_n} p_k \leq c_2/K_n.$$

假设对于 $\tau \geq 0$，$K_n = O(n^\tau)$。

C2 $\boldsymbol{X}$ 在维度 p 和类别上均满足次指数尾概率。也就是说，存在一个正的常数 s_0，使得对于所有 $0 \leq s \leq 2s_0$,

$$\sup_r \max_{1\leq l\leq r} \sup_K \max_{1\leq k\leq K} \mathbb{E}\{\exp(s\|\boldsymbol{X}_{(l|k)}\|)\} < \infty.$$

C3 活跃预测变量组的最小基尼相关满足

$$\min_{l\in\mathcal{A}} \rho_l \geq 2cn^{-\kappa}, \quad \text{for some constants } c > 0 \text{ and } 0 \leq \kappa < 1/2.$$

条件 **C1** 要求分类变量 Y 各类别所占的比例不能过小或过大，即不存在某一类别占主导地位的情况。请注意，类别数 $K_n = O(n^\tau)$ 可以趋于无穷大。若 **C2**满足 $\boldsymbol{X}$ 一致有界或服从多元正态分布，则次指数条件可直接成立。它比 [19]中使用的次高斯条件更弱，并允许具有更重尾的分布 [31]。假设 **C3** 是独立性筛选程序中的典型条件，要求活跃预测变量组的 GDC 不能过小。其阶为 $n^{-\kappa}$，这允许最小真实信号随着样本量 n 趋于无穷大而趋于零。

我们在以下定理中证明分组筛选程序具有确定性筛选性质。

定理 2.1 在条件 ***C1-C2*** 下，对于任意 $0 < \beta < 1/2 - \kappa$ 和 $0 \le \tau < 1 - 2\kappa$，存在正数常量 b_1、b_2、b_3，它们依赖于 c、c_1 和 c_2，使得

$$P(\max_{1\le l\le r} |\hat{\rho}_l - \rho_l| \ge cn^{-\kappa}) \le$$
$$O\left\{r\left[\exp(-b_1 n^{1-2(\kappa+\beta)} + \tau\log n) + \exp(-b_2 n^{\beta} + (1+\tau)\log n) + \exp(-b_3 n^{1-2\kappa-\tau} + \tau\log n)\right]\right\}. \tag{5}$$

在条件***C1-C3***下，我们有

$$P(\mathcal{A} \subseteq \hat{\mathcal{A}}) \ge$$
$$1 - O\left(d_n\left[\exp(-b_1 n^{1-2(\kappa+\beta)} + \tau\log n) + \exp(-b_2 n^{\beta} + (1+\tau)\log n) + \exp(-b_3 n^{1-2\kappa-\tau} + \tau\log n)\right]\right), \tag{6}$$

其中 $d_n = |\mathcal{A}|$ 是 $\mathcal{A}$ 的基数。

这表明，我们基于 GDC 的筛选程序可应用于非多项式（NP）维数情形。我们的意思是，$\log p \ge \log r = O(n^\omega)$，其中 $\omega < \min\{1-2(\kappa+\beta)\ \beta,\ 1-2\kappa-\tau\}$，其中 $0 \le \kappa < 1/2$、$0 \le \beta < 1/2 - \kappa$ 和 $0 \le \tau < 1 - 2\kappa$，这取决于最小真实信号强度和响应类别的数量。

若 K_n 在 $\tau = 0$下固定，则定理2.1 的结果可改进为

$$P(\max_{1\le l\le r} |\hat{\rho}_l - \rho_l| \ge cn^{-\kappa}) \le O\left\{r\left[\exp(-b_1 n^{1-2(\kappa+\beta)}) + \exp(-b_2 n^{\beta} + \log n)\right]\right\}. \tag{7}$$

因此，我们甚至可以处理更大的 NP 维数，其中 $\log p \ge \log r = O(n^\omega)$，其中 $\omega <$ 最小$\{1-2(\kappa+\beta)\ \beta\}$，且 $0 \le \kappa < 1/2$ 并且 $0 \le \beta < 1/2 - \kappa$。平衡式（7）中的两项后，我们选择最优阶 $\beta = (1-2\kappa)/3$，于是对某个常数 $b > 0$,

$$P(\max_{1\le l\le r} |\hat{\rho}_l - \rho_l| \ge cn^{-\kappa}) \le O\left\{r\left[\exp(-bn^{(1-2\kappa)/3} + \log n)\right]\right\}.$$

独立性筛选的另一项性质是 Zhu *et al.* [35]所考察的排序一致性性质。为研究排序一致性性质，我们还需作出以下条件假设:

C4 对于某个正常数 c_3,

$$\liminf_{r\to\infty}\{\min_{l\in\mathcal{A}}\rho_l-\max_{l\in\mathcal{I}}\rho_l\}\geq c_3.$$

条件C4要求根据总体Gini距离相关将活跃组与非活跃组分离。显然，C4弱于部分正交条件。这是因为，如果 $\{\boldsymbol{X}_{(l)}:\ l\in\mathcal{A}\}$ 与 $\{\boldsymbol{X}_{(l)}:\ l\in\mathcal{I}\}$相互独立，那么 $\rho_l>0$ 如果 $l\in\mathcal{A}$ 且 $\rho_l=0$ 对于 $l\in\mathcal{I}$，从而C4成立。

定理 2。2 *[排序一致性性质]* 若条件 ***C1-C4*** 对于 $K_n\log_{n^\alpha}{}^{n}=o(1)$ 和$K_n\log_{n^\alpha}{}^{p}=o(1)$成立，且对于 $\alpha=\min\{1-2\beta+\tau,\beta+\tau,1\}$，则有

$$\liminf_{n\to\infty}\{\min_{l\in\mathcal{A}}\hat{\rho}_l-\max_{l\in\mathcal{I}}\hat{\rho}_l\}\geq 0,a.s.$$

定理 2.2 的结果强于确定性筛选性质。该结果表明，以高概率来看，活跃预测变量的样本 GDC 始终排在非活跃预测变量之后。若采用理想的阈值，只要活跃组和非活跃组在总体层面上可分，所提出的筛选程序便能够根据样本数据识别活跃预测变量组。

备注 2.1 当 $r=p$ 满足 $q_l=1$ 对所有 $l=1,...,r$ 时，该筛选程序属于边际筛选。我们根据每个单个协变量与响应变量之间的边际 *GDC* 对其进行排序。定理 *2.1* 和定理 *2.2* 的结果对边际筛选同样成立。

注 2.2 *Zhang* 等人（*2019*）*[34]* 将 *Gini* 距离协方差和相关推广到了再生核希尔伯特空间（*RKHS*）。通过选择有界核，基于核基尼相关的筛选程序能够放宽条件*C2*，同时仍具有确定性筛选性质和秩一致性性质。他们推论 *7* 中样本核基尼相关的一致指数收敛速率，可用于建立确定性独立筛选性质。

3 模拟

我们通过四项模拟研究评估所提出筛选方法的性能，该方法记为 GN。对于单变量筛选，我们将 GN 与以下四种方法进行比较。

- R^2：单变量方差分析中的 Pearson R^2；
- MV：[3]提出的均值-方差方法；
- AD：[12]研究的 Anderson-Darling 型均值-方差方法；
- DT：基于距离相关的筛选程序 [19, 5]。

请注意，所有方法都可以看作是根据分类变量与数值变量之间的依赖度量对预测变量进行排序。距离方法（DN）基于分类变量的集合差度量。更多详情请参阅 [5]。

与 GN 一样，DN 也用于分组筛选。我们将它们与另外三种组筛选方法进行比较。

- EN: [13]提出的熵信息增益法;
- ENR: [32]使用的熵信息比率法;
- GI: [33]介绍的基尼不纯度方法。

采用包含所有活跃预测变量或所有活跃预测变量组的模型的最小模型规模（MMS），来衡量各筛选方法的有效性。对于每个示例，我们重复该程序 $M = 500$ 次，并计算 MMS 的中位数以及标准差的稳健估计（IQR/1.34，记为 RSD）。此外，对于给定的模型大小 $d = [n/\log\ n]$，我们计算仅包含单个活跃预测变量 X_l 或一个活跃预测变量组 $\boldsymbol{X}_{(l)}$ 的比例（记为 P_l），并报告包含所有活跃预测变量的比例（记为 P_{all}），其中 n 为样本量，$[x]$ 表示 x 的整数部分。$P_{\min}$ 和 $P_{\max}$ 分别是 P_l 的最小值和最大值。

例***3.1***（超高维线性判别分析）
我们参照 Cui、Li和Zhong [3] 以及 Pan、Wang和Li [23]中的类似设定，考虑一个预测变量维度超高的线性判别分析（线性判别分析）问题。我们考虑$K = 3$ 和 $K = 10$ 个类别。对于每个 i^{th} 观测，分类响应 Y_i 从三种不同的离散分布中生成，这些分布包含 K 个类别，其参数为 $\boldsymbol{p} = (p_1, p_2, ...p_K)$，其中 $p_k = P(Y_i = L_k)$: （I）均衡，$p_k = 1/K$，其中 $k = 1, ..., K$；（II）轻度不均衡，$\boldsymbol{p} = (3/12, 4/12, 5/12)$，适用于 $K = 3$，而 $\boldsymbol{p} = (6, 7, 8, 9, 10, 10, 11, 12, 13, 14)/100$ 适用于 $K = 10$；（III）高度不均衡，$\boldsymbol{p} = (0.1, 0.3, 0.6)$适用于 $K = 3$，而 $\boldsymbol{p} = (2, 4, 6, 8, 10, 10, 12, 14, 16, 18)/100$ 适用于 $K = 10$。此设计旨在考察类别不平衡和类别数如何影响各边际筛选方法的表现。

给定 $Y_i = L_k$，随后通过令 $\boldsymbol{X}_i = \boldsymbol{\mu}_k + \boldsymbol{\epsilon}_i$ 来生成 i^{th} 预测变量 $\boldsymbol{X}_i$，其中均值向量 $\boldsymbol{\mu}_k = (\mu_{k1}, \mu_{k2}, \cdots \mu_{kp})^T$ 是一个 p 维向量，其第 k^{th} 个元素为 $\mu_{kk} = 3$，其余元素均为零；而 $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \epsilon_{i2}, \cdots \epsilon_{ip})^T$ 是一个 p 维误差项。这里，我们考虑误差项的两种情形：（1）ϵ_{ij} 独立同分布于 $N(0, 1)$；（2）ϵ_{ij} 独立同分布于 $t(2)$。情况（2）用于考察独立性筛选方法的稳健性。我们考虑 $p = 2000$ 个预测变量，其中 X_1、X_2 和 X_3 是 $K = 3$ 时的活跃预测变量，而 $K = 10$ 时，X_1、X_2、.....和 X_{10} 是活跃预测变量。对于 $K = 3$，样本量为 $n = 60$；对于 $K = 10$，样本量为 $n = 200$。MMS 的中位数和 RSD，以及包含每个活跃预测变量和所有活跃预测变量的比例，均列于表 1。为节省篇幅，对于 $K = 10$，我们仅报告 $P_{\min}$ 和 $P_{\max}$，而不报告全部10个比例。

从表1可以看出，在正态误差项下，当样本量均衡或仅轻度不均衡时，各种筛选方法的表现都很好。严重不均衡会显著影响DT。可以看到，DT中识别出活跃预测变量 X_1 的比例降至52.4%，而 $K = 3$的最小模型规模则从实际的3增至15。5。对于10个分类LDA问题，严重不均衡也会影响MV的性能。这些影响会越来越严重

表1：例3.1中的MMS和选择百分比。

$\boldsymbol{p}$		$K=3(n=60)$						$K=10(n=200)$				
		MMS	RSD	P_1	P_2	P_3	P_{all}	MMS	RSD	P_{min}	P_{max}	P_{all}
						$\varepsilon \sim N(0,\ 1)$						
I	R^2	3	0	1	1	1	1	10	0	1	1	1
	MV	3	0	1	1	1	1	10	0	1	1	1
	AD	3	0	1	1	1	1	10	0	1	1	1
	DT	3	0	1	1	1	1	10	0	1	1	1
	GN	3	0	1	1	1	1	10	0	1	1	1
II	R^2	3	0	1	1	1	1	10	0	1	1	1
	MV	3	0	1	1	1	1	10	0	1	1	1
	AD	3	0	1	1	1	1	10	0	1	1	1
	DT	3	0	1	1	1	1	10	0	0.992	1	0.992
	GN	3	0	1	1	1	1	10	0	1	1	1
III	R^2	3	0	1	1	1	1	10	0	0.984	1	0.984
	MV	3	0.75	0.998	1	1	0.998	65	77.6	0.328	1	0.328
	AD	3	0	1	1	1	1	11	3.73	0.924	1	0.924
	DT	15.5	20.3	0.524	1	1	0.524	641	579	0.016	1	0.016
	GN	3	0	1	1	1	1	10	2.98	0.924	1	0.924
						$\varepsilon \sim t(2)$						
I	R^2	5	45.5	0.834	0.876	0.886	0.626	286	473	0.136	0.848	0.136
	MV	3	0	1	0.998	0.998	0.996	10	0.75	0.994	1	0.994
	AD	3	0	1	0.998	0.998	0.996	10	0	0.996	1	0.996
	DT	3	0	0.996	0.994	0.994	0.988	10	0	0.944	0.998	0.944
	GN	3	0	0.998	0.998	0.996	0.994	10	0	0.994	1	0.994
II	R^2	5	46.6	0.828	0.880	0.870	0.634	314	623	0.130	0.912	0.130
	MV	3	0	0.990	1	1	0.990	12	6.16	0.876	1	0.876
	AD	3	0	0.994	1	1	0.994	10	2.24	0.934	1	0.934
	DT	3	0	0.966	0.988	0.952	0.952	49	90.5	0.416	1	0.416
	GN	3	0	0.996	0.998	0.998	0.998	10	1.49	0.950	1	0.950
III	R^2	26	128	0.564	0.828	0.880	0.428	641	661	0.008	0.914	0.008
	MV	5	10.4	0.726	1	1	0.726	174	211	0.092	1	0.092
	AD	4	5.22	0.816	0.996	1	0.812	88	158	0.286	1	0.286
	DT	105	164	0.090	0.990	0.996	0.088	999	617	0.000	1	0.000
	GN	3	2.98	0.838	0.996	1	0.838	50	134	0.444	1	0.444

表2：例3.2中的 MMS 和选择百分比。

n	Σ_{ij}	方法	MMS	RSD	P_1	P_2	P_3	P_4	P_5	P_{all}
60	0	R^2	1178	633	0.268	0.310	0.308	0.270	0.008	0.000
		MV	571	558	0.250	0.268	0.290	0.302	0.064	0.000
		AD	554	553	0.260	0.278	0.292	0.300	0.192	0.000
		DT	716	612	0.210	0.230	0.230	0.272	0.178	0.002
		GN	556	543	0.248	0.286	0.292	0.298	0.188	0.000
	$0.5^{\lvert j-i\rvert}$	R^2	501	719	0.732	0.920	0.902	0.636	0.044	0.024
		MV	130	157	0.674	0.888	0.878	0.668	0.136	0.056
		AD	83	109	0.692	0.902	0.884	0.664	0.194	0.076
		DT	66	113	0.608	0.850	0.854	0.630	0.536	0.144
		GN	72	120	0.700	0.900	0.894	0.664	0.288	0.130
200	0	R^2	1034	719	0.948	0.952	0.944	0.906	0.024	0.018
		MV	14	23.1	0.918	0.922	0.926	0.950	1.000	0.738
		AD	12	21.8	0.930	0.930	0.930	0.944	1.000	0.754
		DT	19	42.0	0.882	0.898	0.898	0.938	1.000	0.656
		GN	13	22.4	0.924	0.932	0.928	0.954	1.000	0.756
	$0.5^{\lvert j-i\rvert}$	R^2	122	277	1	1	1	1	0.308	0.308
		MV	6	2.24	1	1	1	1	0.978	0.978
		AD	5	0.00	1	1	1	1	1.000	1.000
		DT	5	0.00	1	1	1	1	1.000	1.000
		GN	5	0.75	1	1	1	1	1.000	1.000

在重尾误差项下。如预期， R^2 方法不具备稳健性，在重尾 t_2 误差下表现不佳。当 $K = 10$ 且误差为重尾时，没有任何方法表现良好，但 GN 的估计值非常接近真实值，且将单个/所有活跃预测变量纳入模型的概率最高。GN 的表现最佳，在本例中优于其他方法。

例***3.2***（二元逻辑回归模型）

在本例中，我们参照 He、Ma和Xu [12] 以及 Mai和Zou [20]中的类似设置，考虑一个逻辑回归模型。二元响应变量由以下逻辑回归模型生成。

$$\log\left\{\frac{P(Y=1|\boldsymbol{X})}{P(Y=-1|\boldsymbol{X})}\right\} = -3 + 2X_1 + 2X_2 + 2X_3 + 3\sin(X_4) + 4X_5^2,$$

其中 $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$，其中 $\boldsymbol{\Sigma} = (\rho_{ij})_{p\times p}$ 有两种情形： $\rho_{ij} = 0$ 和 $\rho_{ij} = 0.5^{|j-i|}, i \neq j$。活跃预测变量的数量为五。考虑两种样本量： $n = 60$ 和 $n = 200$ 。

从表2可以观察到，预测变量之间的依赖结构有助于识别活跃预测变量，而样本量 $n = 60$ 并不充足。R^2 方法在检测方面表现较差X_5 其属于响应变量对数几率的二次形式 Y。所提出的 GN 方法和

表3：例3.3中的MMS和选择百分比。

ϵ	方法	MMS	RSD	P_1	P_2	P_{10}	P_{20}	P_{100}	P_{all}
$N(0,1)$	R^2	8.0	20.9	0.938	0.934	0.996	0.984	0.880	0.766
	MV	12.5	42.9	0.904	0.908	0.982	0.976	0.846	0.678
	AD	10.0	32.1	0.922	0.920	0.986	0.982	0.854	0.712
	DT	35.5	127.8	0.790	0.760	0.940	0.950	0.854	0.506
	GN	11.0	36.7	0.918	0.914	0.982	0.980	0.854	0.698
$t(1)$	R^2	1451.0	851.8	0.284	0.268	0.388	0.410	0.278	0.036
	MV	50.5	146.3	0.834	0.810	0.948	0.928	0.774	0.458
	AD	51.0	142.5	0.818	0.814	0.944	0.928	0.766	0.446
	DT	358.5	584.9	0.506	0.478	0.720	0.716	0.652	0.116
	GN	64.0	175.0	0.796	0.788	0.922	0.894	0.760	0.412

AD方法是表现最好的方法之一，具有最高的 P_{all} 和最小的MMS。

例***3.3***（全基因组关联研究）

本例中，我们沿用Cui、Li和Zhong [3]中的相同设定，研究超高维全基因组关联研究（GWAS）。在GWAS中，研究人员可以采集遗传数据，其中通常包含数量极其庞大的单核苷酸多态性（SNP）。作为预测变量的SNP通常是具有三个类别的分类变量，记为 AA， Aa， aa。在本例中， Z_{ij} 表示该 j^{th} SNP对 i^{th} 受试者的显性效应指示变量，其生成方式如下

$$Z_{ij} = \begin{cases} 1, & \text{if} \quad X_{ij} < q_1 \\ 0, & \text{if} \quad q_1 \le X_{ij} < q_3, \\ -1, & \text{if} \quad X_{ij} \ge q_3 \end{cases}$$

其中 $\boldsymbol{X}_i = (X_{i1},\ X_{i2},\ \ldots\ldots,\ X_{ip}) \sim N(\mathbf{0},\ \boldsymbol{\Sigma})$ 对于 $i = 1,\ 2,\ \ldots\ldots,\ n$，且 $\boldsymbol{\Sigma} = (\rho_{jl})_{p\times p}$ 满足 $\rho_{jl} = 0.5^{|j-l|}$，对于 $j,\ l = 1,\ 2,\ \ldots\ldots,\ p$； q_1 和 q_3 分别是标准正态分布的第一四分位数和第三四分位数。响应变量随后按以下方式生成

$$Y = \beta_1 Z_1 + \beta_2 Z_2 + 2\beta_3 Z_{10} + 2\beta_4 Z_{20} - 2\beta_5 |Z_{100}| + \epsilon,$$

其中，对于 $j = 1$=2, ..., 5， $\beta_j = (-1)^U (2\log(n)/\sqrt{n} + |Z_j|)$， $U \sim$服从伯努利分布(0.4)，且 Z_j 独立同分布于 $N(0, 1)$。误差项 ϵ 服从 $N(0, 1)$ 或 $t(1)$。有五个活跃的 SNP： Z_1、 Z_2、 Z_{10}、 Z_{20} 和 Z_{100}。我们设定 $n = 200$ 和 $p = 2000$。该程序重复 500 次。结果见表3。

从表3可以得出结论： R^2 在正态误差下表现最佳，但在 t_1 误差项下表现最差。DT方法对重尾误差并不稳健。DT似乎难以识别弱信号预测变量 Z_1 和 Z_2，其中 P_1 和 P_2 与其他指标相比都相当低。我们的GN方法与AD和MV方法表现相似，能够有效识别与响应变量呈线性或非线性相关的活跃分类预测变量。

例***3.4***（组预测变量筛选）

我们比较了分类中的组筛选方法的性能。预测变量 $\boldsymbol{X}$ 被分组为$(\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}, ..., \boldsymbol{X}_{(r)})$，其中 $\boldsymbol{X}_{(l)} = (X_{3(l-1)+1}, X_{3(l-1)+2}, X_{3l})^T$，适用于 $l = 1, ..., r$。在此示例中，有两个活跃预测变量组 $\boldsymbol{X}_{(1)} = (X_1, X_2, X_3)^T$和 $\boldsymbol{X}_{(2)} = (X_4, X_5, X_6)^T$。同样，MMS和RSD分别是包含这两个组的模型的模型大小的中位数和稳健标准差。 P_1、 P_2 和 P_{all} 分别是在给定模型大小 $d = [n/\log\ n]$的模型中包含 $\boldsymbol{X}_{(1)}$、 $\boldsymbol{X}_{(2)}$ 及两者的比例。

我们考虑 $K = 3$。对于每个 i^{th} 观测，分类响应变量 Y_i 从三个具有3个类别的离散分布中生成，其 $\boldsymbol{p} = (p_1, p_2, p_3)$具有 $p_k = P(Y_i = L_k)$：(I) 均衡，$\boldsymbol{p} = (1/3, 1/3, 1/3)$；(II) 轻度不均衡，$\boldsymbol{p} = (3/12, 4/12, 5/12)$；以及 (III) 高度不均衡，$\boldsymbol{p} = (0.1, 0.3, 0.6)$。我们考虑两种情形。在情形（a）中，两个组在预测分类变量 Y时发挥相同作用。在情形（b）中，$\boldsymbol{X}_{(2)}$ 的信号比 $\boldsymbol{X}_{(1)}$弱。具体而言，给定$Y_i = L_k$，i^{th} 预测变量 $\boldsymbol{X}_i$ 由位置模型 $\boldsymbol{X}_i = \boldsymbol{\mu}_G^{(k)} + \boldsymbol{\epsilon}_i$生成，其中 $\boldsymbol{\epsilon}_i$ 是一个d维误差项，独立服从 $N(0, 1)$或 $t(2)$，且 $\boldsymbol{\mu}_G^{(k)} = (\boldsymbol{\mu}_1^{(k)}, \boldsymbol{\mu}_2^{(k)}, 0, ..., 0)^T$。令$\boldsymbol{\mu}^{(1)} = (1, 0, 0)^T$、 $\boldsymbol{\mu}^{(2)} = (0, 1, 0)^T$以及 $\boldsymbol{\mu}^{(3)} = (0, 0, 1)^T$。在情形（a）中，$\boldsymbol{\mu}_1^{(k)} = \boldsymbol{\mu}_2^{(k)} = 1.5\boldsymbol{\mu}^{(k)}$。在情形（b）中，$\boldsymbol{\mu}_1^{(k)} = 2\boldsymbol{\mu}^{(k)}$，而 $\boldsymbol{\mu}_2^{(k)} = \boldsymbol{\mu}^{(k)}$。我们将样本量 n 设为200，组数 r 设为1000。重复次数 M 为500。

由于 EN、ENR 和 GI 是为分类组筛选设计的，因此 $\boldsymbol{X}_i$的每个分量都必须进行离散化，并作为分类变量处理。正如提出这些方法的作者所建议的，使用标准正态分位数（-0.67、0、0.67）将每个变量分成 4 个类别。

从表 4 可以看出，EN、ENR 和 GI 在所有情形下都完全失效，这并不令人意外，因为它们在分箱过程中损失了过多信息。此外，它们的计算负担极大，计算复杂度随组大小呈指数增长，并随切片数量呈多项式增长。在正态误差项下，DT 和 GN 在情形（a）中表现完美，而在情形（b）III 中，GN 略优于 DT，其 P_2 和 P_{all}百分比略高。在 t（2）误差项下，GN 相较于 DT 的优势变得十分明显，其 MMS 小得多，而识别百分比更高。

4 真实数据分析

我们比较了两组癌症数据集上的筛选方法在基因选择和分类方面的表现。其中一组是 Bhattacharjee 等人 [1] 以及 Cui、Li 和 Zhang [3]研究过的人类肺癌数据。另一组是来自 UCSC Xena 数据库 [10]的 TCGA 乳腺癌微阵列数据集。

4.1 肺癌分类

肺癌数据集包含 203 个样本，每个样本测量了 12,600 个 mRNA 表达水平。这 203 个标本分为五个亚类：139 个肺腺癌（ADEN）、21 个肺鳞状细胞癌（SQUA）、6 个小细胞肺癌（SCLC）、20 个肺类癌（COID），其余 17 个为正常肺样本（NORMAL）。数据已针对每个基因标准化，使其达到零均值和单位方差。

表4：例3.4中的 MMS 和选择百分比。

			$\varepsilon \sim N(0,\ 1)$					$\varepsilon \sim t(2)$				
Case	$\boldsymbol{p}$	方法	MMS	RSD	P_1	P_2	P_{all}	MMS	RSD	P_1	P_2	P_{all}
(a)	I	EN	27	57.1	0.634	0.636	0.392	39.5	86.7	0.566	0.554	0.304
		ENR	18	43.3	0.662	0.690	0.462	19	48.7	0.658	0.694	0.452
		GI	712	391	0.112	0.104	0.010	620	387	0.116	0.108	0.008
		DT	2	0	1	1	1	16	47.0	0.692	0.696	0.486
		GN	2	0	1	1	1	2	0.75	0.968	0.962	0.930
	II	EN	133	209	0.310	0.330	0.114	159	248.9	0.250	0.298	0.076
		ENR	22	49.4	0.630	0.640	0.406	32	74.3	0.566	0.610	0.336
		GI	912	173	0.022	0.030	0.000	882	257	0.024	0.034	0.000
		DT	2	0	1	1	1	20	50.9	0.678	0.686	0.430
		GN	2	0	1	1	1	2	1.50	0.952	0.950	0.904
	III	EN	640	325.4	0.046	0.040	0.002	591	358	0.058	0.050	0.004
		ENR	7	18.7	0.796	0.796	0.654	20	44.8	0.658	0.644	0.426
		GI	998	11.9	0.000	0.000	0.000	991	33.6	0.002	0.000	0.000
		DT	2	0	1	1	1	25	59.7	0.600	0.610	0.378
		GN	2	0	1	1	1	3	4.48	0.908	0.906	0.830
(b)	I	EN	64	166	0.872	0.288	0.240	94	192	0.726	0.244	0.168
		ENR	86	150	0.926	0.218	0.200	93.5	188	0.850	0.222	0.192
		GI	698	383	0.142	0.048	0.006	671	386	0.120	0.056	0.012
		DT	2	0	1	0.990	0.990	62	134	0.890	0.238	0.212
		GN	2	0	1	0.990	0.990	7	23.9	1.000	0.632	0.632
	II	EN	195	272	0.514	0.130	0.064	230	277	0.426	0.102	0.048
		ENR	77	154	0.924	0.244	0.222	94	177	0.824	0.202	0.168
		GI	912	169	0.034	0.018	0.002	849	238	0.036	0.024	0.000
		DT	2	0	1	0.994	0.994	57.5	120	0.926	0.236	0.214
		GN	2	0	1	0.994	0.994	6	23.5	0.998	0.644	0.642
	III	EN	623	338	0.048	0.026	0.000	592	324	0.048	0.028	0.000
		ENR	52	118	0.986	0.290	0.286	66.5	153	0.882	0.296	0.254
		GI	1000	4.48	0.000	0.002	0.000	993	25.7	0.002	0.004	0.000
		DT	2	0.75	1	0.910	0.910	76.5	144	0.900	0.198	0.176
		GN	2	0	1	0.970	0.970	17	51.9	0.990	0.476	0.468

表5：采用线性判别分析和随机森林分类方法时，各筛选方法选出的模型大小及误差（%）。括号内为标准差。

α	方法	LDA			随机森林		
		训练误差	测试误差	Size	训练误差	测试误差	Size
0.9	R2	6.76(1.79)	14.33(6.58)	33.5(4.26)	11.35(1.45)	12.43(4.88)	34.7(3.61)
	MV	3.44(1.02)	8.67(5.92)	20.8(6.10)	5.74(0.68)	8.20(5.44)	25.4(7.60)
	AD	4.25(1.41)	9.10(5.79)	24.7(9.10)	6.05(0.69)	8.86(5.58)	20.5(8.21)
	DT	4.80(1.86)	10.57(5.85)	29.2(7.44)	9.56(1.25)	11.14(5.63)	20.9(10.4)
	GN	4.59(1.11)	8.95(5.76)	31.3(5.15)	6.35(1.14)	7.98(4.53)	30.9(4.62)
0.8	R2	6.45(2.08)	14.8(4.35)	31.7(5.39)	11.38(1.86)	11.90(3.28)	34.9(3.91)
	MV	3.55(1.41)	7.98(4.06)	20.6(7.09)	5.01(0.84)	6.88(3.73)	25.3(8.82)
	AD	4.29(1.60)	8.88(4.57)	22.6(9.09)	6.12(0.82)	7.55(3.99)	20.7(7.53)
	DT	4.09(1.95)	9.83(4.34)	29.8(6.45)	9.70(1.43)	10.78(3.44)	20.0(8.15)
	GN	3.91(1.72)	8.78(4.41)	31.0(6.45)	6.50(1.23)	7.15(3.36)	30.6(5.34)

按照 Cui、Li 和 Zhang 的类似程序 [3], 我们从每个子类中随机抽取大约100α% 的观测值作为训练样本，其余观测值作为测试样本。 α 设为 0.9 和 0.8。将各筛选方法与线性判别分析结合应用于训练集，并使用测试样本评估其性能。具体而言，每种筛选方法首先基于训练数据将维数降至 $d = \lfloor n/\log n \rfloor = 38$。然后通过线性判别分析中的留一交叉验证，从 2 到 d 中选择最优模型大小。该程序重复 100 次。训练分类误差、测试分类误差和模型大小的均值与标准差列于表5。采用线性判别分析分类方法时，MV 在测试误差最低且模型大小较小方面表现最佳。GN 的测试误差次低，但其模型大小相对较大。 R^2 表现最差，误差最大，模型大小也最大。我们还可以观察到，对于每种筛选方法，测试误差都远大于训练误差；测试误差超过训练误差的两倍。这表明线性判别分析分类方法存在过拟合问题。为克服这一局限，我们采用常用的随机森林（RF）作为分类方法来评估筛选方法。

根据训练数据，第一步仍是通过每种筛选方法选取 d_1 个基因。第二步是进一步将模型大小缩减至最优值。我们没有采用留一法交叉验证，而是针对大小从 2 到 d_1的每个模型，使用 200 棵树的随机森林重复运行 20 次。选择袋外分类误差平均值最小的模型作为最优模型。随后，在 500 棵随机树下计算最优模型的训练误差和测试误差。整个程序重复 100 次。分类误差和模型大小的均值与标准差列于表5。

在随机森林中，各筛选方法的测试误差与训练误差之差远小于线性判别分析中的差值。除 DT 外，随机森林在所有筛选方法下的测试误差均小于线性判别分析。DT 在随机森林中的测试误差高于线性判别分析，可能是因为随机森林选出的模型较小。对于 GN 筛选方法，线性判别分析和随机森林选出的模型大小相近，但随机森林的测试分类误差小于线性判别分析。AD 的分类改进甚至是在随机森林中使用更小的模型实现的。总体而言，GN 表现

良好，有时甚至是最佳方法，但其模型大小相对较大。

4.2 乳腺癌基因选择

来自 UCSC Xena 数据库的 TCGA 乳腺癌微阵列数据集 [10] 包含 506 名患者的 17278 个基因表达水平。患者被分为四种亚型：管腔型 A、管腔型 B、HER2 富集型和基底样型。对于乳腺癌，PAM50 是由 50 个基因组成的基因特征，被视为乳腺癌亚型预后和预测的金标准[24]。该数据集包含 PAM50 中的 48 个基因。

我们展示两种分组筛选方法 DT 和 GN，用于选择称为 PAM48 的特征基因组。

我们考虑剩余 17230 个基因中的四种比较组。第一种是按照数据集中基因的原始顺序构建分组。组大小分别设为 50、100、150 和 200。相应的组数 r 分别为 344、172、114 和 86。第二种是从 1 到 17230 的基因索引中进行无放回抽样，以构建随机分组；第三种则是有放回抽样构建随机分组。第四种是根据每个基因与癌症亚型类别的边际相关性排序来构建分组。对于每种类型，选取相关性最高的组与 PAM48 进行比较。由这些分组类型选出的基因所构建的模型分别记为 Original、Random1、Random2 和 Marginal。

此外，我们比较了采用所选基因组的模型的分类性能。我们随机留出 20% 的数据作为测试数据。基于其余 80% 的训练数据，我们通过基尼方法或距离方法选出排名最高的组。我们采用随机森林（RF）方法进行分类。对于每个模型，我们使用 RF 和 200 棵树重复运行 10 次，然后根据训练和测试分类误差评估所选模型的性能。

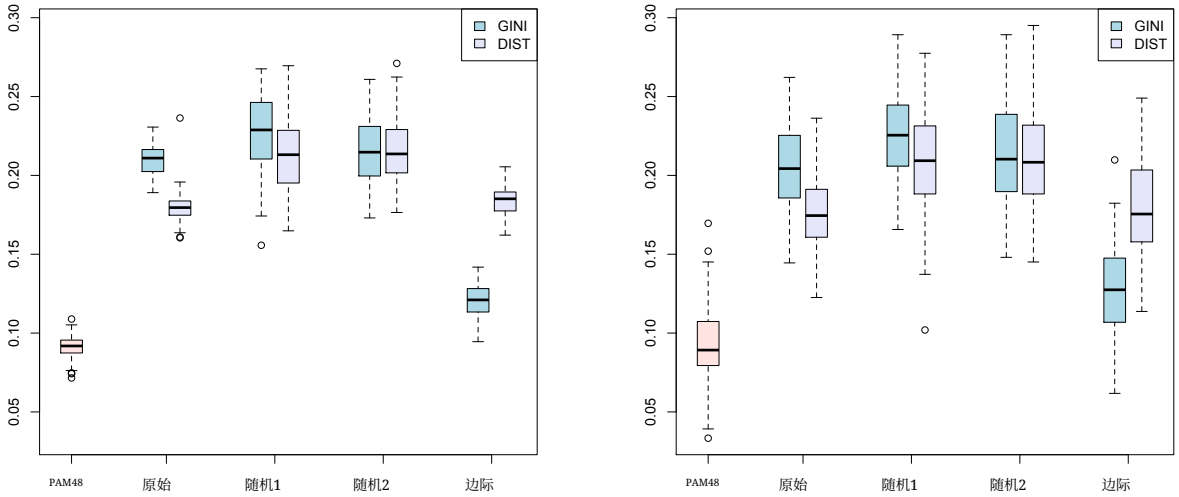
整个程序重复 100 次。我们汇总训练和测试分类误差的均值与标准差，以及所选组与 PAM48 的相关性，结果见表 6。图 1. 展示了组大小为 50 的模型的分类误差分布。

首先，就基尼相关或距离相关而言，PAM48 在所有考察的组中排名最高。从这个意义上说，两种组筛选方法都成功选出了特征基因组。从表 6 可以看出，PAM48 的平均基尼相关为 0.269，而相关性次高的组是边际组，其平均值从组大小为 50 时的 0.260 到组大小为 200 时的 0.215 不等。两个随机组的基尼相关最低。距离相关也呈现出类似的结果。然而，边际组与其他组之间的距离相关差异较小。

在分类性能方面，PAM48 表现最佳，训练误差和测试误差均为最低。边际组表现次佳。预排序确实有助于减少误分类。尤其是在基尼相关排序中，改进幅度很大：在大小为 50 的模型中，误差几乎降低了 8%。即使在大小为 150 或 200 的模型中，与原始组相比，也能看到 3% 的降幅。距离相关排序对较大的模型有益，但对大小为 50 的模型没有帮助。两个随机组在所有大小下的表现都不如其他组。测试误差的波动大于训练误差是合理的，而且两个随机模型的误差标准差也大于其他模型。

表6：与PAM48相比，选定模型的相关性和随机森林分类误差（%）。括号内为标准差。

模型 Size	模型	组相关性		训练误差		测试误差	
		GN	DT	GN	DT	GN	DT
	PAM48	.269(.004)	.750(.007)	9.128(.696)		9.280(2.531)	
50	原始	.186(.004)	.695(.008)	21.0(0.93)	18.0(0.89)	20.5(2.70)	17.6(2.31)
	随机1	.147(.012)	.688(.012)	22.7(2.37)	21.4(2.32)	22.5(2.84)	21.0(3.07)
	随机2	.153(.017)	.693(.011)	21.6(2.13)	21.6(2.07)	21.2(3.26)	21.0(3.15)
	边际	.260(.005)	.698(.006)	12.0(1.03)	18.4(0.90)	12.7(2.97)	18.1(3.00)
100	原始	.133(.002)	.686(.008)	19.8(1.01)	19.3(1.48)	19.3(2.51)	18.6(3.21)
	随机1	.118(.008)	.697(.011)	20.9(1.94)	20.0(2.14)	20.0(3.33)	19.0(2.71)
	随机2	.124(.014)	.699(.012)	20.4(1.91)	20.0(1.88)	19.8(2.82)	19.5(2.95)
	边际	.245(.005)	.703(.006)	11.3(0.72)	14.9(0.98)	11.8(2.80)	15.0(2.71)
150	原始	.112(.002)	.700(.007)	15.8(0.87)	15.8(0.71)	15.1(2.55)	15.0(2.39)
	随机1	.104(.007)	.698(.010)	19.4(1.77)	19.5(1.89)	18.8(2.86)	18.5(3.07)
	随机2	.108(.009)	.704(.011)	19.5(1.67)	19.3(1.69)	18.5(2.95)	18.6(2.74)
	边际	.228(.004)	.705(.005)	11.5(0.74)	13.2(0.92)	11.8(2.71)	13.4(2.96)
200	原始	.103(.002)	.706(.007)	16.4(0.99)	15.8(0.70)	15.9(2.64)	15.1(2.36)
	随机1	.095(.006)	.699(.010)	19.3(1.74)	19.1(1.74)	18.7(3.11)	18.2(2.68)
	随机2	.099(.008)	.704(.012)	19.1(1.71)	18.8(1.53)	18.4(2.95)	18.0(2.93)
	边际	.215(.005)	.706(.006)	11.6(0.73)	12.4(0.74)	12.1(2.73)	12.5(2.89)



(a) 训练误分类率　　(b) 测试误分类率

图1：基尼方法与距离方法的分组筛选比较。(a) 选定模型大小为50时的误分类训练误差箱线图，并与PAM48模型进行比较。(b) 选定模型大小为50时的误分类测试误差箱线图，并与PAM48模型进行比较。

模型大小为50时的结果可见图1。两幅图的刻度相同，因此可以比较。(b)中的箱线图具有更大的四分位距和更长的须。看来，原始排序有助于DT方法进行模型选择，但在GN方法中，相比随机分组，其改进幅度很小。同样，除PAM48外，GN方法选出的边际分组模型都明显优于其他模型。

5 结论与讨论

在相关文献中，自Fan和Lv [6]以来，针对边际预测变量的确定性筛选程序已取得丰硕成果。然而，针对超高维数据中分组特征的筛选程序仍十分有限。我们提出了一种基于GDC的筛选程序GN，适用于响应变量为分类变量时的个体特征和分组特征。该方法易于实现、无需依赖模型，并具有确定性筛选性质。在更强的可分性条件下，我们还为该筛选程序建立了更强的理论结果，即排序一致性性质。模拟研究表明，对于个体特征和分组特征筛选，所提出的GN方法优于目前大多数已有方法。

虽然我们的 GN 程序主要针对数值协变量，而 [33] 中的基尼不纯度方法针对分类协变量，但探究二者之间的联系颇有意思。定义两个分类变量之间的离散距离为

$$\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| = \mathbb{1}(\boldsymbol{x}_1 \neq \boldsymbol{x}_2),$$

其中 $\mathbb{1}(\boldsymbol{x})$ 是指示函数。于是，分类随机变量的基尼平均差就是其基尼不纯度，这是评估分类性能时常用的指标。也就是说，对于响应分类变量 Y，

$$\mathbb{E}\|Y_1 - Y_2\| = \sum_{k=1}^{K} P(Y_1 = L_k)(1 - P(Y_2 \neq L_k)) = 1 - \sum_{k=1}^{K} p_k^2 = \mathrm{GI}(Y).$$

此外，如果组　　协变量 $\boldsymbol{X}_{(l)}$ 是分类变量，我们在（3）中定义的基尼相关就变为　　e

$$\mathrm{gCor}(\boldsymbol{X}_{(l)}, Y) = \rho_l = \frac{\mathrm{GI}(\boldsymbol{X}_{(l)}) - \mathrm{GI}(\boldsymbol{X}_{(l)}|Y)}{\mathrm{GI}(\boldsymbol{X}_{(l)})}, \tag{8}$$

而 [33]中使用的指标为

$$\mathrm{GI}(X_{(l)}, Y) = \frac{\mathrm{GI}(Y) - \mathrm{GI}(Y|\boldsymbol{X}_{(l)})}{\log N_l},$$

其中 N_l 是 $\boldsymbol{X}_{(l)}$的类别总数。直观而言，（8）中的标准化项更加自然，并且利用了更多关于 $\boldsymbol{X}_{(l)}$的信息。因此，当组协变量为分类变量时，所提出的筛选方法即可应用。

6 附录

我们首先给出用于证明定理的引理。引理 6.1 和 6.2 出自 [27]和 [19], ；引理6.3 出自 [3] 中的引理 A.3，该引理由伯恩斯坦不等式推导而来。

引理 6.1 设 W 是一个随机变量，满足 $P(a \leq W \leq b) = 1$ 且 $\mu = \mathbb{E}W$，则

$$\mathbb{E}[\exp\{s(W - \mu)\}] \leq \exp\{s^2(b-a)^2/8\}, \quad for\ any\ s > 0.$$

引理 6.2 设 $h(X_1, X_2, ..., X_m)$ 是 U-统计量 U_n 的核，且 $\theta = \mathbb{E}h(X_1, ..., X_m)$。若$a \leq h(X_1, X_2, ..., X_m) \leq b$，则

$$Pr(U_n - \theta \geq t) \leq \exp\{-2[n/m]t^2/(b-a)^2\}, \quad for\ any\ t > 0,\ and\ n \geq m,$$

其中 $[n/m]$ 表示 n/m 的整数部分。

引理 6.3 对于 $k = 1, .., K$，设 $p_k = F(Y = L_k)$ 和 $\hat{} p_k = \frac{1}{n}\sum_{i=1}^{n} I(Y = L_k)$，我们有

$$P(|\hat{p}_k - p_k| \geq \varepsilon) \leq 2\exp\{-\frac{n\varepsilon^2}{2(p_k + \varepsilon/3)}\}$$

对于任意 $\varepsilon \in$（0、1）。

我们通过分别证明分母和分子的一致相合性来证明不等式（5）。ρ_l。

对于 $l = 1, \ldots\ldots, r$，式（ $\boldsymbol{X}_{(l)}$与 Y之间的吉尼协方差可由1）估计为

$$\begin{aligned}\hat{\mathrm{gCov}}(\boldsymbol{X}_{(l)}, Y) &= \binom{n}{2}^{-1} \sum_{1\le i<j\le n} \|\boldsymbol{X}_{(l)i} - \boldsymbol{X}_{(l)j}\| - \sum_{k=1}^{K} \hat{p}_k \binom{n_k}{2}^{-1} \sum_{1\le i<j\le n_k} \|\boldsymbol{X}_{(l|k)i} - \boldsymbol{X}_{(l|k)j}\| \\ &:= U_{n,l} - \sum_{k=1}^{K} \hat{p}_k U_{n_k,l}.\end{aligned}$$

令 $h(\boldsymbol{X}_{(l)1}, \boldsymbol{X}_{(l)2}) = \|\boldsymbol{X}_{(l)1} - \boldsymbol{X}_{(l)2}|$ 为 $U_{n,\ l}$的核，它可以分解为：$h = h\mathbb{1}(h > M) + h\mathbb{1}(h \le M)$，其中 M将在后文给出。于是

$$\begin{aligned}U_{n,l} &= \binom{n}{2}^{-1} \sum_{1\le i<j\le n} h(\boldsymbol{X}_{(l)i}, \boldsymbol{X}_{(l)j})\mathbb{1}(h \le M) + \binom{n}{2}^{-1} \sum_{1\le i<j\le n} h(\boldsymbol{X}_{(l)i}, \boldsymbol{X}_{(l)j})\mathbb{1}(h > M) \\ &:= \hat{R}_{l,1} + \hat{R}_{l,2}\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}\|\boldsymbol{X}_{(l)1} - \boldsymbol{X}_{(l)2}\| \\ &= \mathbb{E}\left\{\|\boldsymbol{X}_{(l)1} - \boldsymbol{X}_{(l)2}\|\mathbb{1}(h(\boldsymbol{X}_{(l)1}, \boldsymbol{X}_{(l)2}) \le M)\right\} + \mathbb{E}\left\{\|\boldsymbol{X}_{(l)1} - \boldsymbol{X}_{(l)2}\|\mathbb{1}(h(\boldsymbol{X}_{(l)1}, \boldsymbol{X}_{(l)2}) > M)\right\} \\ &:= R_{l,1} + R_{l,2}.\end{aligned}$$

ˆ ˆ 有 $\mathbb{E}R_{l,\ 1} = R_{l,\ 1}$ 以及 $\mathbb{E}R_{l,\ 2} = R_{l,\ 2}$。应用马尔可夫不等式可得，对于任意 $\varepsilon > 0$和 $\gamma > 0$

$$P(\hat{R}_{l,1} - R_{l,1} \ge \varepsilon) \le \exp(-\gamma\varepsilon)\exp(-\gamma R_{l,1})\mathbb{E}\{\exp(\gamma\hat{R}_{l,1})\}.$$

根据任意 U 统计量都可以表示为 [27],中独立同分布随机变量的平均值的平均值这一事实，我们可以得到

$$\hat{R}_{l,1} = (n!)^{-1} \sum_{n!} \Lambda_1(\boldsymbol{X}_{(l)i_1}, \boldsymbol{X}_{(l)i_2}, ..., \boldsymbol{X}_{(l)i_n}),$$

其中 $\sum_{n!}$表示对所有 $n!$ 个排列 $(i_1, \ldots i_n)$ of $(1, \ldots\ldots, n)$ 求和，并且每个 Λ_1 都是 $m = [n/2]$ 个独立同分布随机变量的平均值。更具体地说，

$$\Lambda_1(\boldsymbol{x}_1, ..., \boldsymbol{x}_n) = \frac{1}{m}\sum_{t=1}^{m} h(\boldsymbol{x}_{2t-1}, \boldsymbol{x}_{2t})\mathbb{1}(h(\boldsymbol{x}_{2t-1}, \boldsymbol{x}_{2t}) \le M).$$

根据詹森不等式，对于 $1 < \gamma \le 2s_0$,

$$\begin{aligned}\mathbb{E}\left\{\exp(\gamma\hat{R}_{l,1})\right\} &= \mathbb{E}\left\{\exp\left(\gamma(n!)^{-1}\sum_{n!}\Lambda_1(\boldsymbol{X}_{(l)1}, \boldsymbol{X}_{(l)2}, ..., \boldsymbol{X}_{(l)n})\right)\right\} \\ &\le (n!)^{-1}\sum_{n!}\mathbb{E}\left\{\exp\left(\gamma\Lambda_1(\boldsymbol{X}_{(l)1}, \boldsymbol{X}_{(l)2}, ..., \boldsymbol{X}_{(l)n})\right)\right\} \\ &= \mathbb{E}^m\left\{\exp\left(\frac{1}{m}\gamma h(\boldsymbol{X}_{(l)1}, \boldsymbol{X}_{(l)2})\mathbb{1}(h(\boldsymbol{X}_{(l)1}, \boldsymbol{X}_{(l)2}) \le M)\right)\right\},\end{aligned}$$

结合引理1可立即得出

$$\begin{aligned}
P(\hat{R}_{l,1} - R_{l,1} \geq \varepsilon) &\leq \exp(-\gamma\varepsilon)\exp(-\gamma R_1)\mathbb{E}\{\exp(\gamma \hat{R}_{l,1})\} \\
&\leq \exp(-\gamma\varepsilon)\exp(-\gamma R_{l,1})\mathbb{E}^m\left\{\exp\left(\frac{1}{m}\gamma h^{(t)}\mathbb{1}(h^{(t)} \leq M)\right)\right\} \\
&= \exp(-\gamma\varepsilon)\mathbb{E}^m\left\{\exp\left(\frac{1}{m}\gamma[h^{(t)}\mathbb{1}(h^{(t)} \leq M) - R_{l,1}]\right)\right\} \\
&\leq \exp\Big(-\gamma\varepsilon + M^2\gamma^2/(8m)\Big).
\end{aligned}$$

取 $\gamma = 4\varepsilon m/M^2$，则有 $P(\hat{R}_{l,1} - R_{l,1} \geq \varepsilon) \leq \exp(-2\varepsilon^2 m/M^2)$。因此，

$$P(|\hat{R}_{l,1} - R_{l,1}| \geq \varepsilon) \leq 2\exp(-2\varepsilon^2 m/M^2). \tag{9}$$

接下来我们证明 $\hat{R}_{l,2}$ 的一致性。由柯西-施瓦茨不等式和马尔可夫不等式，

$$\begin{aligned}
R_{l,2}^2 &\leq \mathbb{E}\big[h^2(\boldsymbol{X}_{(l)i}, \boldsymbol{X}_{(l)j})\big]P\big(h(\boldsymbol{X}_{(l)i}, \boldsymbol{X}_{(l)j}) > M\big) \\
&\leq \mathbb{E}\big[h^2(\boldsymbol{X}_{(l)i}, \boldsymbol{X}_{(l)j})\big]\mathbb{E}\big[\exp\big(s_1 h(\boldsymbol{X}_{(l)i}, \boldsymbol{X}_{(l)j})\big)\big]/\exp(s_1 M) \\
&\leq 2\mathbb{E}\big[\|\boldsymbol{X}_{(l)i}\|^2\big]\{\mathbb{E}\big[\exp(s_1\|\boldsymbol{X}_{(l)i}\|)\big]\}^2/\exp(s_1 M)
\end{aligned}$$

对任意 $s_1 > 0$。

若对 $0 < \beta < 1/2 - \kappa$ 取 $M = cn^\beta$，在条件 **C2** 下，当 n 足够大时，$R_{l,2} \leq \varepsilon/2$。因此我们有

$$P(|\hat{R}_{l,2} - R_{l,2}| > \varepsilon) \leq P(|\hat{R}_{l,2}| > \varepsilon/2).$$

此外，对某个 $1 \leq i \leq n\}$，有 $\{|R_{l,2}| > \varepsilon/2\} \subset \{\|\boldsymbol{X}_{(l)i}\| > M/2,$。事实上，若对所有 $1 \leq i \leq n$ 都有 $\|\boldsymbol{X}_{(l)i}\| \leq M/2$，则 $h(\boldsymbol{X}_{(l)i}, \boldsymbol{X}_{(l)j}) \leq \|\boldsymbol{X}_{(l)i}\| + \|\boldsymbol{X}_{(l)j}\| \leq M$ 且 $R_{l,2} = 0$，这与 $\hat{R}_{l,2} > \varepsilon/2$ 的事实矛盾。由条件 **C2** 和马尔可夫不等式，我们有

$$P(\|\boldsymbol{X}_{(l)i}\| > M/2) \leq C\exp(-sM/2).$$

因此

$$\begin{aligned}
\max_{1\leq l\leq r} P(|\hat{R}_{l,2}| > \varepsilon/2) &\leq n \max_{1\leq l\leq r} P(\|\boldsymbol{X}_{(l)i}\| > M/2) \\
&\leq Cn\exp(-sM/2)
\end{aligned}$$

$$\begin{aligned}
P(|\hat{R}_{2,l} - R_{l,2}| > \varepsilon) &\leq P(|\hat{R}_{2,l}| > \varepsilon/2) \\
&\leq nC\exp(-sM/2).
\end{aligned} \tag{10}$$

结合（9）和（10），

$$P(|U_{n,l} - \Delta_l| \geq 2\varepsilon) \leq 2\exp(-\varepsilon^2 n^{1-2\beta}) + nC\exp(-sn^\beta/2), \tag{11}$$

其中 $0 < \beta < 1/2 - \kappa$。用相同的论证，对于 $k = 1, \ldots, K$，我们有

$$P(|U_{n_k,l} - \Delta_{l|k}| \geq 2\varepsilon) \leq 2\exp(-\varepsilon^2 n_k^{1-2\beta}) + n_k C\exp(-sn_k^\beta/2). \tag{12}$$

根据（11）和（12）的结果，我们准备证明下述引理中所述的样本基尼协方差的集中性结果。

引理 6.4 在条件 ***C1*** 和 ***C2*** 下，对于任意 $\varepsilon \in$（0，1）、 $l = 1$、...、 r 和 $0 < \beta < 1/2 - \kappa$，有

$$
\begin{aligned}
&P(|g\hat{C}ov(\boldsymbol{X}_{(l)}, Y) - gCov(\boldsymbol{X}_{(l)}, Y)| \geq \varepsilon) \\
&\leq 2K_n \exp(-\varepsilon^2 n^{1-2\beta}) + nK_n \exp(-sn^{\beta}/2) + 2K_n \exp\{-C\frac{n}{K_n}\varepsilon^2\}.
\end{aligned}
$$

引理的证明6.4。

$$
\begin{aligned}
\mathrm{g\hat{C}ov}(\boldsymbol{X}_{(l)}, Y) - \mathrm{gCov}(\boldsymbol{X}_{(l)}, Y) &= U_{n,l} - \sum_{k=1}^{K} \hat{p}_k U_{n_k,l} - (\Delta_l - \sum_{k=1}^{K} p_k \Delta_{l|k}) \\
&= U_{n,l} - \Delta_l - \sum_{k=1}^{K} \hat{p}_k U_{n_k,l} + \sum_{k=1}^{K} p_k \Delta_{l|k} \\
&= (U_{n,l} - \Delta_l) - \sum_{k=1}^{K} \hat{p}_k (U_{n_k,l} - \Delta_{l|k}) - \sum_{k=1}^{K} (\hat{p}_k - p_k)\Delta_{l|k} \\
&:= \hat{I}_{l,1} + \hat{I}_{l,2} + \hat{I}_{l,3}
\end{aligned}
$$

由（11），

$$
P(|\hat{I}_{l,1}| \geq \varepsilon) \leq 2\exp(-\varepsilon^2 n^{1-2\beta}/4) + nC\exp(-sn^{\beta}/2), \tag{13}
$$

ˆ ˆ 接下来我们处理$I_{l,\ 2}$ 和 $I_{l,\ 3}$。容易得到

$$
\begin{aligned}
|\hat{I}_{l,2}| &\leq \sum_k \hat{p}_k |U_{n_k,l} - \Delta_{l|k}| \leq \max_k |U_{n_k,l} - \Delta_{l|k}|, \\
|\hat{I}_{l,3}| &\leq \sum_{k=1}^{K_n} |\hat{p}_k - p_k| \Delta_{l|k} \leq K_n \max_k \{|\hat{p}_k - p_k| \Delta_{k,l}\} \leq CK_n \max_k |\hat{p}_k - p_k|.
\end{aligned}
$$

将此式代入（12）可得

$$
\begin{aligned}
P(|\hat{I}_{l,2}| \geq \varepsilon) &\leq P(\max_k |U_{n_k,l} - \Delta_{l|k}| \geq \varepsilon) \\
&\leq K_n P(|U_{n_k,l} - \Delta_{l|k}| \geq \varepsilon) \\
&\leq 2K_n \exp(-\varepsilon^2 n_k^{1-2\beta}/4) + n_k K_n C \exp(-sn_k^{\beta}/2).
\end{aligned} \tag{14}
$$

and

$$
\begin{aligned}
P(|\hat{I}_{l,3}| \geq \varepsilon) &\leq P(CK_n \max_k |\hat{p}_k - p_k| \geq \varepsilon) = P(\max_k |\hat{p}_k - p_k| \geq \varepsilon/(CK_n)) \\
&\leq 2K_n \exp\{-\frac{n\varepsilon^2/(CK_n)^2}{2(p_k + \varepsilon/(3CK_n))}\} \qquad (15) \\
&\leq 2K_n \exp\{-C\frac{n}{K_n}\varepsilon^2\}. \qquad (16)
\end{aligned}
$$

不等式 **(15)** 来自引理 6.3，最后一个不等式由条件 **C1** 推导得出。

结合 (13)、(14) 和 (16)，可得

$$
\begin{aligned}
&P(|\hat{\mathrm{gCov}}(\boldsymbol{X}_{(l)},Y)-\mathrm{gCov}(\boldsymbol{X}_{(l)},Y)|\geq\varepsilon)\\
&\leq P(|\hat{I}_{l,1}|\geq\varepsilon/3)+P(|\hat{I}_{l,2}|\geq\varepsilon/3)+P(|\hat{I}_{l,3}|\geq\varepsilon/3)\\
&\leq 2\exp(-\varepsilon^2n^{1-2\beta}/36)+nC\exp(-sn^{\beta}/2)\\
&+2K_n\exp(-\varepsilon^2n^{1-2\beta}/36)+nCK_n\exp(-sn^{\beta}/2)+2K_n\exp\{-C\frac{n}{K_n}\varepsilon^2\}.\\
&\leq 2K_n\exp(-\varepsilon^2n^{1-2\beta}/36)+nCK_n\exp(-sn^{\beta}/2)+2K_n\exp\{-C\frac{n}{K_n}\varepsilon^2\}.
\end{aligned}
$$

定理2.1的证明。

$\hat{}$ Gini距离相关$\hat{}\rho_l$与 $\mathrm{gCov}(\boldsymbol{X}_{(l)},\ Y)$ $l=1$、2、.....、 r具有相同的收敛速率。令 $\varepsilon=cn^{-\kappa}$，对于 $0<\kappa+\beta<1/2$，在条件 $K_n=O(n^{\tau})$ 下，我们有

$$
\begin{aligned}
&P\big(\max_{1\leq l\leq r}|\hat{\rho}_l-\rho_l|\geq cn^{-\kappa}\big)\leq r\max_{1\leq l\leq r}P\big(|\hat{\rho}_l-\rho_l|\geq cn^{-\kappa}\big)\\
&\leq O\left\{r\left[K_n\exp(-b_1n^{1-2(\kappa+\beta)})+nK_n\exp(-b_2n^{\beta})+2K_n\exp\{-C\frac{n}{K_n}n^{-2\kappa}\}\right]\right\}\\
&\leq O\left\{r\left[\exp(-b_1n^{1-2(\kappa+\beta)}+\tau\log n)+\exp(-b_2n^{\beta}+(1+\tau)\log n)+\exp(-b_3n^{1-2\kappa-\tau}+\tau\log n)\right]\right\},
\end{aligned}
$$

这表明不等式（5）。最后一个不等式由条件 **C1** 得到，即 $K_n=O\left(n^{\tau}\right)$。

如果 $\mathcal{A}\nsubseteq\hat{}\mathcal{A}$，则必定存在某个 $l\in\mathcal{A}$，使得 $\hat{}\rho_l<cn^{-\kappa}$。根据条件 **C2**，$|\hat{}\rho_l-\rho_l|\geq cn^{-\kappa}$ 对于某个 $l\in\mathcal{A}$成立。这意味着 $\{\mathcal{A}\nsubseteq\hat{}\mathcal{A}\}\subseteq\{|\hat{}\rho_l-\rho_l|>cn^{-\kappa},$ 对于某个 $l\in\mathcal{A}\}$ 成立，从而 $\hat{}\mathcal{B}_n=\{\max_{l\in\mathcal{A}}|\hat{}\rho_l-\rho_l|\leq cn^{-\kappa}\}\subseteq\{\mathcal{A}\subseteq\hat{\mathcal{A}}\}$。

$$
\begin{aligned}
P(\mathcal{A}\subseteq\hat{\mathcal{A}})&\geq P(\mathcal{B}_n)=1-P(\mathcal{B}_n^c)=1-P(\min_{l\in\mathcal{A}}|\hat{\rho}_l-\rho_l|\geq cn^{-\kappa})\\
&=1-d_nP(|\hat{\rho}_l-\rho_l|\geq cn^{-\kappa})\\
&\geq 1-O\Big(d_n\left[\exp(-b_1n^{1-2(\kappa+\beta)}+\tau\log n)+\exp(-b_2n^{\beta}+(1+\tau)\log n)\right.\\
&\left.+\exp(-b_3n^{1-2\kappa-\tau}+\tau\log n)\right]\Big),
\end{aligned}
$$

其中 d_n 是 $\mathcal{A}$的基数。我们已完成式（6）的证明。

定理 2.2 的证明。 我们将通过证明2.2 来证明定理

$$
P\left\{\limsup_{n\to\infty}\big(\min_{l\in\mathcal{A}}\hat{\rho}_l-\max_{l\in\mathcal{I}}\hat{\rho}_l\big)<c_3/2\right\}=0.
$$

根据条件 **C4** 和引理 6.4，我们有

$$
\begin{aligned}
P\left\{\big(\min_{l\in\mathcal{A}}\hat{\rho}_l-\max_{l\in\mathcal{I}}\hat{\rho}_l\big)<c_3/2\right\}&\leq P\left\{\big(\min_{l\in\mathcal{A}}\hat{\rho}_l-\max_{l\in\mathcal{I}}\hat{\rho}_l\big)-\big(\min_{l\in\mathcal{A}}\rho_l-\max_{l\in\mathcal{I}}\rho_l\big)<-c_3/2\right\}\\
&\leq P\left\{\big|\big(\min_{l\in\mathcal{A}}\hat{\rho}_l-\max_{l\in\mathcal{I}}\hat{\rho}_l\big)-\big(\min_{l\in\mathcal{A}}\rho_l-\max_{l\in\mathcal{I}}\rho_l\big)\big|>c_3/2\right\}\\
&\leq P\left\{2\max_{1\leq l\leq r}|\hat{\rho}_l-\rho_l|>c_3/2\right\}\\
&\leq 2rK_n\exp(-b_4n^{1-2\beta})+nrK_n\exp(-b_5n^{\beta}/2)+rK_n\exp(-b_6\frac{n}{K_n})
\end{aligned}
$$

对于某些正 常数 b_4, b_5 和 b_6。根据 Borel–Cantelli引理, 只需证明

$$\sum_{n=1}^{\infty}\left\{2rK_n\exp(-b_4n^{1-2\beta})+nrK_n\exp(-b_5n^{\beta}/2)+rK_n\exp(-b_6\frac{n}{K_n})\right\}<\infty.$$

事实上, 由条件**C4**, $K_n\underline{\log_{n^{\alpha}}}^{p}=o(1)$ 和 $K_n\underline{\log_{n^{\alpha}}}^{n}=o(1)$, $\alpha=\min\{1-2\beta+\tau,\beta+\tau,1\}$ 可得

$$r\le p\le\exp(\frac{Cn^{\alpha}}{2K_n})\le\min\{\exp(b_4n^{1-2\beta}/2),\exp(b_5n^{\beta}/2),\exp(b_6n^{1-\tau}/2)\},$$
$$\log K_n\le\log n,\quad 3\log n\le\min\{n^{1-2\beta},\frac{n}{K_n}\},\quad 4\log n\le n^{\beta}$$

当 n 足够大时。因此, 存在 n_0 使得

$$\begin{aligned}
&\sum_{n=n_0}^{\infty}\left\{2rK_n\exp(-b_4n^{1-2\beta})+nrK_n\exp(-b_5n^{\beta}/2)+rK_n\exp(-b_6\frac{n}{K_n})\right\}\\
&\le\sum_{n=n_0}^{\infty}\left\{\exp\Big(\log K_n-b_4n^{1-2\beta}/2\Big)+\exp\Big(\log(nK_n)-b_5n^{\beta}/2\Big)+\exp\Big(\log K_n-b_6n^{1-\tau}/2\Big)\right\}\\
&\le\sum_{n=n_0}^{\infty}\left\{2\exp\Big(\log K_n-3\log n\Big)+\exp\Big(\log(nK_n)-4\log n\Big)\right\}\\
&\le 3\sum_{n=n_0}^{\infty}n^{-2}<\infty.
\end{aligned}$$

证明至此完成。

致谢

感谢 Wei Zhong 分享 MV方法的 R代码, 并与我们分享肺癌数据。

参考文献

[1] Bhattacharjee, A., Richards, W., Staunton,J., Li, C., Monti, S., Vasal, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Lodal, M., Weber, G., Mark, E., Lander, E., Wong, W., Johnson, B., Golub, T., Sugarbaker, D., and Meyerson, M. (2001). 通过mRNA表达谱分析对人类肺癌进行分类, 揭示了不同的腺癌亚类。《美国国家科学院院刊》,**98**, 13790-13795.[2] Cheng, G., Li, X., Lai, P., Song, F. and Yu, J. (2017). 面向超高维判别分析的稳健秩筛选。《统计与计算》, **27** (2), 535-545.[3] Cui, H., Li, R. and Zhong, W. (2015). 面向超高维判别分析的无模型特征筛选。《美国统计学会会刊》, **110**, 630-641.

[4] Curry, J., Dang, X. and Sang, H. (2018). 一种用于两样本的基于秩的 Cram´er-von-Mises 型检验。《巴西概率与统计学杂志》，已录用。arxiv.org/abs/1802.06332.[5] Dang, X., Nguyen, D., Chen, X. and Zhang, J. (2021). 一种新的定量变量与定性变量之间的基尼相关。Scand. J. Stat., 48 (4), 1314-1343.[6] Fan, J. and Lv, J. (2008). 超高维特征空间的确定性独立筛选（附讨论）。英国皇家统计学会杂志 B 辑, 70, 849-911.[7] Fan, J., Feng, Y. and Song, R. (2011). 稀疏超高维加性模型中的非参数独立筛选。《美国统计学会会刊》, 106, 544- 557.[8] Fan, J. and Song, R. (2010). 具有 NP 维数的广义线性模型中的确定性独立筛选。《统计年刊》, 38, 3567-3604.[9] Fan, J., Samworth, R., and Wu, Y. (2009). 通过独立学习进行超高维变量选择：超越线性模型。Journal of Machine Learning Research, 10, 1829-1853.[10] Goldman, M., Craft, B., Brooks, A.N., Zhu, J. and Haussler, D. (2018). 用于癌症基因组学数据可视化与解读的 UCSC Xena 平台。bioRxiv.[11] Hall, P. and Miller, H. (2009). 使用广义相关在超高维问题中进行变量选择。计算与图形统计学杂志, 18, 533-550.[12] He, S., Ma, S. and Xu, W. (2019). 一种改进的均值-方差特征筛选程序，用于超高维判别分析。计算统计学与数据分析, 137, 155 -169.[13] He, H. and Deng, G. (2022). 面向分类模型的超高维数据分组特征筛选。J. Stat. Comput. Simul., 92, 972 -997.[14] Huang, Q. and Zhu, Y. (2016). 通过最大相关进行无模型确定性筛选。多元分析杂志, 148, 89-106.[15] Kendall, M.G. (1938). 一种新的秩相关度量。Biometrika, 60 (2), 185-196.[16] Kendall, M.G. and Gibbons, J.D. (1990). 秩相关方法， 5^{th} 版，Griffin，伦敦。[17] Lai, P., Song, F., Chen, K. and Liu, Z. (2017). 超高维二分类中含因变量的无模型特征筛选。统计与概率快报, 125, 141-148.[18] Li, G., Peng, H., Zhang, J, and Zhu, L. (2012). 基于稳健秩相关的筛选，《统计年刊》, 40 (3), 1846-1877.[19] Li, R., Zhong, W. and Zhu, L. (2012). 通过距离相关学习进行特征筛选。《美国统计学会会刊》, 107, 1129-1139.

[20] Mai, Q. 和 Zou, H. (2013). 用于高维二分类方差筛选的柯尔莫哥洛夫筛选器。Biometrika, 100, 229-234。[21] Ni, L. 和 Fang, F. (2016). 基于熵的无模型特征筛选，用于超高维多分类。Journal of Nonparametric Statistics, 28 (3), 515-530。[22] Niu, Y., Zhang, R., Liu, J. 和 Li, H. (2020). 线性模型下超高维特征的分组筛选。Stat. Theor. Relat. Field., 4 (1), 43-54。[23] Pan, R., Wang, H. 和 Li, R. (2013). 关于类别数不断增加的超高维线性判别分析问题。未发表手稿。[24] Parker, J.S. 等 (2009). 基于内在亚型的乳腺癌监督风险预测变量。Journal of Clinical Oncology, 27 (8), 1160-1167。[25] Qiu, D. 和 Ahn, J. (2020). 线性模型超高维数据的分组变量筛选。Comput. Stat. Dta Anal., 178, 1-15。[26] Sang, Y. 和 Dang, X. (2023). 高维下基尼相关的渐近正态性及其在K样本问题中的应用。https://doi.org/10.48550/arXiv.2203.00081。[27] Serfling, R. (1980). 《数理统计近似定理》。Wiley，纽约。[28] Shao, X. 和 Zhang, J. (2014). 鞅差相关及其在高维变量筛选中的应用。《美国统计学会杂志》，109, 1302-1318。[29] Song, W. 和 Xie, J. (2019). 基于F统计量的分组特征筛选。Commun. Stat. Simul. Comput., 48, 1921-1931。[30] Sz´ekely, G. J., Rizzo, M. L. 和 Bakirov, N. (2007). 通过距离相关测量和检验依赖关系。《统计年刊》，35 (6), 2769-2794。[31] Vershynin, R. (2018). 《高维概率：数据科学应用导论》，剑桥大学出版社，剑桥，英国。[32] Wang, Z., Deng, G. 和 Yu, J. (2022). 基于信息增益率的超高维数据分组特征筛选，《数学杂志》，第2022卷，文章编号1600986。[33] Wang, Z., Deng, G. 和 Xu, H. (2023). 基于基尼不纯度的超高维多分类分组特征筛选，《AIMS数学》，8 (22), 4342-4362。[34] Zhang, S., Dang, X., Nguyen, D., Wilkins 和 Chen, Y. (2021). 使用基尼距离统计量估计特征与标签的依赖关系，IEEE Trans. Pattern Anal. Mach. Intell., 43 (6), 1947-1963。[35]Zhu, L., Li, L., Li, R. 和 Zhu, L. (2011). 超高维数据的无模型特征筛选，《美国统计学会杂志》，106, 1464-1475。