arXiv:2603.02160v1 [stat.ME] 2 Mar 2026

集合式分层变量选择与用于错误发现率控制的广义线性逐步向上程序

Sarah Organ*、Toby Kenney*、Hong Gu* 同等共同作者 达尔豪斯大学数学与统计学系，加拿大新斯科舍省哈利法克斯

2026年3月3日

摘要

当预测变量相关时，在变量选择中控制错误发现率（FDR）变得具有挑战性，因为现有方法往往会排除相关组的所有成员，从而导致预测表现不佳。我们提出一种新的集合式变量选择框架，该框架识别潜在预测变量的簇，而不是强制选择单个变量。通过允许所选集合中的任一成员作为替代预测变量，我们的方法在保持严格错误发现率控制的同时支持较强的预测性能。我们基于相关性对预测变量进行层次聚类来构建集合，然后检验每个集合是否包含任何非零效应。类似的聚类和集合式选择已应用于家族错误率（FWER）控制情形，但以往研究未能克服将其扩展到错误发现率控制框架所固有的挑战。

为控制错误发现率，我们对线性逐步递增程序进行了重要的推广，将 Benjamini–Hochberg 和 Benjamini–Yekutieli 方法扩展至适用于这些复合假设之间的逻辑依赖关系。我们证明，这些程序能够在名义水平上控制错误发现率，并强调了其更广泛的适用性。模拟研究和真实数据分析表明，与现有方法相比，我们的方法在保持错误发现率控制的同时，具有更高的检验效能，从而得到信息更丰富的变量选择结果和效果更好的预测模型。

关键词

错误发现率控制、多重假设检验、线性逐步递增程序、层次聚类、变量选择

1 引言

1.1 背景

变量选择是统计学中一个广泛且核心的问题，对于科学解释和构建有效的预测模型都至关重要。鉴于其重要性，人们已经开发出种类繁多的方法。这些方法大致可分为两类：旨在优化预测的方法，以及旨在控制错误发现率（FDR）的方法。

在以预测为导向的方法中，收缩技术占据主导地位，其中最著名的是 LASSO（Tibshirani，1996年）及其众多扩展，这些扩展通过纳入额外结构（Yuan 和 Lin，2006年）、改善理论性质（Zou，2006年），或增强稀疏性和稳定性（Meinshausen

和 Bühlmann，2010年；Liu 等人，2021年）。其他重要方法基于搜索策略，例如最佳子集选择（Bertsimas 等人，2016年；Zhang 和 Cavenaugh，2016年）。

一些控制错误发现率（FDR）或家族错误率（FWER）的方法，依赖于诸如 BH（Benjamini 和 Hochberg，1995年）和 BY（Benjamini 和 Yekutieli，2001年）等假设检验程序。较新的方法包括结合数据分割（DS）或多重数据拆分（MDS）的镜像方法（Xing 等人，2021年；Dai 等人，2023年），以及 model-X knockoff 框架（Barber 和 Candes，2015 年）。这些方法通过生成虚假的零假设统计量，并将其与观测检验统计量进行比较，从而识别显著预测变量。虽然这使得这些方法能够应用于高维情形，但也将其限制在特定的模型族中，并且要求了解所有特征的联合分布——而这类信息通常无法获得，也难以估计。该分布的错误设定可能导致检验效能大幅损失（Dai 等人，2022年），或使错误发现率膨胀（Barber 等人，2020年）。

乍看之下，以预测为重点的方法与旨在控制错误发现的方法之间的区别似乎令人费解：真实变量往往是强预测变量，而纳入虚假变量通常会损害预测性能。因此，我们可能会认为，旨在选出真实预测变量并避免选入虚假变量的程序，也能带来出色的预测性能。然而，当两个或更多预测变量高度相关时，即使我们确信其中至少有一个携带真实信号，也很难确定究竟是哪一个。控制错误发现率的方法通常会因此选择两个变量都不选。反复有研究记录了这种检验效能损失，以及在存在代理变量时错误发现率可能膨胀的现象；相关研究既涉及基于 p-值的线性逐步递增程序（Schwartzman 和 Lin，2011年），也涉及较新的基于镜像的方法（Dai 等人，2022年；Rastaghi 等人，2024年）。

为弥补这一局限，我们考虑一种不同的范式：除了选出我们确信是真实预测变量的单个变量外，我们还选出我们确信至少包含一个真实预测变量的预测变量集合，即使无法确定集合中的哪个预测变量是真实预测变量。就预测而言，这类集合中的任何预测变量都可以作为有效预测变量。由于我们并不声称能够识别集合中的真实变量，因此只有当所选集合中的变量无一是真实预测变量时，才会产生错误发现。这种表述既保留了错误发现率控制，也提供了有关真实信号位置的有用信息。

这一范式此前已应用于控制 FWER 的变量选择。具体而言，Meinshausen（2008）提出了一种控制 FWER 的程序，该程序对基于相关性的层次聚类所形成的预测变量簇进行检验，从而在无法精确确定真实预测变量时，仍可选择预测变量簇。检验按簇的规模从大到小依次进行；只有在所有上位簇都已被拒绝后，才会检验某个簇（该簇也可能只包含一个预测变量）。Goeman 和 Finos（2012）对这一方法进行了改进和推广。Goeman 和 Finos（2012）的核心思想可视为将 Bonferroni-Holm 方法（Holm，1979）扩展到更一般的“继承原则”：Bonferroni-Holm 方法会随每次拒绝动态调整拒绝阈值，而继承原则则依据假设背后的结构进行调整。

这项研究与用于控制 FWER 的闭合检验程序（Marcus 等人，1976）有关。这些程序会检验零假设的所有交集，从最强的零假设开始；只有在所有更强的假设都已被拒绝后，才会检验较弱的假设。这等价于将每个假设的 p值替换为所有更强假设中的最大 p值。Hommel（1986）指出，若拒绝所有调整后的 p值低于q的交集假设，则可将 FWER 控制在 q水平。

应用于变量选择时，一组预测变量的原假设是其各个预测变量原假设的交集。在闭合检验中，只有当所有超集对应的假设都能在阈值 q 下被拒绝时，才会选择该预测变量集合。然而，对所有可能的预测变量子集分别进行检验，很快就会变得计算上不可行。因此，闭合检验通常仅限于以下情形：直接计算交集假设的 p 值

而不是根据各个假设的 p 值计算；这并不适用于变量选择情境 t.

虽然与 FWER 控制相比，变量选择通常更倾向于较宽松的 FDR 控制，但文献中尚未出现将 FDR 控制用于集合式变量选择的研究。我们在本文中克服的主要技术障碍，是如何恰当地定义发现的数量。这一点至关重要，因为恰当的发现计数方式是 FDR 控制定义的一部分；相比之下，FWER 控制是根据一个二元事件（是否存在任何错误发现）来定义的。为我们的研究情境定义了恰当的 FDR 后，我们还需要构建能够处理这一 FDR 的广义线性逐步上调程序。

据我们所知，文献中尚无适用于集合式变量选择的 FDR 控制程序，不过，FDR 控制领域的一些进展确实为此迈出了重要一步。尤其是，Yekutieli 等人（2006年）针对先验排列在层次树中的假设给出了三种 FDR 定义。该论文提出的方法既受限于独立 p 值这一假设，也受限于基于最坏情形树结构的截断值；后者往往导致过于保守的 FDR 控制，即使在假设独立的情况下也是如此。尽管如此，“外部节点 FDR”的定义仍是重要的一步。其核心思想是：由于拒绝较弱的假设会使拒绝较强假设变得多余，我们应当只将被拒绝的最弱零假设计为发现。然而，外部节点 FDR 对所有假设赋予相同权重，而不考虑其层级，因此，提高已有发现的分辨率不计作一次发现。为了构建适用于变量集选择的 FDR 定义，需要将这种“外部节点 FDR”与加权 FDR（Benjamini 和 Hochberg，1997年；Basu 等人，2018年）结合起来；加权 FDR 允许为更重要的假设赋予更高权重。其他针对假设树的 FDR 定义，例如树的逐步 FDR（Bogomolov 等人，2021年），或针对一般偏序集的定义（Ramdas 等人，2019a），都不适合作为变量选择的 FDR 定义。

该领域另一个重要的研究方向是对多种不同的 FDR 进行控制，例如 p-filter 方法（Ramdas 等人，2019b）可同时控制假设的多个划分上的错误发现率。该方法的常见应用情形是，这些划分对应于层次树中的分辨率层级（例如分类学层级）。这一方法已被推广到 knockoff 方法（Katsevich 和 Sabatti，2019年；Sesia 等人，2020年）。尽管这些方法适用于某些问题，但利用层次聚类得到的集合式假设并不能自然地落入各个层级。

上述针对层次树的错误发现率和 FWER 控制方法有一个共同特征：仅当所有更强的零假设都具有较小的 p 值时，它们才会拒绝假设。这是合理的，因为如果我们拒绝某个假设，就隐含地拒绝了任何更强的假设。然而，相应的检验统计量未必遵循这一模式，因此这些方法可能无法拒绝较低层级的零假设，尽管有强有力的证据反对它们，从而导致较低的检验效能。因此，我们将同时开发带有和不带有这一要求的方法。

1.2 示例

我们通过一个包含 $n = 300$ 个样本和 $p = 50$ 个预测变量的小型模拟线性回归数据集，展示 FDR 控制的变量选择的缺陷，以及选择变量集可能带来的好处；其中 $X \sim N(0, \Sigma)$，协方差矩阵 Σ 定义如下：

$$\Sigma_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0.9 & \text{if } \{i, j\} = \{1, 2\} \\ 0 & \text{otherwise} \end{cases}$$

变量 $X_2 - X_6$ 是真实变量，其系数为 $\beta_i \sim N(0, 1)$。

表 1 展示了在该模拟数据集上、FDR 控制水平为 $q = 0.05$ 的各种变量选择方法的结果，包括本文稍后提出的方法，以及选定模型在规模为 $n = 3000$ 的测试数据集上的预测均方误差。现有的 FDR 控制方法均无法

表 1：示例数据上不同错误发现率控制变量选择方法的性能

方法	所选变量	FDR	MSE
BH	X_3, X_4, X_5, X_6	0%	1.93
镜像方法 DS	None	0%	NA
镜像方法 MDS	X_3, X_4, X_5, X_6	0%	1.93
Knockoff	None	0%	NA
集合式选择	$\{X_1, X_2\}$, X_3, X_4, X_5, X_6	0%	1.72
套索回归	X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_{12}, X_{13}, X_{14}, X_{15}, X_{16}, X_{17}	61.5%	1.69

由于与 X_1的高度相关性而选出 X_2。LASSO方法不进行错误发现率控制，它选出了所有真实变量，因此尽管选出了若干虚假变量，仍取得了良好的预测效果。由于无法区分 X_1 和 X_2 中哪一个是真实变量，我们的方法选择了集合$\{X_1, X_2\}$，这意味着其中之一是真实变量。这使得该方法能够维持对错误发现率的控制，同时仍取得良好的预测效果。尽管该示例是为说明这一点而人为构造的，但预测变量之间这种程度的相关性在真实数据中很常见。本研究稍后给出的微生物组数据示例即属此类。其他示例包括遗传数据，其中带有错误发现率控制的生物标志物筛选往往是关键步骤，而当预测变量不仅相关而且呈非高斯分布时，这一步骤尤其具有挑战性。

1.3 预备知识

现在我们来探讨该方法的技术细节——例如：应考虑哪些预测变量集合、应如何选择这些集合，以及在此情境下如何定义错误发现率?

我们首先明确问题背景的具体设定。考虑一个一般的预测建模问题。我们有一组预测变量 $\{X_1, \ldots, X_p\}$，以及一个响应变量 Y。我们的目标是预测 $Y|X_1$、.....、X_p 的条件分布。一些文献将真实预测变量定义为：在给定所有其他预测变量的条件下，与 Y 不条件独立的预测变量。然而，当一个预测变量是其他变量的函数时（例如，在线性回归中包含非线性项或交互项的情形），这一定义并不成立。因此，我们给出一个更严谨、适用于此类情形的定义。设 $\mathcal{D}$ 为可能的条件分布族——通常 $\mathcal{D}$ 是一个参数族，例如所有正态分布构成的族。我们的模型空间由一列函数空间 $(\mathcal{M}_n)_{n\in\mathbb{N}}$ 组成，其中每个 $\mathcal{M}_n$ 都是从 $\mathbb{R}^n$ 到 $\mathcal{D}$ 的函数集合。例如，对于高斯线性回归，$\mathcal{D}$ 是均值为 μ、方差为 σ^2 的正态分布族，而 $\mathcal{M}_n$ 由形式为 $\mu = \beta_0 + \sum_{i=1}^n X_i\beta_i$ 的函数 $f\colon \mathbb{R}^n \longrightarrow \mathcal{D}$ 构成，且 σ^2 为常数。通常，我们假设对于任意 $J \subseteq \{1, \ldots, n\}$ 和任意 $f \in \mathcal{M}_{|J|}$，由 $f_J(x_1、\ldots\ldots、x_n) = f((x_i)_{i\in J})$ 定义的函数 f_J 都属于 $\mathcal{M}_n$，即模型是嵌套的，但这一假设并非关键。我们假设真实条件分布由某个唯一的最小 $T \subseteq \{1, \ldots, p\}$ 和 $f \in \mathcal{M}_{|T|}$ 对应的函数 $(Y|X_1 = x_1、\ldots\ldots、X_p = x_p) = f((x_i)_{i\in T})$ 给出。若预测变量 X_i 满足 $i \in T$，则它是真实变量；若满足 $i \notin T$，则它是零变量。对于任意子集 $C \subseteq \{1, \ldots, p\}$，我们可以进行假设检验，例如似然比检验，来检验假设 $H_C\colon C \cap T = \emptyset$。

由于这些问题出现在高度相关的变量中，我们根据预测变量之间的相关性进行层次聚类。令 $\mathcal{C}$ 表示通过这种层次聚类识别出的所有簇或 $\{1, \ldots, p\}$的子集，包括单元素集合。对于每个簇 $C \in \mathcal{C}$，我们使用响应变量 Y检验原假设 H_C。如果检验统计量的分布基于给定 X时 Y 的条件分布，那么我们根据预测变量 X 之间的相关性来选择要检验的子集这一事实，不会导致所得 p值产生选择偏差；这与基于响应变量 Y的筛选方法不同。

接下来，我们将探讨如何在当前情境下定义错误发现率，并据此构建控制错误发现率的方法。

首先回顾 BH 和 BY 方法所采用的标准线性逐步递增程序。对于该程序，给定一组零假设 $\{H_1, \ldots, H_m\}$ 及其对应的 p 值 $P_1, \ldots, P_m$，使得真实零假设对应的 p 值边际上服从均匀分布。错误发现率控制方法是一个函数，它针对每个 p 值向量，选出一组 R 假设予以拒绝。若 I_N 是真实零假设的集合，则错误发现率定义为

$$\text{FDR} = \mathbb{E}\left(\frac{|R \cap I_N|}{|R|}\right)$$

线性逐步上升法最初表述为对 p值 $P_{(1)} \leqslant P_{(2)} \leqslant \cdots \leqslant P_{(m)}$进行排序，然后拒绝最小的 k p值，其中 k 是满足 $P_{(k)} \leqslant \frac{k}{\alpha}$ 的最大秩，其中 $\alpha = \frac{m}{q}$ 对应 BH方法，而 $\alpha = \frac{m}{q}\sum_{i=1}^{m}\frac{1}{i}$ 对应 BY方法。该程序可以改写为："拒绝所有小于或等于 p的 c值，其中 c 是满足 $|\{i|P_i \leqslant c\}| \geqslant \alpha c$的最大截断值。" Finner等人（2009）采用这一表述，以非线性阈值函数替代线性αc，这更便于将其推广到我们的情境中。

错误发现率可以理解为

$$\mathbb{E}\left(\frac{\text{number of false discoveries}}{\text{total number of discoveries}}\right)$$

因此，要在我们的情境中定义它，只需确定如何计算发现的数量。问题在于，例如，如果我们检验以下零假设：

$H_{\{1\}}$：变量 X_1 不是真实预测变量。

$H_{\{2\}}$：变量 X_2 不是真实预测变量。

$H_{\{1,2\}}$：变量 X_1 和 X_2 都不是真实预测变量。

我们不能理所当然地将拒绝这三个假设都计为三项发现，因为拒绝这三个假设相当于只选择了两个变量。此外，由于 $H_{\{1,2\}}$是一个更强的原假设，因此更容易被拒绝，我们应认为拒绝它所代表的发现少于拒绝 $H_{\{1\}}$ 或 $H_{\{2\}}$所代表的发现。

对于集合式变量选择问题，我们采用以下方法来计算发现。首先，根据 C中的变量数量，为每个假设 H_C赋予权重 $\phi_C \geqslant 0$ 。本文中，我们使用 $\phi_C = \frac{1}{|C|}$，因此选择包含 k 个变量的集合相当于 $\frac{1}{k}$ 项发现，不过我们的方法也适用于更一般的 ϕ_C。这样做有一个优点：如果 C 是单元素集合，则 $\phi_C = 1$，因此选择一个变量计为一项发现。

如果两个被拒绝的零假设之间存在蕴含关系，我们只将较弱的被拒绝假设计为一项发现，因为拒绝较弱的假设后，再拒绝较强的假设就不再算作一项发现。（对于未加权假设，Yekutieli 等人（2006）将此称为"外节点错误发现率"。）因此，我们的（错误）发现数为

$$\sigma(\mathcal{S}) = \sum_{C \in \text{Min}(\mathcal{S})} \phi_C$$

其中，$\mathcal{S} \subseteq \mathcal{C}$ 是（错误）被拒绝簇的集合，Min($\mathcal{S}$) 是 $\mathcal{S}$ 中极小簇的集合（即不包含 $\mathcal{S}$ 中任何其他簇的簇）。由于聚类$\mathcal{C}$的层次性，对于任意 $\mathcal{S} \subseteq \mathcal{C}$，$\mathcal{S}$ 中的极小簇彼此不相交。我们将把函数 σ 称为"规模函数"，它用于统计发现的数量。

另一个需要考虑的问题是，由于假设之间存在蕴含关系，拒绝一个较弱的假设并不合理 H_{C_1} 而不同时拒绝一个更强的假设 H_{C_2}，后者蕴含 H_{C_1}。因此，我们对可以拒绝的假设组合施加约束。也就是说，我们事先给定一个$\mathcal{R}$ 可被拒绝的假设集的集合。集合 $\mathcal{R}$ 对交运算封闭，因此，对于任意假设集 J，都存在一个最小的$J \in \mathcal{R}$使得 $J \subseteq J$，称为 J的"闭包"。我们提出两种程序，以确保被拒绝假设的集合属于 $\mathcal{R}$。这两种程序都需要扩展现有的错误发现率控制方法理论。

这一通用的错误发现率控制框架涵盖了原始的 BH 和 BY 方法，以及加权 BH 方法（见第 2.1 节）。这种高度一般化的框架具有广泛的潜在应用。例如，与我们在此提出的规模函数类似的通用规模函数，可能比 Yekutieli 等人（2006）的层级错误发现率控制方法中隐含使用的规模函数更适用于微生物组数据分析和全基因组关联研究。

1.4 贡献与概述

本文有两项主要贡献。第一，我们对多重假设检验的错误发现率控制方法进行了广泛推广，使其涵盖以下情形："发现数量"由一般的"规模函数"衡量，且假设之间的关系会对可拒绝的假设集合施加限制。该框架不仅将原始的 BH 和 BY 程序以及加权 BH 程序（Benjamini 和 Hochberg，1997）作为特例纳入其中，还为我们提出的集合式变量选择及错误发现率控制问题提供了解决方案，并有望应用于一系列其他多重假设检验问题。

第二项主要贡献是一种新颖的集合式变量选择框架，它克服了在高相关性情境下兼顾错误发现率控制与足够检验效能的挑战。以往的变量选择研究分为两类：一类旨在控制错误发现率，另一类旨在构建良好的预测模型。我们的集合式变量选择方法有望同时实现这两个目标。

本文其余部分的结构如下。第 2 节中，我们针对一般规模函数，构建用于控制错误发现率的广义线性逐步提升框架。随后，我们在关于p值之间依赖关系的不同假设下，推导出用于控制错误发现率的适当阈值斜率的理论结果。第 3 节中，我们采用广义线性逐步向上程序，使集合层级错误发现率（SHRED）方法能够以控制错误发现率的方式选择变量集。此方法适用于任何能够为任意预测变量集提供有效 p值的变量选择问题。第 4 节中，我们通过一系列模拟研究，将 SHRED 方法与其他最先进的变量选择方法进行比较。最后，在第 5 节中，我们将这些变量选择方法应用于一个真实的微生物组数据集，该数据集测量了污水处理厂中的分类单元丰度（Dueholm 等，2022）。第 6 节总结全文。

2 广义线性逐步上升程序（GLSUP）

2.1 一般程序

对于我们的广义线性逐步向上程序，需要以下输入：

定义 1 应用 GLSUP 方法所需的输入包括以下内容：

- 我们有一组零假设 $H_1, \ldots, H_m$，以及与之对应的保守 p 值 $P_1, \ldots, P_m$，它们是随机变量。（保守的 p 值是指，在零假设 H_i 下，P_i 服从超均匀分布。）我们用 $\mathcal{P}$ 表示 $\{1, \ldots, m\}$ 的幂集，即 $\mathcal{P} = \{J \mid J \subseteq \{1, \ldots, m\}\}$。

- 我们给定了一个集合族 $\mathcal{R} \subseteq \mathcal{P}$，其中包含 $\{1,\ldots,m\}$的“可拒绝”子集，即可以被一致拒绝的假设集合。我们假设 $\mathcal{R}$ 在任意交下闭合。这意味着存在一个相应的闭包算子，它会将任意 $J \subseteq \{1,\ldots,m\}$ 映射到包含$J \in \mathcal{R}$ 的最小集合 J。

- 我们有一个规模函数 σ： $\mathcal{R} \longrightarrow \mathbb{R}_{\geqslant 0}$，它关于集合包含关系单调递增（不一定严格递增）$ct$ 到集合包含。

在此框架中，斜率为 α 的广义线性逐步上升法如下：

过程 1（GLSUP）

- 对于任意截断值 $c \in [0,1]$，我们定义 $I_c = \{i|P_i \leqslant c\}$。
- 定义 $c_{\text{最大值}}$ ＝上确界 $\left\{c \in [0,1] \middle| \sigma(I_c) \geqslant \alpha c\right\}$

- 拒绝$I_{c_{max}}$中的所有假设。

这被称为线性步升程序，因为阈值 αc 是截断值c的线性函数。我们也可以仿照 Finner等人（2009）设计一个具有非线性阈值的程序；不过，这超出了本文的研究范围。线性阈值的斜率 α 是为控制错误发现率而选取的，并取决于我们对零假设 p值的假设。我们在第 2.3 节和第 3 节中推导了 α 在各种常见变量选择情形下的取值。

除上述记号外，我们用 $I_N \subseteq \{1,\ldots,m\}$ 表示真实零假设的指标，用 $I_A = \{1,\ldots,m\} \setminus I_N$ 表示虚假原假设的指标。广义错误发现率（gFDR）和广义检验功效（gPower）定义如下：

$$\text{gFDR} = \mathbb{E}\left(\frac{\sigma\left(\overline{I_{c_{\max}}} \cap I_N\right)}{\sigma\left(\overline{I_{c_{\max}}}\right)}\right)$$

and,

$$\text{gPower} = \mathbb{E}\left(\frac{\sigma\left(\overline{I_{c_{\max}}} \cap I_A\right)}{\sigma\left(\overline{I_A}\right)}\right)$$

现有的线性步升错误发现率控制程序是 GLSUP 的特例，具有特定的规模函数和经过适当选择的 α 值。当 $\mathcal{R} = \mathcal{P}$ 和 $\sigma(J) = |J|$时，令 $\alpha = \frac{m}{q}$ 即可得到 BH 程序，而令 $\alpha = \frac{m\sum_{j=1}^{m}\frac{1}{j}}{q}$ 即可得到 BY程序，其中 q 是期望的错误发现率控制水平。加权 BH（wBH）程序（Benjamini 和 Hochberg，1997）也是 GLSUP 的特例，其 $\sigma(J) = \sum_{i\in J} w_i$ 且 $\alpha = \frac{\sum_{i=1}^{m} w_i}{q}$。显然，若 $\sigma(J) = |J|$ 且 $\mathcal{R} = \mathcal{P}$，则 BH 和 BY 程序下的 gFDR 与 gPower 分别等同于标准错误发现率和检验效能；而若$\sigma(J) = \sum_{i\in J} w_i$ 且 $\mathcal{R} = \mathcal{P}$，则 gFDR 和 gPower 分别等同于 Benjamini 和 Hochberg（1997）提出的加权 FDR 和加权功效。

2.2 修正的 GLSUP 程序

在我们关注的变量选择问题中，假设的可拒绝集合基于蕴含。逆蕴含 $H_i \Leftarrow H_j$ 在集合 $\{H_1,\ldots,H_m\}$ 上定义了一个偏序。可拒绝集合是该偏序下的上闭集。在这种情况下，我们可以修改 GLSUP 程序，使小于截断值 c 的 p 值集合始终是可拒绝的。

程序 2（mGLSUP）

- *Define* $P'_1, \ldots, P'_m$ *by* $P'_i = \sup\{P_j, j \in \{1, \ldots, m\} | (H_j \implies H_i)\}$
- *For any cut-off* $c \in [0,1]$, *define* $I_c = \{i | P'_i \leqslant c\}$.
- *Define* $c_{max} = \sup\left\{c \in [0,1] \middle| \sigma\left(\overline{I_c}\right) \geqslant \alpha c\right\}$
- *Reject all hypotheses in* $\overline{I_{c_{max}}}$.

请注意，由于 P'_i 的定义方式，对于所有 $c \in [0,1]$，我们都有 $I_c = I_c$。

该程序与文献中的其他若干程序效果相似（Meinshausen，2008年；Yekutieli 等人，2006年；Bogomolov 等人，2021年）：只有在其父假设被拒绝时，某个假设才可能被拒绝。这种方法的一个优点是，对于我们的规模函数 σ（即极小元的权重之和），以及我们的变量选择问题（其中偏序具有二叉树结构），我们可以将 σ 表示为成对权重 $\sigma(A) = \sum_{i,\ j \in A} w_{ij}$的和，其中 $w_{ij} \geqslant 0$满足某些条件。该结论在附录 ?? 中得到证明。借助这一点，我们可以在一定条件下证明广义错误发现率控制，并采用更小的截断值 α；因此，在某些情况下，修正的 GLSUP 能够获得比 GLSUP 更高的广义检验功效。

2.3 广义线性逐步向上程序的理论

本节中，我们将证明，只要适当选择 α，我们的 GLSUP 和 mGLSUP 就能控制广义错误发现率。我们首先证明，在不对 p 值作任何假设的情况下，广义错误发现率仍受控。

定理1 对于定义 1 中的输入，假设对于任意 $J \subseteq \{1, \ldots, m\}$，$\sigma\left(J\right) \leqslant \sum_{i \in J} w_i$，存在某些权重 w_1，……，w_m。那么，截断值为

$$\alpha = \frac{\left(\sum_{i \in I_N} w_i\right)\left(1 + \log(\sigma(\{1, \ldots, m\}))\right) - \sum_{i \in I_N} w_i \log(\sigma_i)}{q}$$

其中 $\sigma_i = \sigma(\{i\})$，可在水平 q控制广义错误发现率。

定理1的证明是对 Benjamini 和 Yekutieli（2001）中证明的直接推广，详见补充附录 ??。

注1 若假设之间不存在结构，因此 $\sigma(J) = |J|$ 为基数，且所有集合均可被拒绝（即 $\mathcal{R} = \mathcal{P}$），则令 $w_i = 1 = \sigma_i$，得到 $\alpha = \frac{m_0}{q}(1 + \log(m))$。由于 m_0 未知，将 m 用作上界是标准做法。于是，定理1给出 $\alpha = \frac{m}{q}(1 + \log(m))$。这与 BY 方法（Benjamini 和 Yekutieli，2001）所用的截断值 $\alpha = \frac{m}{q}\sum_{i=1}^{m} \frac{1}{i}$ 非常接近，但略高一些。造成这一差异的原因是，在 Benjamini 和 Yekutieli（2001）的原始问题中，规模函数是被拒绝假设的数量，因此必须为整数。

与原始 BH 方法一样，若假设 p值在真实零假设集合 I_N 上满足子集上的正回归依赖性（PRDS）条件（Benjamini 和 Yekutieli，2001年），我们便可以使用更小的截断值 α 来控制广义错误发现率（gFDR）（从而使检验不那么保守）：

定义 2（PRDS）一组 $[0,1]$值随机变量 $\mathbf{P} = (P_1, \ldots, P_m)$ 在子集 $I_N \subseteq \{1, \ldots, m\}$ 上满足 PRDS 条件，当且仅当对于任意 $i \in I_N$和任意上闭集 $U \subseteq [0,1]^{m-1}$，$P(\mathbf{P}_{\hat{i}} \in U | P_i = x)$ 是关于 x的单调递增函数，其中 $\mathbf{P}_{\hat{i}}$表示移除第 i个索引后的 P。

在此假设下，我们可以在较低截断值下证明广义错误发现率控制。

定理 2 假设保守的 p值 P_1、.....、 P_m 满足关于 I_N 的 PRDS 条件，并且规模函数 σ: $\mathcal{P} \longrightarrow \mathbb{R}_{\geqslant 0}$ 受权重和限制： $(\forall J \subseteq \{1,\ldots,m\})(\sigma\left(J\right) \leqslant \sum_{i\in J} w_i)$ 对于某些权重 $w_i \geqslant 0$。那么，截断值为 $\alpha = \frac{\sum_{i=1}^{m} w_i}{q}$ 的广义线性逐步上升法可在 q 水平控制广义错误发现率。

在 $\sigma(J) = |J|$ and $\mathcal{R} = \mathcal{P}$ 的情况下，这与经典 BH 方法中的 α 值相同。定理 2 的证明是对 Benjamini 和 Yekutieli（2001）证明的相当直接的扩展。在 $\sigma(J) = \sum_{i\in J} w_i$ 的情况下，这与加权 BH 方法（Benjamini 和 Hochberg，1997）中的界相同。最初，wFDR 控制仅在独立 p 值的情况下针对该方法得到证明，因此我们的定理改进了相关理论，即使对于现有方法也是如此。

对于我们修正的 GLSUP，无法将 $\sigma(J)$ 表示为权重和 $\sum_{i\in J} w_i$，但可以将其表示为成对权重之和，$\sum_{i,\ j\in J} w_{ij}$。为了将我们关于较小 α 的结果扩展到这种情况，我们将 PRDS 条件推广为成对 PRDS 条件：

定义 3（PPRDS）一组 $[0,1]$ 值随机变量 $\mathbf{P} = (P_1,\ldots,P_m)$ 在子集 $J \subseteq \{1,\ldots,m\}$ 上满足 PPRDS 性质，当且仅当其在 J 上满足 PRDS 条件，并且对于任意 $i \neq j \in J$ 和任意上集 $U \subseteq [0,1]^{m-2}$，都有

$$P(\mathbf{P}_{\hat{i},\hat{j}} \in U | P_i = x_i, P_j = x_j)$$

是关于 x_i 和 x_j 的单调递增函数，其中 $\mathbf{P}_{\hat{i},\ \hat{j}}$ 是移除 P_i 和 P_j 后由 p 值构成的向量 $\mathbf{P}$。

定理 3 对于定义 1 中的输入，假设保守的 p 值在 I_N 上满足成对正回归依赖于子集（PPRDS）条件，且规模函数由成对权重之和所界定，即 $(\forall J \subseteq \{1,\ldots,m\})(\sigma\left(J\right) \leqslant \sum_{i,\ j\in J} w_{ij})$，其中 $w_{ij} \geqslant 0$。那么，阈值为 $\alpha = \frac{\sum_{i,\ j\in I_N} w_{ij}}{q}$ 的广义线性逐步上升法可将 gFDR 控制在水平 q。

当 PPRDS 成立时，定理 2 是定理 3 的一个特例，可通过令 $w_{ij} = 0$ 对所有$i \neq j$成立而得到。然而，PPRDS 条件强于 PRDS 条件。事实上，定理 2 和定理 3 都可以视为一个更一般定理的特例，其中 σ 满足容斥原理的不等式形式：对任意 J_1，.....，$J_n \subseteq \{1,\ldots,m\}$，

$$\sigma(J_1 \cup \cdots \cup J_n) \geqslant \sum_{S\subseteq\{1,\ldots,n\}, S\neq\emptyset} (-1)^{|S|-1} \sigma(\cap_{k\in S} J_k)$$

在这种情况下，我们在补充附录 ?? 中证明，σ 可以表示为和 $\sigma(J) = \sum_{J'\subseteq J} w_{J'}$，其中权重为某些 $w_{J'} \geqslant 0$。为陈述这一更一般的定理，我们首先将 PRDS 推广为集合式 PRDS 条件：

定义 4（SPRDS）一组随机变量 $\mathbf{P} = (P_1,\ldots,P_m) \in [0,1]^m$ 对$\{1,\ldots,m\}$ 的子集构成的集合 $\mathcal{J} \subseteq \mathcal{P}$ 满足子集上的集合式正回归依赖（SPRDS）条件，如果对任意 $C \in \mathcal{J}$、任意满足 $(\forall i \in C)(x_i \leqslant y_i)$ 的 $x, y \in [0,1]^{|C|}$ 以及任意向上闭$U \subseteq [0,1]^{m-|C|}$，则有

$$P(\mathbf{P}_{\hat{C}} \in U | \mathbf{P}_C = x) \leqslant P(\mathbf{P}_{\hat{C}} \in U | \mathbf{P}_C = y)$$

其中，$\mathbf{P}_{\hat{C}}$ 表示移除指标 C 后的$\mathbf{P}$，$\mathbf{P}_C$ 表示将$\mathbf{P}$限制在 C上的结果。

不难看出，PRDS 和 PPRDS 是 SPRDS 条件的特例：分别令$\mathcal{S}$ 为单元素集，或单元素集与二元组的集合。

定理 4 设 $H_1, \ldots, H_m$ 为一组零假设。设 $I_N \subseteq \{1, \ldots, m\}$ 为真实零假设集合。设规模函数有界于 $\sigma(J) \leqslant \sum_{J' \subseteq J} w_{J'}$，其中 $w_{J'} \geqslant 0$ 是 $\{1, \ldots, m\}$的子集上的权重。此外，假设保守的 p-值 $\mathbf{P}$ 满足 I_N 上权重非零的子集集合的 SPRDS 条件 $w_J > 0$。那么，阈值为$\alpha = \frac{W}{q}$ 的广义线性逐步上升法（其中 $W = \sum_{J \subseteq \{1,\ldots,m\}} w_J$）可将 gFDR 控制在水平 q。

当只有 $\{1, \ldots, m\}$ 的单元素子集具有非零权重时，这就得到定理 2；当只有大小至多为 2 的子集具有非零权重时，则简化为定理 3。定理 4 的证明是对 Benjamini 和 Yekutieli（2001）所给证明的推广，见补充附录??。

3 用于错误发现率控制的 SHRED 集合式变量选择程序

回顾我们的变量选择问题。我们有一组预测变量 $\{X_1, \ldots, X_p\}$，以及一个响应变量 Y。我们的模型空间由一族可能的条件分布 $\mathcal{D}$ 和一列（通常是嵌套的）函数空间 $(\mathcal{M}_n)_{n\in\mathbb{N}}$ 构成，其中每个 $\mathcal{M}_n$ 都是函数 $f\colon \mathbb{R}^n \longrightarrow \mathcal{D}$ 的集合。我们假设，对于某个唯一的最小 $T \subseteq \{1, \ldots, p\}$ 和 $f \in \mathcal{M}_{|T|}$，真实条件分布由 $(Y|X_1 = x_1, \ldots, X_p = x_p) = f((x_i)_{i\in T})$ 给出。若预测因子 X_i 满足 $i \in T$，则它是真实变量；若满足 $i \notin T$，则它是零变量。对于任意子集 $C \subseteq \{1, \ldots, p\}$，都有一个针对原假设$H_C$: $C \cap T = \emptyset$ 的假设检验。我们的目标是以控制广义错误发现率的方式，从 $\{1, \ldots, p\}$ 中选出一个子集集合。

3.1 SHRED 程序

我们的集合层级错误发现率（SHRED）方法如下:

（1）对所有预测变量构成的集合 $\{X_1, \ldots, X_p\}$ 进行基于相关性的层次聚类。令 $\mathcal{C}$ 表示该方法识别出的所有簇的集合。（2）使用 X 中的所有变量拟合预测 Y 的完整模型。（3）对于每个簇 $C_i \in \mathcal{C}$，移除 C_i 中的预测变量后重新拟合模型，并据此获得一个 p 值，用于检验原假设 H_{C_i}，即 C_i 中的所有预测变量都是零变量。我们使用似然比检验，或在适当情况下使用 F 检验，将该简化模型与完整模型进行比较。不过，也可以使用任何有效的假设检验。（4）在假设集合上定义一个“规模函数”: $\sigma(J) = \sum_{C_i \in \mathrm{Min}(J)} \phi_i$，其中 $\mathrm{Min}(J) = \{C_i \in J | (\forall C_j \in J)(C_j \subseteq C_i \Rightarrow i = j)\}$ 是 J 中的最小子集集合。（5）对这些假设对应的 p 值应用 GLSUP 或 mGLSUP，使用第 3.2 节定义的 α 值之一，拒绝其中一组假设。（6）选出的集合是 GLSUP 或 mGLSUP 方法得到的最小被拒绝簇。

我们将采用层次聚类的 mGLSUP 方法称为“降序 E 值缩减的集合层级错误发现率”（ SHREDDER）。（“E 值”一词（Wang 和 Ramdas，2022年）是一个替代框架，与保守的 p值等价。）

3.2 阈值斜率

在本节中，我们应用定理1、定理2和定理3来选择截断值 α，以保证第1.3节中定义的规模函数 σ 满足广义错误发现率控制。回顾我们的情形：零假设 H_1、.....、H_m按逆蕴含关系构成偏序，并且 σ 由 $\sigma(J) = \sum_{i\in \mathrm{Min}(J)} \phi_i$ 定义，其中 $\phi_i \geqslant 0$ 是假设的权重，而 Min(*J*) 表示在 J 中按逆蕴含关系排列的极小假设集合。（这些假设对应于变量的极小子集。）权重 ϕ_i 的选择使得 σ成为递增集合函数。（对于变量选择而言，这仅意味着如果 $H_i \Leftarrow H_j$，那么 $\phi_i \geqslant \phi_j$，但该理论适用于更一般的蕴含结构。）令 $I_N \subseteq \{1,\ldots,m\}$ 表示真实零假设的指标集合。

如果我们不对零假设 p 值之间的依赖关系作任何假设，则由定理1，取 $w_i = \phi_i$，可证明：

推论 1 对于任意 $q > 0$，广义线性逐步向上程序满足

$$\alpha = \alpha_{SHRED} = \frac{\left(\sum_{i\in I_N} \phi_i\right)(1 + \log(p)) - \sum_{i\in I_N} \phi_i \log(\phi_i)}{q}$$

其中 $p = \sigma(\{1,\ldots,m\})$，可将 gFDR 控制在水平 q。

在实践中，$\sum_{i\in I_N} \phi_i$ 未知，因此我们使用其上界 $\sum_{i=1}^m \phi_i$。对于集合式变量选择，我们对所有单元素集设定 $\phi_i = 1$，使得 p 为预测变量总数。

如果假设 p 值满足 PRDS 条件，则可将上界 $\sigma(J) \leqslant \sum_{i\in J} \phi_i$ 与定理2结合使用，以获得更小的截断值来控制 gFDR：

推论 2 如果 p 值在真实零假设上满足 PRDS 条件，那么，广义线性逐步向上程序

$$\alpha = \alpha_{SHRED_{PRDS}} = \frac{\sum_{i=1}^m \phi_i}{q}$$

将广义错误发现率控制在 q 水平。

显然，如果为所有大小大于 1 的集合赋予权重 $\phi_i = 0$，并为所有单元素集赋予权重 $\phi_i = 1$，那么这就变成了 BH 程序。

对于 SHREDDER 方法，我们先回顾一些初等偏序理论：

定义 5 对于一个偏序集 $(X, \Leftarrow)$，若元素 $x \in X$ 覆盖元素 $y \in X$，则 $y \Leftarrow x$，并且对于任意满足 $y \Leftarrow z \Leftarrow x$ 的 $z \in X$，都有 $z = x$ 或 $z = y$。例如，如果 $(X, \Leftarrow)$ 具有树结构，那么一个节点只覆盖其子节点。

推论 3 设 $\Leftarrow$ 是假设上的偏序，并具有如下性质：任意 H_i 至多覆盖该偏序中的两个 H_j。如果由 $p'_i = \sup\{p_j | H_j \Longrightarrow H_i\}$ 给出的 p' 值 $p'_1,\ldots,p'_m$ 在真实零假设上满足 PPRDS 条件，那么带有 $\alpha_{SHREDDER} = {}^{\sigma}{}_{(1,\ldots,m)}{}^{q}$ 的 mGLSUP 程序将广义错误发现率控制在 q 水平。

C推论 3 在补充附录 ?? 中得到证明。其思路是证明，对于任意 $J \in \mathcal{R}$, $\sigma(J$) 是一个 s um 的成对权重，然后使用定理 3。

注 2 推论 1 和推论 2 在 $\sigma(J) \leqslant \sum_{i\in \mathrm{Min}(J)} \phi_i$ 时也成立——也就是说，只要规模函数具有这种形式的上界，这些方法就适用。对于推论 3，若 $\sigma(J) \leqslant \sum_{i\in \mathrm{Min}(J)} \phi_i$，则该推论在将 $\alpha_{SHREDDER}$ 替换为 $\sum_{i\in \mathrm{Min}(\{H_j | j=1,\ldots,m\})} \phi_i$ 后仍成立。

实践中，SHRED 使用 PRDS 截断值所达到的经验 gFDR 几乎总是远低于名义水平。因此，可以通过选择较小的 α 值，在维持可接受的 gFDR 的同时获得更高的 gPower。我们发现，启发式方法 $\alpha_{\text{SHRED}_h} = \frac{p}{q}$ 在实践中效果良好。这种方法使用与 BH方法 下选择单个变量时相同的 gFDR 截断值。虽然无法保证使用此截断值时能够控制 gFDR，但在我们的模拟中，使用该截断值时，gFDR 始终在 q 水平上得到经验控制。我们还进行了小规模模拟研究，比较经验 FDR 与多种理论 FDR 水平。结果见补充附录 **??**。

实现这些方法的 SHRED 软件包可在 CRAN 上获取。

4 模拟

4.1 模拟设计

我们通过高斯回归、逻辑回归和泊松回归（采用对数链接）的模拟，评估所提方法相对于 BH、BY、基于 MDS 的镜像方法、基于 DS 的镜像方法以及 knockoff 方法的性能。我们模拟如下形式的数据 $X \sim N(\mu, \Sigma)$, μ =(0, ..., 0)^T, diag(Σ) $= 1$，并采用三种不同的相关结构：

（1）共同成分相关：

$$\Sigma_{ij} = \begin{cases} 1 & \text{if } i = j \\ \rho & \text{otherwise} \end{cases}$$

（2）聚类相关：将变量随机划分为互不相交的簇 $C_1, \ldots, C_K$，簇大小分别为$|C_1| = 5, |C_2| = 10, |C_3| = 15, \ldots$，其中 C_K包含所有剩余变量，因此，当 $p = 300$,$K = 11$和 $|C_K| = 25$时，而当 $p = 200$, $K = 9$ 和 $|C_K| = 20$时。

$$\Sigma_{ij} = \begin{cases} 1 & \text{if } i = j \\ \rho_k & \text{if } i \neq j \in C_k \\ 0 & \text{otherwise} \end{cases}$$

其中每个 ρ_k 都独立地从某个范围内的均匀分布中抽取。

（3）自回归（一阶自回归）相关性：

$$\Sigma_{ij} = \rho^{|i-j|}$$

对于每种设置，高斯回归和泊松回归使用 100 次模拟重复，逻辑回归则使用 200 次，这是因为逻辑回归往往会产生更宽的置信区间。

使用 simstudy v0.8.1 R 软件包（Goldfeld 和 Wujciak-Jens，2020年）生成协方差矩阵。在每种情景的每次重复实验中，都会生成一个新的相关矩阵。在所有模拟中，$T = 100$ 都会随机选取真实预测变量。系数 β_i 模拟自独立同分布的标准正态分布，适用于 $i = 1$、.....、T。高斯回归的数据集规模固定为 $p = 300$、$n = 1000$，逻辑回归和泊松回归的数据集规模固定为 $p = 200$、$n = 5000$。逻辑回归模拟中的截距项设定为使每个数据集的边际概率为 0.5；泊松计数的截距固定为 -1，通常会产生介于 0 到 10 之间的响应。

Mirror 方法遵循 Dai 等（2023）：数据驱动使用单次数据划分，而 MDS 使用 50 次数据划分。数据驱动和 MDS 的结果均使用 Li 等（2023）给出的 R 代码获得。Knockoff 结果使用 knockoff v0.3.6 R 软件包（Patterson and Sesia，2022年）生成。

对于每次模拟和每次重复实验，我们模拟一个相同规模的测试集，用于评估均方误差和分类准确率。对于泊松回归，均方误差采用与高斯情形相同的方法计算。所有方法的目标都是将广义错误发现率控制在 $q = 0.05$。

我们所有模拟所用的 R 代码均可在以下 GitHub 代码仓库中获取：https://github.com/s-organ/SHRED-R。

4.2 结果

图 1–3 展示了各场景下每种方法的 gPower、gFDR，以及均方误差或分类准确率。对于所有方法，均方误差（MSE）和分类准确率的计算方式是：首先仅使用该方法在训练数据中选出的变量重新拟合模型，然后在独立测试数据集上评估预测性能。对于 SHRED 方法，在重新拟合时，会从每个选出的集合中随机选取一个变量。

图中还列出了 LASSO 的均方误差和分类准确率结果，以便比较预测性能；LASSO 的 gPower 和 gFDR 结果则另见附录 ??。LASSO 变量选择采用交叉验证调优参数，该参数使交叉验证均方误差或负对数似然达到最小；预测性能则通过使用相应 LASSO 系数的模型进行评估。

对于 SHRED 方法，分别将函数 σ 应用于正确拒绝和错误拒绝的零假设，以计算广义检验功效和广义错误发现率。由于检验效能较低，逻辑斯蒂和泊松模拟的图中省略了 knockoff 方法；同样，由于性能不佳，若干图中未展示采用数据驱动（DS）的镜像方法。这些方法的完整结果以及各图对应的模拟表格见附录 ??。

4.2.1 高斯回归

从图 1 可以看出，在所有情景中，采用 PRDS（或启发式方法）截断值的 SHRED 和 SHREDDER，其广义检验功效始终显著高于选择单个变量的方法；例外是套索回归，因为它不控制错误发现率。在大多数情景中，采用 PRDS 截断值的 SHRED 和 SHREDDER 的广义错误发现率也显著低于 Benjamini–Hochberg 方法。镜像方法和 knockoff 方法的性能不稳定，在聚类相关性设置中往往检验效能很低，且无法控制广义错误发现率。SHRED 在所有情景中的表现均优于由（BY），实现了更高的广义检验功效和更低的广义错误发现率。

尽管在这些模拟中无法保证满足 PRDS 假设，但旨在 PRDS 条件下控制广义错误发现率的方法在实践中确实能将广义错误发现率控制在名义水平。正如预期，不对 p 值施加任何依赖性假设的广义错误发现率控制方法，在实践中表现出更严格的控制效果——通常远低于目标水平，但代价是检验效能降低。然而，在某些情形下，SHRED 能够选择变量集，因此在保持广义错误发现率显著更低的同时，其广义检验功效可与 Benjamini–Hochberg 方法相当。

就预测性能而言，在较低相关性或公共分量相关结构下，各方法的均方误差相近。在较高相关性下，SHRED方法——尤其是采用 PRDS 截断值的 SHRED——往往优于其他方法。几乎在所有情况下，错误发现率控制程序的均方误差都不及套索回归。我们对所有模拟进行了单样本 t 检验，以确定均方误差差异是否显著。结果见补充附录，表 ??. 在大多数情况下，均方误差差异并不显著；但对于聚类相关、低相关性以及带有 $\rho = 0.9$ 的 AR(1)相关性，SHRED-PRDS 和 SHREDDER 均显著优于 Benjamini–Hochberg 方法，而 SHRED 显著优于 BY方法。对于逻辑回归，也有若干情形下 SHRED方法显著优于相应的 Benjamini–Hochberg 方法或 BY方法。

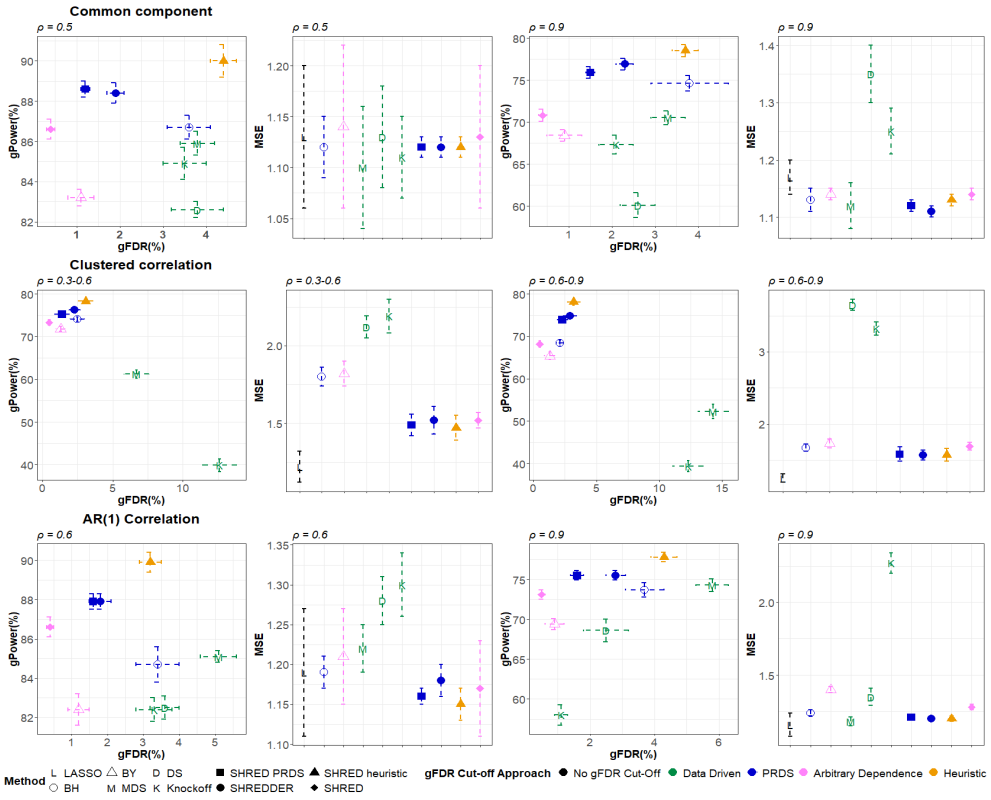


图 1：高斯回归模拟结果。误差棒表示一个标准误。

对 SHRED-PRDS 和 SHREDDER 结果的详细考察表明，所选集合中大多数都是单元素集，真实单元素集的数量与 BH 识别出的数量十分接近。值得注意的是，错误选出的单元素集数量低于 BH 产生的错误选择数量。其余所选集合大多包含 2 个元素；在预测变量相关性更强的情形下，大小超过 1 的集合所占比例会增加。

我们还进行了模拟研究，其中预测变量是从标准高斯分布中独立生成的。结果见补充附录 ??。由于传统的错误发现率控制变量选择方法在预测变量相关时通常表现不佳，我们预期 BH 和 BY 在预测变量相互独立的情形下表现良好，而 SHRED 为适应对预测变量集合进行检验所需的较大 α 值可能会降低其检验效能。然而，即使预测变量是通过模拟独立生成的，我们也没有观察到 SHRED 与 BY 之间或 SHRED-PRDS 与 BH 之间存在显著差异。

即使在独立情形下，我们的方法表现也没有明显逊于 BH 或 BY，原因有二。首先，由于检验预测变量集合而导致的 α 增幅有限。由于额外的集合权重较小，且数量少于单元素集，α 的增幅受到限制——在完美二叉树下为 $\sim 33\%$——因此，即使没有选出任何多预测变量集合，有效 FDR 控制水平也仅下降 $\sim 25\%$。其次，即使预测变量实际上相互独立，某些二元组之间仍可能出现中等程度的样本相关性，这会给识别真实预测变量带来挑战。

SHRED 在相关情形下表现出色，在独立情形下也具有竞争力，这表明无须顾虑预测变量之间的底层相关结构，SHRED 就能广泛应用于变量选择问题。

4.2.2 逻辑回归

对于逻辑回归（图 2），所有方法的广义检验功效均低于高斯回归情形，因为二项响应所包含的信息少于高斯响应。不过，各方法的相对性能与高斯情形非常相似：采用 PRDS 截断值的 SHRED 和 SHREDDER 通常比 BH 具有更高的广义检验功效，且广义错误发现率往往更低；在所有情形下，它们的广义错误发现率都低于名义水平。正如预期，没有一种错误发现率控制方法能达到套索回归的检验效能。

SHRED 的广义检验功效始终高于 BY，广义错误发现率则始终低于 BY，尽管标准误较大，因此在单个情景中，这些差异往往并不显著。广义检验功效和广义错误发现率的改善带来了更高的分类准确率。在许多情景中，这种改善具有显著性（见补充表 ??）。特别是在两个一阶自回归情景中，SHRED-PRDS 和 SHREDDER 的准确率均显著高于 BH，而 SHRED 的准确率显著高于 BY。在许多情况下，尽管 SHRED 的广义检验功效较低，其表现甚至优于 BH。镜像方法在所有情景中都保持了错误发现率控制，但检验功效的变异性很大；使用多个数据划分可将其检验功效提高到与 BH 相当的水平。

与高斯情形一样，SHRED 方法主要选择单元素集，仅选择少量二元组和极少数更大的集合。S HRED_PRDS 和 SHRED- DER 选中的真实单元素集数量，与 BH 选中的真实变量数量相比，差异并不显著。

我们还进行了具有独立预测变量的逻辑回归模拟，并发现，与高斯情形一样，SHRED方法的表现并未显著差于相应的Benjamini–Hochberg 方法或由程序。结果见补充附录??。

4.2.3 泊松回归

对于泊松回归模拟（图3），在公共分量情景和聚类相关情景下，所有方法的广义检验功效甚至低于逻辑回归。尽管如此，

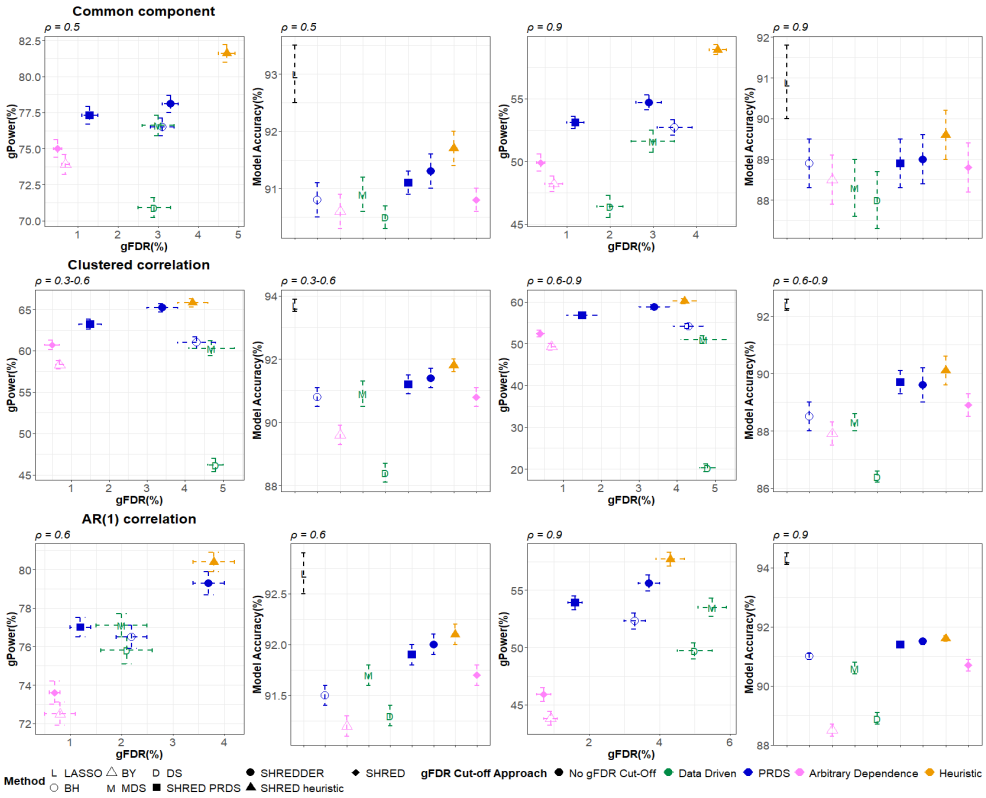


图 2：逻辑回归模拟结果。误差棒表示一个标准误。

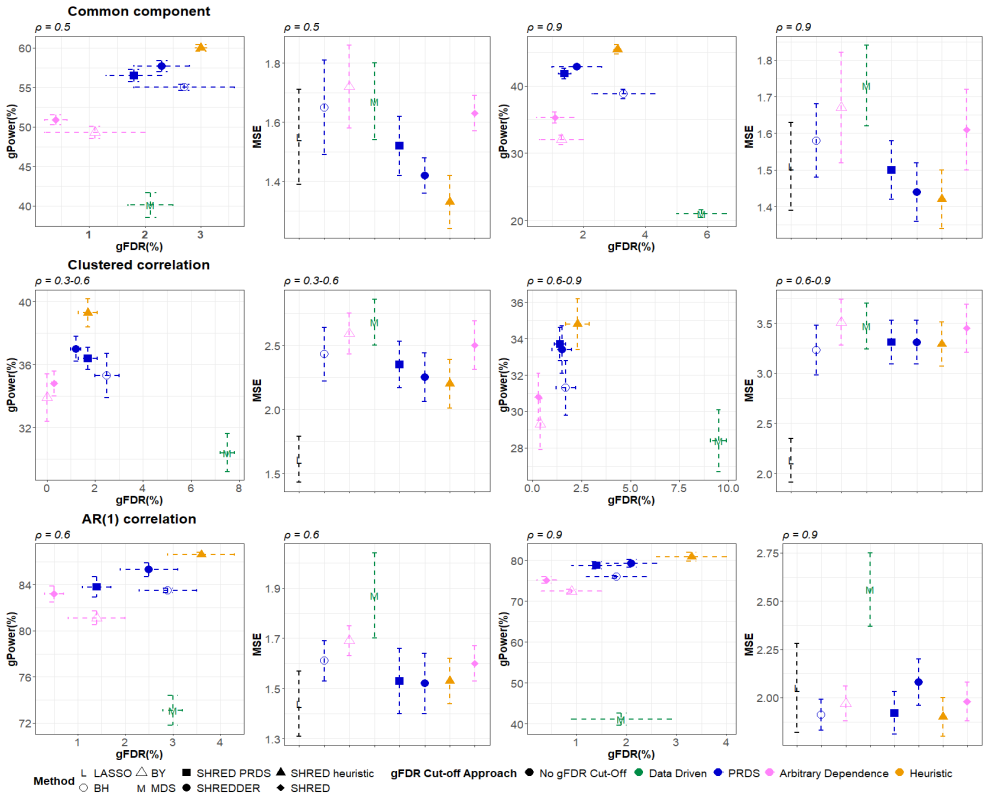


图3：泊松回归模拟结果。误差棒表示一个标准误。

表2：在模拟高斯 Y数据上的《微生物组》 X 数据的 SHRED 程序

方法	gFDR Cut-off	广义检验功效	SE 广义检验功效	广义错误发现率	SE 广义错误发现率	MSE 标准误	均方误差
套索回归	无广义错误发现率截断值	88.6%	0.2%	46.3%	0.9%	1.17	0.04
镜像	MDS（数据驱动）	70.2%	0.5%	6.7%	0.2%	1.53	0.06
镜像	DS（数据驱动）	66.5%	0.5%	3.3%	0.2%	2.28	0.10
Knockoff	数据驱动	69.0%	0.8%	3.7%	0.3%	2.68	0.05
BH	PRDS	82.3%	0.7%	3.6%	0.4%	1.33	0.03
SHRED	PRDS	85.5%	0.3%	1.5%	0.2%	1.29	0.02
SHRED	启发式方法	88.3%	0.2%	4.2%	0.5%	1.18	0.05
SHREDDER	PPRDS	85.8%	0.5%	3.5%	0.4%	1.22	0.04
SHRED	任意	80.5%	0.4%	0.2%	0.1%	2.05	0.08
BY	任意	77.7%	0.8%	0.7%	0.3%	1.87	0.08

相对性能与其他模拟中的模式相似。SHRED 通常比由实现更高的广义检验功效和更低的广义错误发现率，而 SHRED_PRDS 和 SHREDDER 通常比 Benjamini–Hochberg 方法实现更高的广义检验功效和更低的广义错误发现率。采用 MDS 的镜像方法再次表现不可靠，往往性能较差，且经常无法控制错误发现率。

与高斯和逻辑斯蒂情形一样，没有任何错误发现率控制方法的检验效能能与套索回归相匹敌。在公共分量和聚类相关情景中，套索回归的均方误差显著低于所有错误发现率控制方法。然而，在 AR(1) 相关情景中，SHRED_PRDS 和 SHREDDER 的均方误差与套索回归相当。SHRED_PRDS 和 SHREDDER 的均方误差通常小于 Benjamini–Hochberg 方法，但只有在单一公共成分的 $\rho = 0.5$ 情形下，SHREDDER 的差异才达到显著水平（见补充表 ??）。

与先前结果类似，SHRED方法主要选择单元素集，同时选择少量大小为2的集合以及极少更大的集合。SHREDPRDS与SHREDDER所选真实单元素集的数量，与Benjamini–Hochberg 方法所选真实变量的数量没有显著差异。

我们还开展了含独立预测变量的泊松回归模拟，并发现与其他设定一样，SHRED方法的表现并未显著差于相应的Benjamini–Hochberg 方法或BY程序。结果见补充附录??。

4.3 基于真实微生物组数据的模拟

我们还利用MIDAS 4全球污水分类数据集（Dueholm 等，2022），评估了我们的方法在具有真实世界相关结构的数据上的性能。该数据集包含从全球污水处理厂采集的$n = 1$,278个微生物组样本，其中 $p = 604$个分类单元（属级）的丰度以每个样本的整数读数计数报告。这些计数代表绝对测序量，表示分配给每个扩增子序列变体（ASV）或分类单元的读数。

在我们的分析中，我们首先将原始计数转换为相对丰度，然后应用中心对数比（CLR）变换，遵循微生物组文献中关于处理测序深度差异的建议（Quinn 等，2019；Yerke 等，2024）。零计数通过在变换前向每个计数添加一个小常数（10^{-6}）来处理。以每个属经 CLR 变换后的相对丰度作为预测变量，我们模拟了正态分布响应变量 Y 的 100 个重复样本。对于每个重复样本，我们随机选择 $T = 200$ 个真实信号，然后模拟$Y \sim N(X\beta, 1)$，其中真实信号的 β 系数独立地从 $N(0, 1)$ 分布生成。在完整数据集上进行变量选择，并通过 10 折交叉验证估计均方误差（MSE）。对于所有方法，目标广义假发现率水平设定为 $q = 0.05$。

结果如表2所示。除采用启发式截断值的 SHRED 外，在保持广义错误发现率控制的同时，就广义检验功效而言表现最佳的方法是 SHREDDER。尽管采用 PRDS 截断值的 SHRED 与 SHREDDER 的广义检验功效均显著高于 Benjamini–Hochberg 方法，但只有 SHREDDER 的均方误差相较 Benjamini–Hochberg 方法有显著改善。

4.4 进一步提升 SHRED 方法选择性能的潜在调整

SHRED 的基本思路是：对于难以区分真实变量的情况，通过检验代理变量集合来获得更高的广义检验功效。代价是，检验更多假设会迫使我们增大 α，进而可能降低检验效能。因此，检验相关性较低的预测变量集合可能会降低性能。我们或许可以通过在指定相关性值处截断层次树来提升性能，即只检验相关性高于某个阈值的集合。这样，方法便只需检验单元素集，以及包含高度相关代理变量的较小集合，从而减小 α。我们采用第4.1节中用于高斯回归的设计生成模拟数据，其中自回归（AR(1)）相关结构的强度分别为 $\rho = 0.6$、0.9 和 0.99。每种情形下均进行 100 次重复实验。在每个相关性截断值下，将广义检验功效与标准 BH方法进行比较。

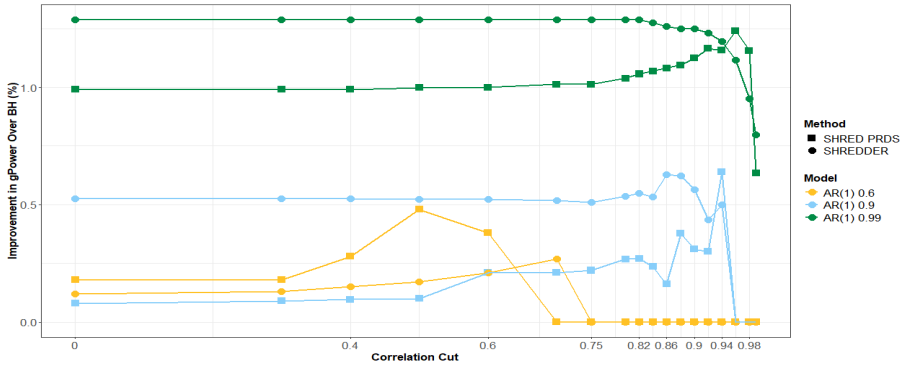


图4：在层次树的不同相关性截断位置下，SHRED方法与Benjamini–Hochberg 方法的广义检验功效差异，$p = 300, n = 1000, T = 100$。

图4显示，通过截断层次树，广义检验功效仍有进一步提升的可能。然而，提升幅度相对较小，最佳截断位置因情景而异，而且许多错误的截断位置表现不如使用完整的层次树，因此在对此方法提出任何建议之前，还需要开展更多研究。

5 真实数据应用

我们研究 MIDAS 全球污水分类学数据集（Dueholm 等，2022）。与使用该数据进行的模拟一样，我们按照现有文献的建议（Quinn 等，2019；Yerke 等，2024），对相对丰度应用中心对数比（CLR）变换。零计数通过向每个计数添加一个小常数（10^{-6}）来处理。Dueholm等人（2022）发现，整体微生物群落受到大陆的强烈影响。据推测，这是由于污水处理厂采样不均衡，并受到不同工厂类型和工业负荷效应的混杂。对样本进行的 Bray-Curtis β 多样性主坐标分析表明，来自欧洲和北美的样本群落

表3：真实数据应用中的模型选择和性能

方法	FDR 截止值	$\sigma(I_{c_{max}})$	准确率	LL
套索回归	无 FDR 截断值	116	88.5%	-139.1
镜像	MDS（数据驱动）	8	76.5%	-602.1
镜像	DS（数据驱动）	4	73.4%	-627.3
Knockoff	数据驱动	0	NA	NA
BH	PRDS	45	83.8%	-285.8
SHRED	PRDS	45.75	84.4%	-237.4
SHRED	启发式方法	52	85.1%	-233.9
SHREDDER	PPRDS	46.75	84.8%	-235.9
SHRED	任意	16.9	80.5%	-436.9
BY	任意	15	79.3%	-458.7

彼此之间差异并不过大，而这些群落与其他大陆的群落则明显不同。因此，我们基于分类学数据拟合逻辑回归，以预测样本来源大陆是否位于欧洲/北美。为拟合逻辑回归，通过去除任何稀有或条件性高丰度分类群，将属级分类学数据缩减至 $p = 162$、$n = 1278$ 。对于所有方法， q 设为0.05。为评估模型准确率，将数据划分为分别占数据90%和10%的训练集和测试集。将测试集中每个样本都归入最大类别所对应的默认准确率为72.4%。

表3比较了各种变量选择方法所选属的数量，以及所得模型的测试准确率和对数似然。SHREDDER 和 SHRED PRDS 选择的变量比 BH 更多_，因此测试准确率和对数似然也更高。与 BH 和 SHRED 方法相比，采用 MDS 的镜像方法选择的分类单元更少，因此预测准确率和对数似然较低。这与表2的结果一致。我们在模拟中发现，镜像方法并不总能控制错误发现率，因此镜像方法选出的属中很可能有一部分是错误发现。

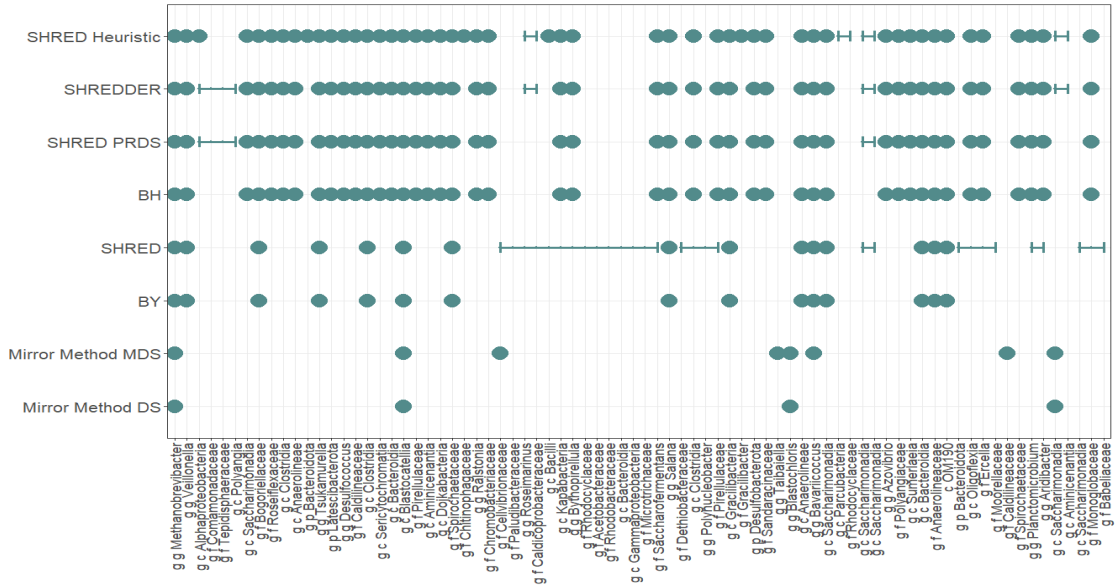
图5展示了各方法选出的分类单元。对于 SHRED 方法，区间表示一组被选中的分类单元。分类单元按可识别到的最低分类等级标注。

SHRED_PRDS 的结果与 BH 非常相似，仅在选择两组属时有所不同：一组是来自不同纲和门的四个未命名属，另一组是来自 Saccharamonadia 纲的两个未命名属。Saccharamonadia 已在多种不同环境中被发现，包括人唾液，这表明它们在生活污水中的丰度可能高于工业废水。这可以解释其作为预测因子的价值：Dueholm等人（2022）指出，工业负荷是大陆的一个混杂因素。在测试数据中，每组中加入一个属都能提升模型性能。

SHREDDER 选出了一组非常相似的预测变量，但还额外选出了两对属：一对是 Roseimarinus 和 Caldicoprobacteraceae；另一对是来自 Saccharamonadia 纲和 Aminicenantia 纲的两个未命名属。镜像方法将来自 Saccharamonadia 纲的这个属识别为真实预测变量。这个属与来自 Aminicenantia 纲的属之间的相关性可能需要进一步研究。

SHRED 选择了与由方法相同的各个属，但还选择了若干属的集合。其中一组是 SHRED_PRDS 选出的、来自 Saccharamonadia 的一对属。即使采用 BH临界值，我们也无法确定哪个属是真实预测变量，但至少有一个属是真实预测变量的证据已足以达到 strictSHRED临界值。对于其他集合，BH 至少选出了一个属，这支持了我们的主张：每个集合都包含一个真实预测变量。

图5：Midas 案例研究中各方法选出的分类单元



6 结论

我们提出了一种新的变量选择框架，可在相关设置下选择预测变量集合；在这些情况下，无法识别单个真实预测变量，但我们确信集合中至少有一个预测变量具有相关性。这解决了 FDR 控制方法的一项关键局限：与套索回归等面向预测的方法相比，这类方法的检验效能往往较低。在这一框架下，我们基于对变量集进行假设检验所得的 p-值，开发了若干变量选择程序。

这些假设之间的逻辑关系要求对现有多重检验方法进行大幅推广，由此形成了我们的广义线性逐步向上程序。这一推广具有广泛的适用性，并不限于变量选择问题。由于我们的 SHRED 和 SHREDDER 方法仅依赖有效且保守的 p 值，因此不像镜像方法和 model-X knockoff 程序那样局限于广义线性模型。

我们证明了广义线性逐步向上程序能够控制广义假发现率（gFDR）。此外，当 p 值满足 PRDS 条件时，可以降低拒绝阈值 α，从而在维持 gFDR 控制的同时提升检验效能。

大量模拟研究表明，我们的方法在各种设定下均优于现有方法。尤其是，基于真实数据开展的模拟中表现出色，这说明我们的方法在实际应用中的优势可能更为显著。

这项工作在变量选择和多重假设检验方面提出了若干概念性进展，并开辟了多个未来研究方向。首先，虽然层次聚类能够方便地定义预测变量集合，但未必能充分描述所有预测变量结构。在非层次设定中，可能会出现重叠的选定集合，这使得合适的规模函数 σ 难以定义。另一个相关方向是开发数据依赖的规模函数，将相关性信息纳入考量，而非仅仅依赖集合大小。

另一个有趣的方向是 SHRED 的自适应版本。回顾一下，GLSUP 中实际的广义错误发现率控制依据的是真实零假设的规模函数，而非所有假设的规模函数；我们使用的是

针对所有假设的取值，并将其作为上界，以确保广义错误发现率控制。对于原始 BH 程序，通过使用数据驱动的真实零假设数量估计值，而非这个宽松的上界，检验效能已得到提升。若将类似方法用于 GLSUP，在假零假设占比很高的情况下，检验效能可能会大幅提升。

最后，GLSUP 框架可以扩展到非线性阈值函数。为方便起见，我们采用了线性阈值$\sigma\left(I_c\right) \geqslant \alpha c$，但也可以选择其他阈值。探索这些替代方案可能会催生新的广义错误发现率控制方法类别。

参考文献

Barber, R. F. 和 Candes, E. J.（2015）。通过 knockoff 控制错误发现率。《统计年鉴》，43(5)。Barber, R. F.、Candes, E. J. 和 Samworth, R. J.（2020）。利用 knockoff 进行稳健推断。《统计年鉴》，48(3)。Basu, P.、Cai, T. T.、Das, K. 和 Sun, W.（2018）。大规模多重检验中的加权错误发现率控制。《美国统计学会会刊》，113(523):1172–1183。Benjamini, Y. 和 Hochberg, Y.（1995）。错误发现率的控制：一种实用且有效的多重检验方法。《英国皇家统计学会会刊B辑：统计方法论》，51(1):289–300。Benjamini, Y. 和 Hochberg, Y.（1997）。带权重的多重假设检验。《斯堪的纳维亚统计学杂志》，24(3):407–418。Benjamini, Y. 和 Yekutieli, D.（2001）。依赖关系下多重检验中错误发现率的控制。《统计年鉴》，29(4)。Bertsimas, D.、King, A. 和 Mazumder, R.（2016）。以现代优化视角进行最佳子集选择。《统计年鉴》，44:813–852。Bogomolov, M.、Peterson, C. B.、Benjamini, Y. 和 Sabatti, C.（2021）。树上的假设：新的错误率和检验策略。Biometrika，108(3):575–590。Dai, C.、Lin, B.、Xing, X. 和 Liu, J. S.（2022）。通过数据分割控制错误发现率。《美国统计学会会刊》，118(544):2503–2520。Dai, C.、Lin, B.、Xing, X. 和 Liu, J. S.（2023）。广义线性模型中错误发现率控制的尺度无关方法。《美国统计学会会刊》，118(543):1551–1565。Dueholm, M. K. D.、Nierychlo, M.、Andersen, K. S.、Rudkjøbing, V.、Knutsson, S.、Consortium, M. G.、Albertsen, M. 和 Nielsen, P. H.（2022）。Midas 4：用于污水处理厂细菌群落研究的全长16S rRNA基因序列与分类全球目录。《自然-通讯》，13(1)。Finner, H.、Dickhaus, T. 和 Roters, M.（2009）。关于错误发现率和渐近最优拒绝曲线。《统计年鉴》，37(2)。Goeman, J. J. 和 Finos, L.（2012）。继承程序：对树结构假设进行多重检验。《遗传学与分子生物学中的统计应用》，11(1):文章11。Goldfeld, K. 和 Wujciak-Jens, J.（2020）。simstudy：通过数据生成阐明研究方法。《开源软件杂志》，5(54):2763。

Holm, S.（1979）。一种简单的序贯拒绝式多重检验程序。《斯堪的纳维亚统计学杂志》，6(2):65–70。Hommel, G.（1986）。适用于任意依赖结构的多重检验程序。Metrika，33:321–336。Katsevich, E. 和 Sabatti, C.（2019）。多层 knockoff 筛选器：在多个分辨率下进行受控变量选择。《应用统计年鉴》，13(1):1–33。Li, S.、Yao, Y. 和 Zhang, C.-H.（2023）。对“广义线性模型中错误发现率控制的尺度无关方法”的评论。《美国统计学会会刊》，118(543):1586–1589。Liu, L.、Gu, H.、Van Limbergen, J. 和 Kenney, T.（2021）。Surf：一种用于稀疏变量选择的新方法及其在微生物组数据分析中的应用。《医学统计学》，40(4):897–919。Marcus, R.、Peritz, E. 和 Gabriel, K. R.（1976）。关于闭合检验程序，特别是有序方差分析。Biometrika，63:655–660。Meinshausen, N.（2008）。变量重要性的层级检验。Biometrika，95(2):265–278。Meinshausen, N. 和 Bühlmann, P.（2010）。稳定性选择。《皇家统计学会会刊：B辑（统计方法论）》，72(4):417–473。Patterson, E. 和 Sesia, M.（2022）。knockoff：用于受控变量选择的 Knockoff筛选器。R 软件包版本 0.3.6。Quinn, T. P.、Erb, I.、Richardson, M. F. 和 Crowley, T. M.（2019）。将测序数据理解为组成数据：展望与综述。《生物信息学》，34(16):2870–2878。Ramdas, A.、Chen, J.、Wainwright, M. J. 和 Jordan, M. I.（2019a）。一种用于有向无环图上错误发现率控制的序贯算法。Biometrika，106(1):69–86。Ramdas, A. K.、Barber, R. F.、Wainwright, M. J. 和 Jordan, M. I.（2019b）。利用 p-filter 结合先验知识进行多重检验的统一处理。《统计年鉴》，47:2790–2821。Rastaghi, S.、Saki, A. 和 Tabesh, H.（2024）。基于信息论修正错误发现率程序：任意相关结构下的应用及其在高维基因组数据中的性能。《BMC生物信息学》，25(1)。Schwartzman, A. 和 Lin, X.（2011）。相关性对错误发现率估计的影响。Biometrika，98(1):199–214。Sesia, M.、Katsevich, E.、Bates, S.、Candès, E. 和 Sabatti, C.（2020）。基因组中因果变异的多分辨率定位。《自然-通讯》，11:1093。Tibshirani, R.（1996）。利用套索回归进行回归收缩与变量选择。《英国皇家统计学会会刊B辑：统计方法论》，58(1):267–288。Wang, R. 和 Ramdas, A.（2022）。利用 e 值控制错误发现率。《英国皇家统计学会会刊B辑：统计方法论》，84(3):822–852。Xing, X.、Zhao, Z. 和 Liu, J. S.（2021）。利用高斯镜像控制错误发现率。《美国统计学会会刊》，118(541):222–241。

Yekutieli, D.、Reiner-Benaim, A.、Benjamini, Y.、Elmer, G. I.、Kafkafi, N.、Letwin, N. E. 和 Lee, N. H.（2006年）。微阵列分析中复杂研究的多重性问题处理方法。《荷兰统计学》，60：414–437。Yerke, A.、Fry Brumit, D. 和 Fodor, A. A.（2024年）。基于比例的归一化在机器学习应用中的表现优于组成数据变换。《微生物组》，12(1)：45。Yuan, M. 和 Lin, Y.（2006年）。含分组变量的回归模型选择与估计。《皇家统计学会会刊：B辑（统计方法论）》，68(1)：49–67。Zhang, T. 和 Cavenaugh, J. E.（2016年）。一种基于 Kullback–Leibler差异的最佳子集模型选择多阶段算法。《计算统计学》，31：643–669。Zou, H.（2006年）。自适应套索及其预言机性质。《美国统计学会杂志》，101(476)：1418–1429。

《用于错误发现率控制的集合式分层变量选择与广义线性逐步向上程序》的补充附录

A. 定理的证明

In t本节将证明，GLSUP 在名义水平上控制广义错误发现率。我们首先
est建立我们的记号并回顾 GLSUP。

A.1. GLSUP

设 $H_1, \ldots, H_m$ 为待检验的假设。设 $I_N \subseteq \{1, \ldots, m\}$ 表示真实零假设指标的（未知）集合，I_A 为其补集。设 $\mathcal{P}$ 表示 $\{1, \ldots, m\}$ 的幂集，即 $\mathcal{P} = \{J | J \subseteq \{1, \ldots, m\}\}$。

用于广义错误发现率控制的广义线性逐步向上程序（GLSUP）需要以下输入：

- 保守的 p值随机向量 $\mathbf{P} = \{P_1, \ldots, P_m\}$。（所谓保守的 p值，是指对于每个 $i \in I_N$，P_i 的边缘分布均为超均匀分布，即 P_i 的边缘分布随机占优于均匀分布，或者用公式表示为，$(\forall c \in [0,1]) P(P_i \leqslant c) \leqslant c$。）

- 一个集合 $\mathcal{R} \subseteq \mathcal{P}$，由一组“可拒绝”的假设集合构成。（这可以是整个幂集。）我们要求 $\mathcal{R}$ 对交和并运算封闭，且包含空集以及$\{1, \ldots, m\}$。这意味着对于任意 $J \subseteq \{1, \ldots, m\}$，都存在一个最小集合 $J \in \mathcal{R}$ 满足$J \subseteq J$。集合 J 称为 的闭包 J。我们还要求 $I_A \in \mathcal{R}$。（这里的想法是，$\mathcal{R}$ 表示所有可能合理地被 I_A的假设集合，因此我们自然假设，无论真实的 I_A 是什么，它都应当属于 $\mathcal{R}$。）

- 规模函数 $\sigma: \mathcal{R} \to \mathbb{R}_{\geqslant 0}$，它关于集合包含关系单调递增。这意味着对于任意 $J_1, J_2 \in \mathcal{P}$，若满足 $J_1 \subseteq J_2$，则有 $\sigma(J_1) \leqslant \sigma(J_2)$。（注意，我们不要求 σ 严格递增。）

- 截断阈值 α。（这是在特定假设下，利用本节中的定理根据所需的广义错误发现率控制水平 q 计算得出的。）

根据这些输入，对于任意 $c \in [0,1]$，我们令 I_c 表示集合 $I_c(\mathbf{P}) = \{i | P_i \leqslant c\}$，并令 I_c表示集合 $\overline{I}_c(\mathbf{P}) = \{i | P_i \leqslant \overline{c}\}$。我们定义 $\overline{c_{\max}(\mathbf{P})} = \underline{\sup}\{c \in [0,1] | \sigma(I_c) \geqslant \alpha c\}$。然后，我们拒绝假设集合 $\{H_i | i \in I_{c_{\max}}\}$。

A.2. 定理的证明

我们的目标是证明，若正确选择 α，该程序便能在指定水平上控制广义错误发现率。我们从一般情形开始，不对 p-值之间的依赖关系作任何假设。

定理1. 对于来自 *A.1* 节的输入，假设对任意 $J \subseteq \{1,\ldots,m\}$，$\sigma(J) \leqslant \sum_{i\in J} w_i$ 对某些权重 w_1，……，w_m。则具有截断的广义线性逐步向上程序

$$\alpha = \frac{\left(\sum_{i\in I_N} w_i\right)\left(1+\log(\sigma(\{1,\ldots,m\}))\right) - \sum_{i\in I_N} w_i \log(\sigma_i)}{q}$$

其中 $\sigma_i = \sigma(\{i\})$，将广义错误发现率控制在水平 q。

证明。令 $S = I_{c_{\max}}$(P) 为具有较小 p 值的被拒绝假设的集合。令 $K = \sigma\left(S\right)$为应用于所有被拒绝假设集合的规模函数，并令 $V = \sigma\left(S \cap I_N\right)$ 为最终被拒绝的真实原假设集合的规模函数。与 Benjamini 和 Yekutieli （ 2001 ）中的证明一样，错误发现率为

$$\begin{aligned}
\mathbb{E}(Q) &= \mathbb{E}\left(\frac{V}{K}\right) \\
&= \mathbb{E}\left(\frac{\sigma\left(\overline{S} \cap I_N\right)}{K}\right) \\
&\leqslant \mathbb{E}\left(\frac{\sum_{i\in S\cap I_N} w_i}{K}\right) \\
&= \mathbb{E}\left(\frac{\sum_{i\in I_N} 1_{i\in S} w_i}{K}\right) \\
&= \sum_{i\in I_N} w_i \mathbb{E}\left(\frac{1_{i\in S}}{K}\right)
\end{aligned}$$

wh此处第三行之所以成立，是因为 $\mathcal{R}$ 在并与交下封闭且 $I_A \in \mathcal{R}$，所以 we have

$$\overline{S} \subseteq \overline{I_A \cup (S \cap I_N)} = \overline{I_A} \cup \overline{(S \cap I_N)} = I_A \cup \overline{(S \cap I_N)}$$

因此 $S \cap \overline{I_N} \subseteq (S \cap \overline{I_N})$。如果令 ϕ_i 为 P_i的边缘分布测度，则

我们有

$$\begin{aligned}
\sum_{i\in I_N} w_i \mathbb{E}\left(\frac{1_{i\in S}}{K}\right) &= \sum_{i\in I_N} w_i \int_0^1 \mathbb{E}\left(\frac{1_{i\in S}}{K}\middle| P_i = p\right) d\phi_i(p) \\
&= \sum_{i\in I_N} w_i \int_0^1 \mathbb{E}\left(\frac{1_{K\geqslant(\alpha p\vee\sigma_i)}}{K}\middle| P_i = p\right) d\phi_i(p) \qquad (1)\\
&\leqslant \sum_{i\in I_N} w_i \int_0^{\frac{\sigma(I)}{\alpha}} \mathbb{E}\left(\frac{1_{K\geqslant(\alpha p\vee\sigma_i)}}{K}\middle| P_i = p\right) d\phi_i(p) \qquad (2)\\
&\leqslant \sum_{i\in I_N} w_i \int_0^{\frac{\sigma(I)}{\alpha}} \left(\frac{1}{\alpha p}\wedge\frac{1}{\sigma_i}\right) d\phi_i(p) \\
&\leqslant \sum_{i\in I_N} w_i \int_0^{\frac{\sigma(I)}{\alpha}} \left(\frac{1}{\alpha p}\wedge\frac{1}{\sigma_i}\right) dp \qquad (3)\\
&= \sum_{i\in I_N} \frac{w_i}{\alpha}\left(1+\int_{\frac{\sigma_i}{\alpha}}^{\frac{\sigma(I)}{\alpha}} \frac{1}{p}\,dp\right) \\
&= \sum_{i\in I_N} \frac{w_i}{\alpha}\left(1-\log(\sigma_i)+\log(\sigma(I))\right) \\
&= \frac{\sum_{i\in I_N} w_i}{\alpha}\left(1+\log(\sigma(I))\right) - \frac{1}{\alpha}\sum_{i\in I_N} w_i \log(\sigma_i)
\end{aligned}$$

（1）是因为 $c_{\max} = \frac{K}{\alpha}$，因此根据 $S = I_{c_{\max}}$ 的定义，只有当 $P_i \leqslant \frac{K}{\alpha}$ 时才有 $i \in S$。（2）是因为 $K \leqslant \sigma(I)$，所以如果 $P_i > \frac{\sigma(I)}{\alpha}$，那么 $i \notin S$。（3）是因为 P_i 的边缘分布是超均匀的。

□

注1。在假设之间没有结构的情况下，因此 $\sigma(J) = |J|$ 是基数，我们令 $w_i = 1 = \sigma_i$，得到 $\alpha = \frac{m_0}{q}(1+\log(m))$。由于 m_0 未知，将 $\overline{m}$ 用作上界是标准做法。因此，定理*1*给出 $\alpha = \frac{m}{q}(1+\log(m))$。这与 ***BY*** 方法（***Benjamini*** 和 ***Yekutieli***，*2001*年）非常接近，但略高；***BY*** 方法使用 $\underline{\alpha} = \frac{m}{q}\sum_{i=1}^{m}\frac{1}{i}$。差异在于，在 ***Benjamini*** 和 ***Yekutieli***（*2001*年）的原始问题中，K 被限制为整数，因此我们将 $\frac{1}{K}$ 以 $\frac{1}{\alpha p}$ 为上界时，也可以改为以 $\frac{1}{\lceil \alpha p\rceil}$ 为上界，从而得到 ***BY*** 方法中的 α 值。

如果我们对 p 值之间的依赖关系作出更强的假设，就可以用更小的 α 证明广义错误发现率控制，从而获得更高的检验效能。具体而言，回顾 Benjamini 和 Yekutieli（2001）提出的子集上的正回归依赖（PRDS）条件：

定义 1（PRDS）。一组 p 值 $\mathbf{P} = \{P_1, \ldots, P_m\}$ *s*满足子集 $J \subseteq \{1, \ldots, m\}$ 上的 PRDS 条件，当且仅当对于任意 $i \in J$和任意上闭集 $U \subseteq [0,1]^{m-1}$, $P(\mathbf{P}_{\hat{i}} \in U | P_i = x)$ 是关于 x的单调递增函数，其中 $\mathbf{P}_{\hat{i}}$ 是从 $\mathbf{P}$ 中移除第 i 个索引后得到的。

定理 2。对于 ***A.1*** 节中的输入，假设保守的 p-值 $P_1, \ldots, P_m$满足关于 I_N 的 ***PRDS*** 条件，且规模函数 $\sigma: \mathcal{P} \longrightarrow \mathbb{R}_{\geqslant 0}$ 可由权重和界定：$(\forall J \subseteq \{1, \ldots, m\})\ \left(\sigma(J) \leqslant \sum_{i\in J} w_i\right)$，其中权重为 $w_i \geqslant 0$。那么，采用截断值 $\alpha = \frac{\sum_{i=1}^{m} w_i}{q}$ 的广义线性逐步上升法可在水平 q 控制广义错误发现率。

在某些情况下，σ 无法由权重和紧密界定，但可由成对权重和更紧密地界定——也就是说，我们或许能够对$\sigma\left(J\right) \leqslant \sum_{i,j\in J} w_{ij}$ 进行界定，其中权重为 $(w_{ij})_{i,j=1}^{m}$。在这些情况下，若假设 PRDS 条件具有成对形式，我们便能以更紧的界证明广义错误发现率控制：

定义 2 (PPRDS)。一组 $\{P_1, \ldots, P_m\}$ 个取值于 $[0,1]$ 的随机变量，如果满足在子集 $J \subseteq \{1, \ldots, m\}$ 上的 *PRDS* 条件，且对于任意$i \neq j \in J$以及任意上集 $U \subseteq [0,1]^{m-2}$，都有以下性质，则称其在子集 $J \subseteq \{1, \ldots, m\}$ 上满足PPRDS性质：

$$P(\mathbf{P}_{\hat{i},\hat{j}} \in U | P_i = x_i, P_j = x_j)$$

is an 个关于 x_i 和 x_j 的单调递增函数，其中 $\mathbf{P}_{\hat{i},\ \hat{j}}$ 是取值于 p 且满足 P_i 的向量$\mathbf{P}$，并且 P_j 剔除。

定理 3。对于 *A.1* 节中的输入，假设保守的 p 值在真实零假设集合上满足 *PPRDS* 条件，且规模函数由成对权重之和所界定，即 $(\forall J \subseteq \{1, \ldots, m\})\left(\sigma\left(J\right) \leqslant \sum_{i,j\in J} w_{ij}\right)$，其中 $w_{ij} \geqslant 0$。那么，阈值为 $\alpha = \frac{\sum_{i,\ j\in I_N} w_{ij}}{q}$ 的广义线性逐步上升法可在水平 q 控制广义错误发现率。

事实上，定理 2 和定理 3 是一个更一般定理的特例，其中我们对$\sigma\left(J\right) \leqslant \sum_{J'\subseteq J} w_{J'}$ 进行界定；这里，$w_{J'} \geqslant 0$ 是分配给 $\{1, \ldots, m\}$所有子集的权重。在 w_J 仅对单元素子集取非零值的特殊情形下，这等价于定理 2；而若 w_J 仅对大小不超过 2 的集合取非零值，则得到定理 3。[这种情形有一点微妙之处：定理 3 并不要求 w_{ij} 是对称的（即 $w_{ij} = w_{ji}$）。不过，不难证明，在定理 3 中，我们可以用对称权重$w'_{ij} = \frac{1}{2}(w_{ij} + w_{ji})$替换 w_{ij}。]

因此，我们只需证明这个更一般的定理，我们在定理中陈述该定理 4 below.

定义 *3*（*SPRDS*）。若随机变量集合 $P = (P_1, \ldots, P_m) \in [0,1]^m$ 满足如下子集上的集合式正回归依赖（*SPRDS*）条件：对于子集 $\{1, \ldots, m\}$ 的集合 $\mathcal{S} \subseteq \mathcal{P}$，任意 $C \in \mathcal{S}$、任意满足 $(\forall i \in C)(v_i \leqslant w_i)$ 的 $v, w \in [0,1]^{|C|}$，以及任意向上闭 $U \subseteq [0,1]^{m-|C|}$，均有 $P(P_{\hat{C}} \in U | P_C = v) \leqslant P(P_{\hat{C}} \in U | P_C = w)$；其中，$P_{\hat{C}}$表示移除指标 C 后的 P，P_C 表示将 P 限制在 C 上所得的结果。

不难看出，PRDS 和 PPRDS 是 SPRDS 条件的特例：分别令 $\mathcal{S}$ 为单元素集的集合，或单元素集与二元组的集合即可。

定理 4. 对于 *A.1* 节中的输入，假设规模函数由$\sigma\left(J\right) \leqslant \sum_{J'\subseteq J} w_{J'}$ 界定，其中 $w_{J'} \geqslant 0$ 是 $\{1, \ldots, m\}$子集上的权重。另假设保守的 p-值 $\mathbf{P}$ 满足针对具有非零权重的 I_N 子集集合的 *SPRDS* 条件。那么，阈值为 $\alpha = \frac{W}{q}$ 的广义线性逐步上升法（其中$W = \sum_{J\subseteq\{1,\ldots,m\}} w_J$）可将 *gFDR* 控制在水平 q。

证明。对于标量 x，我们将用 $\mathrm{P}_C \leqslant x$ 表示 $(\forall i \in C)(P_i \leqslant x)$。令 $S = I_{c_{\max}(\mathbf{P})}(\mathbf{P})$ 表示具有较小 p 值的被拒绝假设集合。令 $K = \sigma\left(S\right)$ 为所有被拒绝零假设集合的规模函数，$V = \sigma\left(S \cap I_N\right)$ 为最终被拒绝的真实零假设集合的规模函数。与 Benjamini 和 Yekutieli（2001）的证明一样，对于任意满足 $w_C > 0$ 的 $C \subseteq I_N$、任意 $x \in \mathbb{R}^+$ 以及任意 $\mathrm{v}: C \longrightarrow [0,1]$，SPRDS 表明 $P(K \leqslant x | \mathrm{P}_C = \mathrm{v})$ 是关于 x 和 v 的单调递增函数。错误发现率为

$$\begin{aligned}\mathbb{E}(Q) &= \mathbb{E}\left(\frac{V}{K}\right)\\ &= \mathbb{E}\left(\frac{\sigma\left(\overline{S}\cap I_N\right)}{K}\right)\\ &\leqslant \mathbb{E}\left(\frac{\sum_{J\subseteq S\cap I_N} w_J}{K}\right)\\ &= \mathbb{E}\left(\frac{\sum_{J\subseteq I_N} w_J 1_{J\subseteq S}}{K}\right)\\ &= \sum_{J\subseteq I_N} w_J \mathbb{E}\left(\frac{1_{J\subseteq S}}{K}\right)\end{aligned}$$

现在，根据 $c_{\max}(\mathbf{P})$ 和 S 的定义，我们有 $\sigma\left(S\right) = \alpha c_{\max}(\mathbf{P})$。当且仅当 $P_i \leqslant c_{\max}(\mathbf{P}) = \frac{K}{\alpha}$ 时，$i \in S$ 成立。因此

$$\mathbb{E}(Q) \leqslant \sum_{J\subseteq I_N} w_J \mathbb{E}\left(\frac{1_{J\subseteq S}}{K}\right) = \sum_{J\subseteq I_N} w_J \mathbb{E}\left(\frac{1_{(\mathbf{P}_J \leqslant \frac{K}{\alpha})}}{K}\right)$$

由于 $K = \sigma\left(S\right)$，其中 $S \subseteq \{1,\ldots,m\}$，$K$ 只有有限多个可能取值。令这些可能取值为 $k_0 \leqslant k_1 \leqslant \cdots \leqslant k_M$。因此，

$$\begin{aligned}\mathbb{E}\left(\frac{1_{\mathbf{P}_J \leqslant \frac{K}{\alpha}}}{K}\right) &= \sum_{l=0}^{M} \frac{P\left(K = k_l \cap \left(\mathbf{P}_J \leqslant \frac{k_l}{\alpha}\right)\right)}{k_l}\\ &= \sum_{l=0}^{M} \frac{P\left(\mathbf{P}_J \leqslant \frac{k_l}{\alpha}\right)}{k_l}\left(P\left(K \leqslant k_l \middle| \mathbf{P}_J \leqslant \frac{k_l}{\alpha}\right) - P\left(K \leqslant k_{l-1} \middle| \mathbf{P}_J \leqslant \frac{k_l}{\alpha}\right)\right)\end{aligned}$$

对于任意 l，p 值的集合 $\{\mathbf{P}|K(\mathbf{P}) \leqslant k_l\}$ 是一个上闭集，因此根据 SPRDS性质，

$$P\left(K \leqslant k_{l-1} \middle| \mathbf{P}_J \leqslant \frac{k_l}{\alpha}\right) \geqslant P\left(K \leqslant k_{l-1} \middle| \mathbf{P}_J \leqslant \frac{k_{l-1}}{\alpha}\right)$$

因此，

$$\mathbb{E}\left(\frac{1_{\mathbf{P}_J \leqslant \frac{K}{\alpha}}}{K}\right) \leqslant \sum_{l=0}^{M} \frac{P\left(\mathbf{P}_J \leqslant \frac{k_l}{\alpha}\right)}{k_l}\left(P\left(K \leqslant k_l \middle| \mathbf{P}_J \leqslant \frac{k_l}{\alpha}\right) - P\left(K \leqslant k_{l-1} \middle| \mathbf{P}_J \leqslant \frac{k_{l-1}}{\alpha}\right)\right)$$

由于$\alpha = \frac{W}{q}$和 P_i都是超均匀的，因此$P\left(\mathbf{P}_J \leqslant \frac{k_l}{\alpha}\right) \leqslant \frac{\binom{k_l}{\alpha}}{k_l} = \frac{1}{\alpha} = \frac{q}{W}$。令$\pi_l^J = P\left(K \leqslant k_l \middle| \mathbf{P}_J \leqslant \frac{k_l}{\alpha}\right)$ 和 $\pi_{-1}^l = 0$，可得

$$\mathbb{E}\left(\frac{1_{\mathbf{P}_J \leqslant \frac{K}{\alpha}}}{K}\right) \leqslant \sum_{l=0}^{M} \frac{q}{W}\left(\pi_l^J - \pi_{l-1}^J\right) = \frac{q}{W}\pi_M^J = \frac{q}{W}$$

因此，

$$\mathbb{E}(Q) \leqslant \sum_{J \subseteq I_N} w_J \mathbb{E}\left(\frac{1_{(\mathbf{P}_J \leqslant \frac{K}{\alpha})}}{K}\right) \leqslant \sum_{J \subseteq I_N} w_J \frac{q}{W} = q$$

□

容易看出，在定理 4 中令 w_J 除单元素集外均为零，即可得到定理 2。下面简要说明如何由定理 4 证明定理 3。

定理 *3* 的证明。对于 $J' \subseteq \{1, \ldots, m\}$，令

$$W_{J'} = \begin{cases} w_{ii} & \text{if } J' = \{i\} \\ w_{ij} + w_{ji} & \text{if } J' = \{i, j\} \\ 0 & \text{otherwise} \end{cases}$$

可以看到，对于任意 $J \subseteq \{1, \ldots, m\}$,

$$\sigma\left(J\right) \leqslant \sum_{i,j \in J} w_{ij} = \sum_{i \in J} w_{ii} + \sum_{i<j \in J} \left(w_{ij} + w_{ji}\right) = \sum_{i \in J} W_{\{i\}} + \sum_{i<j \in J} W_{\{i,j\}} = \sum_{J' \subseteq J} W_{J'}$$

此外，对于所有大小大于 2 的子集， $W_{J'}$ 均为零，因此，关于 I_N的 PPRDS 条件蕴含了对大小至多为 2 的 I_N 子集成立的 SPRDS 条件。因此，由定理 4 可知，采用 $\alpha = \frac{\sum_{J \subseteq \{1,\ldots,m\}} W_J}{q}$ 的 GLSUP 能以 q为水平控制广义错误发现率。我们有

$$\sum_{J \subseteq \{1,\ldots,m\}} W_J = \sum_{i=1}^{m} w_{ii} + \sum_{i<j \in \{1,\ldots,m\}} \left(w_{ij} + w_{ji}\right) = \sum_{i,j=1}^{m} w_{ij}$$

□

A.3. 证明 SHREDDER 的规模函数可由 $I \times I$上的成对权重之和界定

我们曾解释，修正的 GLSUP 的一个动机是它使我们能够用$\sigma\left(J\right) \leqslant \sum_{i,\ j \in J} w_{ij}$ 来界定规模函数 σ： $\Sigma \to \mathbb{R}_{\geqslant 0}$，其中权重为 $w_{ij} \geqslant 0$。本节将进一步说明这一点如何应用。

A.3.1. 定理 4 的应用

我们首先考察定理 4 适用的情形。为了应用定理 4，我们需要确定何时规模函数的一个上界可以表示为子集上的权重和。也就是说，若令 $\phi(A) \geqslant \sigma\left(A\right)$，何时可以将 $\phi(A) = \sum_{A' \subset A} w_{A'}$ 表示为权重 $w_{A'} \geqslant 0$? 事实证明，这可以简洁地表述为满足容斥不等式的函数。

命题 1。设 $\mathcal{R}$ 是 $\{1, \ldots, m\}$ 的一个子集集合，它对并和交封闭。递增集合函数 ϕ： $\mathcal{R} \to \mathbb{R}_{\geqslant 0}$ 可表示为 $\phi(A) = \sum_{A' \subseteq A} w_{A'}$，其中权重为 $w_{A'} \geqslant 0$，当且仅当 ϕ 满足容斥不等式：

$$\phi\left(\bigcup_{n=1}^{k} A_n\right) \geqslant \sum_{\emptyset \neq K \subseteq \{1,\ldots,k\}} (-1)^{|K|+1} \phi\left(\bigcap_{m \in K} A_m\right)$$

对于所有集合 A_1，.....， $A_k \in \mathcal{R}$。

此外，在这种情况下，我们有 $w_A = 0$ 只要对于 A的某个分解，不等式取等号，即存在非平凡分解 $A = A_1 \cup \cdots \cup A_k$，使得

$$\phi(A) = \sum_{\emptyset \neq K \subseteq \{1,\ldots,k\}} (-1)^{|K|+1} \phi\left(\bigcap_{m \in K} A_m\right)$$

这等价于以下条件

$$\phi(A) = -\sum_{\substack{A' \in \mathcal{R} \\ A' \subsetneq A}} \mu(A, A') \phi(A')$$

其中， μ 是偏序集 $\{A'' \in \mathcal{R} | A' \subseteq A'' \subseteq A\}$上的莫比乌斯函数，递归定义为 $\mu(A, A) = 1$，且 $\mu(A, B) = -\sum_{A' \in \mathcal{R},\ B \subsetneq A' \subseteq A} \mu(A, A')$。.

证明。 首先，若对于某些权重 $w_{A'} \geqslant 0$， $\phi(A) = \sum_{A' \subseteq A} w_{A'}$，则

$$\begin{aligned}
\phi\left(\bigcup_{n=1}^{k} A_n\right) &= \sum_{\substack{A' \in \mathcal{R} \\ A' \subseteq \bigcup_{n=1}^{k} A_n}} w_{A'} \\
&\geqslant \sum_{\substack{A' \in \mathcal{R} \\ (\exists i \in 1,\ldots,k)(A' \subseteq A_i)}} w_{A'} \\
&= \sum_{A' \in \mathcal{R}} w_{A'} \left(\sum_{\substack{(\emptyset \neq K \subseteq \{1,\ldots,k\}) \\ (\forall i \in K)(A' \subseteq A_i)}} (-1)^{|K|+1} \right) \\
&= \sum_{A' \in \mathcal{R}} w_{A'} \left(\sum_{\emptyset \neq K \subseteq \{1,\ldots,k\}} (-1)^{|K|+1} 1_{A' \subseteq \bigcap_{i \in K} A_i} \right) \\
&= \sum_{\emptyset \neq K \subseteq \{1,\ldots,k\}} (-1)^{|K|+1} \left(\sum_{\substack{A' \in \mathcal{R} \\ A' \subseteq \bigcap_{i \in K} A_i}} w_{A'} \right) \\
&= \sum_{\emptyset \neq K \subseteq \{1,\ldots,k\}} (-1)^{|K|+1} \phi\left(\bigcap_{i \in K} A_i\right)
\end{aligned}$$

反过来，假设 ϕ 满足容斥不等式。我们要证明存在适用于 $A \in \mathcal{R}$ 的权重 $w_A \geqslant 0$，使得对所有 $A \in \mathcal{R}$， $\phi(A) = \sum_{A' \subseteq A} w_{A'}$。我们回顾偏序中的标准结果的证明： w_A 由偏序集 $\mathcal{R}$上的莫比乌斯反演公式给出。

$$w_A = \sum_{\substack{A' \in \mathcal{R} \\ A' \subseteq A}} \mu(A, A') \phi(A')$$

其中， $\mu(A, A')$ 是偏序集 $\{A'' \in \mathcal{R} | A' \subseteq A'' \subseteq A\}$上的莫比乌斯函数，递归定义为 $\mu(A, A) = 1$，且 $\mu(A, B) = -\sum_{A' \in \mathcal{R}} \mu(A, A')$。证明是直接的-$^{B \subsetneq A' \subseteq A}$

接下来验证这些 $w_{A'}$ 对每个 $A \in \mathcal{R}$都满足所需的方程，首先用归纳法证明，对于 $A' \subseteq A$，

$$\sum_{A' \subseteq A'' \subseteq A} \mu(A'', A') = \begin{cases} 1 & \text{if } A' = A \\ 0 & \text{otherwise} \end{cases}$$

这显然对 $A' = A$成立，而对于 $A' \subsetneq A$，我们有

$$\begin{aligned} \sum_{A' \subseteq A'' \subseteq A} \mu(A'', A') &= \mu(A', A') + \sum_{A' \subsetneq A'' \subseteq A} \left(- \sum_{A' \subsetneq A''' \subseteq A''} \mu(A'', A''') \right) \\ &= 1 - \sum_{A' \subsetneq A''' \subseteq A} \sum_{A''' \subseteq A'' \subseteq A} \mu(A'', A''') \\ &= 1 - \sum_{A' \subsetneq A''' \subseteq A} 1_{A'''=A} \\ &= 0 \end{aligned}$$

利用这一点，我们证明上述定义的 w_A 满足所需条件:

$$\begin{aligned} \sum_{\substack{A' \in \mathcal{R} \\ A' \subseteq A}} w_{A'} &= \sum_{\substack{A' \in \mathcal{R} \\ A' \subseteq A}} \sum_{\substack{A'' \in \mathcal{R} \\ A'' \subseteq A'}} \mu(A', A'') \phi(A'') \\ &= \sum_{\substack{A'' \in \mathcal{R} \\ A'' \subseteq A}} \phi(A'') \sum_{\substack{A' \in \mathcal{R} \\ A'' \subseteq A' \subseteq A}} \mu(A', A'') \\ &= \sum_{\substack{A'' \in \mathcal{R} \\ A'' \subseteq A}} \phi(A'') 1_{A''=A} \\ &= \phi(A) \end{aligned}$$

现在只需证明 w_A 对所有 $A \in \mathcal{R}$均为非负。然而，由分解式 $A \supseteq \bigcup \{A' \in \mathcal{R} | A' \subsetneq A\}$可得

$$\phi(A) \geqslant \sum_{\emptyset \neq \mathcal{S} \subseteq \{A' \in \mathcal{R} | A' \subsetneq A\}} (-1)^{|\mathcal{S}|+1} \phi\left(\bigcap \mathcal{S}\right) = \sum_{\substack{A' \in \mathcal{R} \\ A' \subsetneq A}} \phi(A') \left(\sum_{\substack{\emptyset \neq \mathcal{S} \subseteq \{A'' \in \mathcal{R} | A'' \subsetneq A\} \\ \bigcap \mathcal{S} = A'}} (-1)^{|\mathcal{S}|+1} \right)$$

我们可以通过归纳法证明 $\sum_{\mathcal{S} \subseteq \{A'' \in \mathcal{R} | A'' \subsetneq A\}} (-1)^{|\mathcal{S}|} = \mu(A, A')$，其中求和包括空集，且在此语境下，其交集定义为 A。显然，当

$A' = A$。假设这对所有 $A' \subsetneq A'' \subseteq A$都成立。我们有

$$\begin{aligned}
\mu(A, A') &= -\sum_{A' \subsetneq A'' \subseteq A} \mu(A, A'') \\
&= -\sum_{A' \subsetneq A'' \subseteq A} \left(\sum_{\substack{\mathcal{S} \subseteq \{A''' \in \mathcal{R} | A''' \subsetneq A\} \\ \bigcap \mathcal{S} = A''}} (-1)^{|\mathcal{S}|} \right) \\
&= \sum_{\substack{\mathcal{S} \subseteq \{A''' \in \mathcal{R} | A''' \subsetneq A\} \\ A' \subsetneq \bigcap \mathcal{S}}} (-1)^{|\mathcal{S}|+1} \\
&= \sum_{\mathcal{S} \subseteq \{A''' \in \mathcal{R} | A' \subseteq A''' \subsetneq A\}} (-1)^{|\mathcal{S}|+1} - \sum_{\substack{\mathcal{S} \subseteq \{A''' \in \mathcal{R} | A' \subseteq A''' \subsetneq A\} \\ \bigcap \mathcal{S} = A'}} (-1)^{|\mathcal{S}|+1} \\
&= \sum_{\substack{\mathcal{S} \subseteq \{A''' \in \mathcal{R} | A' \subseteq A''' \subsetneq A\} \\ \bigcap \mathcal{S} = A'}} (-1)^{|\mathcal{S}|}
\end{aligned}$$

那么，这种分解的容斥不等式给出 $w_A \geqslant 0$；当且仅当 $\phi(A) = -\sum_{A' \in \mathcal{R}} \mu(A, A')\phi(A')$ 时取等号。由上文可知，这等价于等式$^{A' \subsetneq A}$

$$\phi(A) = \sum_{\emptyset \neq \mathcal{S} \subseteq \{A' \in \mathcal{R} | A' \subsetneq A\}} (-1)^{|\mathcal{S}|+1} \phi\left(\bigcap \mathcal{S}\right)$$

这就是分解 $A = \bigcup \{A' \in \mathcal{R} | A' \subsetneq A\}$的容斥等式。我们希望证明：如果任意分解 $A = A_1 \cup \cdots \cup A_n$ 的容斥不等式取等号，那么对于这个极大分解，该不等式也取等号。令 $A = A_1 \cup \cdots \cup A_n$ 满足

$$\phi(A) = \sum_{\emptyset \neq S \subseteq \{1, \ldots, n\}} (-1)^{|S|+1} \phi\left(\bigcap_{i \in S} A_i\right)$$

我们将证明，对于任意 $A_{n+1} \subseteq A$，都有

$$\phi(A) = \sum_{\emptyset \neq S \subseteq \{1, \ldots, n+1\}} (-1)^{|S|+1} \phi\left(\bigcap_{i \in S} A_i\right)$$

根据容斥不等式，我们有

$$\begin{aligned}
\phi(A) &\geqslant \sum_{\emptyset \neq S \subseteq \{1, \ldots, n+1\}} (-1)^{|S|+1} \phi\left(\bigcap_{i \in S} A_i\right) \\
&= \sum_{\emptyset \neq S \subseteq \{1, \ldots, n\}} (-1)^{|S|+1} \phi\left(\bigcap_{i \in S} A_i\right) + \phi(A_{n+1}) - \sum_{\emptyset \neq S \subseteq \{1, \ldots, n\}} (-1)^{|S|+1} \phi\left(A_{n+1} \cap \bigcap_{i \in S} A_i\right) \\
&= \phi(A) + \phi(A_{n+1}) - \sum_{\emptyset \neq S \subseteq \{1, \ldots, n\}} (-1)^{|S|+1} \phi\left(\bigcap_{i \in S} (A_{n+1} \cap A_i)\right) \\
&\geqslant \phi(A)
\end{aligned}$$

因此，这个分解的容斥不等式也是等式。将 A 的所有子集都加入该分解后，我们可以得出：如果存在某个分解使容斥不等式成为等式，那么对于作为 A 所有真子集之并的最大分解，该不等式也成立为等式。

□

回顾一下，mGLSUP 从 A.1 节给出的数据出发，但对 $\mathcal{R}$ 作了特定选择，从而在修改 p 值之后，可以保证对所有 $c \in [0,1]$ 都有 $I_c \in \mathcal{R}$。在修改 p 值之后，该方法的后续步骤与 GLSUP 完全相同。我们关注的是假设被排列成层级二叉树森林的特定情形，不过我们可以在稍更一般的背景下证明该结果。我们首先回顾一些基本的序理论：

定义 4. 对于偏序集 $(X, \Leftarrow)$，若 $y \Leftarrow x$，且对任意满足 $y \Leftarrow z \Leftarrow x$ 的 $z \in X$，我们有 $z = x$ 或 $z = y$，则称元素 $x \in X$ 覆盖 元素 $y \in X$（记为 $y \prec x$）。例如，若 $(X, \Leftarrow)$ 具有树结构，则一个节点仅覆盖其子节点。

我们使用 $\Leftarrow$ 而不是 $\leqslant$，因为在我们的启发性示例中，假设之间的偏序是逆蕴含。我们可以证明，对于每个元素至多覆盖另外两个元素的偏序，我们的结果成立；可拒绝假设集合由相对于该偏序的上闭集（上集）组成。我们将范围限制在 $\sigma(S) = \sum_{i \in \mathrm{Min}(S)} \phi_i$ 的情形，其中权重为 $\phi_1, \ldots, \phi_m \geqslant 0$，$\mathrm{Min}(S)$ 表示 S 中的极小元素集合。我们假设这些 ϕ_i 满足：对于任意 $A \subseteq B$，$\sigma(A) \leqslant \sigma(B)$。也就是说，$\phi_i \geqslant \sum_{H_i \prec H_j} \phi_j$。

命题 2。 设 $\Leftarrow$ 是假设集合 $H_1, \ldots, H_m$ 上的一个偏序。令 $H_i \prec H_j$ 表示该偏序的覆盖关系。假设对于任意 i，至多存在两个 $j \in \{1, \ldots, m\}$ 满足 $H_j \prec H_i$。

令 $\mathcal{U}$ 为关于偏序 $\Leftarrow$ 的 $\{H_1, \ldots, H_m\}$ 的上闭子集族。设 $\sigma : \mathcal{U} \to \mathbb{R}_{\geqslant 0}$ 由 $\sigma(A) = \sum_{H_i \in \mathrm{Min}(A)} \phi_i$ 给定，其中 $\phi_i \geqslant 0$ 为某个元素，并且满足 $\sigma(A) \leqslant \sigma(B)$，只要 $A \subseteq B$ 成立。于是存在权重 $w_{ij} \geqslant 0$，对所有 $i, j = 1, .., m$ 均成立，使得对任意 $A \in \mathcal{U}$、$\sigma(A) = \sum_{i,j \in A} w_{ij}$。

证明。 根据命题 1，如果我们能证明 σ 在上集上满足容斥不等式，那么就有 $\sigma(A) = \sum_{B \subseteq A} W_B$，其中 $W_B \geqslant 0$ 为某个元素。接着，如果我们能证明当 B 有三个或更多极小元时 $W_B = 0$ 成立，那么对于任意 i、j，令 $\uparrow(\{i, j\})$ 为由 i 和 j 生成的上集，并定义

$$w_{ij} + w_{ji} = \begin{cases} 0 & \text{if } H_i \Leftarrow H_j \text{ or } H_j \Leftarrow H_i \\ W_{\uparrow(\{i,j\})} & \text{otherwise} \end{cases}$$

以及 $w_{ii} = W_{\uparrow(\{i\})}$。由此可得

$$
\begin{aligned}
\sigma(A) &= \sum_{B\subseteq A} W_B \\
&= \sum_{\substack{B\subseteq A \\ |\operatorname{Min}(B)|\leqslant 2}} W_B \\
&= \sum_{\substack{B\subseteq A \\ (\exists i) B=\uparrow(\{i\})}} W_B \quad + \sum_{\substack{B\subseteq A \\ (\exists i,j) H_i\not\preccurlyeq H_j, H_j\not\preccurlyeq H_i, B=\uparrow(\{i,j\})}} W_B \\
&= \sum_{i\in A} w_{ii} + \sum_{i<j\in A} w_{ij} + w_{ji} \\
&= \sum_{i,j\in A} w_{ij}
\end{aligned}
$$

对于上集 $T_1,\ldots,T_n\in\mathcal{U}$，其并集不是 $T_1,\ldots,T_n$ 中的任何一个，我们有

$$
\begin{aligned}
w_{T_1\cup\cdots\cup T_n} &\leqslant \sigma(T_1\cup\cdots\cup T_n) + \sum_{\emptyset\neq A\subseteq\{1,\ldots,n\}} (-1)^{|A|}\sigma\left(\bigcap_{i\in A} T_i\right) \\
&= \sum_{H_i\in\operatorname{Min}(T_1\cup\cdots\cup T_n)} \phi_i \quad + \sum_{\emptyset\neq A\subseteq\{1,\ldots,n\}} (-1)^{|A|} \sum_{H_i\in\operatorname{Min}\left(\bigcap_{j\in A} T_j\right)} \phi_i \\
&= \sum_{H_i\in T_1\cup\cdots\cup T_n} \phi_i \left(\sum_{A\subseteq\{1,\ldots,n\}} (-1)^{|A|} 1_{H_i\in\operatorname{Min}\left(\bigcap_{j\in A} T_j\right)} \right)
\end{aligned}
$$

我们需要证明这始终是非负的，并且对于任意具有三个极小元的 U，存在一个分解 $U=T_1\cup\cdots\cup T_n$，其中 $T_i\neq U$，使得该和为零。我们对每个 H_i考虑项 $\sum_{A\subseteq\{1,\ldots,n\}}(-1)^{|A|}$。我们将证明

$$
\sum_{\substack{A\subseteq\{1,\ldots,n\} \\ H_i\in\operatorname{Min}\left(\bigcap_{j\in A} T_j\right)}} (-1)^{|A|} = \begin{cases} 1 & \text{if } (\exists i_1,i_2,j_1,j_2)(H_{i_1}\prec H_i, H_{i_2}\prec H_i, H_{i_1}\in T_{j_1}, H_{i_2}\in T_{j_2} \text{ and } (\forall j)\{H_{i_1},H_{i_2}\}\not\subseteq T_j) \\ 0 & \text{otherwise} \end{cases} \tag{4}
$$

也就是说，除非 H_i 覆盖另外两个位于不同上集中的假设，否则该和为零；在这种情况下，该和为1。由此立即可知，该界始终非负，即容斥不等式成立；再结合 σ 是单调递增函数，这意味着 σ 可以表示为非负权重 $\{W_U|U\in\mathcal{U}\}$之和。

假设 H_i、 H_j 和 H_k 是上集 $U\in\mathcal{U}$的不同极小元。则我们有$U=(U\setminus\{H_i\})\cup(U\setminus\{H_j\})\cup(U\setminus\{H_k\})$，并且我们有

$$
\begin{aligned}
W_U \leqslant\ & \sigma(U) - \sigma(U\setminus\{H_i\}) - \sigma(U\setminus\{H_j\}) - \sigma(U\setminus\{H_k\}) \\
& + \sigma(U\setminus\{H_i,H_j\}) + \sigma(U\setminus\{H_i,H_k\}) + \sigma(U\setminus\{H_j,H_k\}) - \sigma(U\setminus\{H_i,H_j,H_k\})
\end{aligned}
$$

由此可得

$$
W_U \leqslant \sum_{H_l\in U} \phi_l \sum_{A\subseteq\{H_i,H_j,H_k\}} (-1)^{|A|} 1_{H_l\in\min(U\setminus A)}
$$

由 (4) 可知，除非 H_l 覆盖 U 中的两个元素，且 $U\setminus\{H_i\}$、$U\setminus\{H_j\}$、$U\setminus\{H_k\}$ 中的每一个都排除其中一个元素，否则 $\sum_{A\subseteq\{H_i,H_j,H_k\}}(-1)^{|A|}1_{H_l\in\min(U\setminus A)}$ 为 0。这显然不可能。因此，对所有 l，$\sum_{A\subseteq\{H_i,H_j,H_k\}}(-1)^{|A|}1_{H_l\in\min(U\setminus A)}=0$，这意味着 $W_U=0$。因此，唯一非零的 W_U 是那些 U 至多有两个极小元的情况。

我们只需证明 (4)。显然，求和中只有满足 A 限定为满足 $H_i\in T_j$ 的 j 的项才非零。不失一般性，设对于某个 $l\leqslant n$，$\{j|H_i\in T_j\}=\{1,\ldots,l\}$。因此，我们要确定对于哪些 $A\subseteq\{1,\ldots,l\}$，H_i 是 $\bigcap_{j\in A}T_j$ 的极小元。由于 $\bigcap_{j\in A}T_j$ 向上闭，若 $\bigcap_{j\in A}T_j$ 不包含任何被 H_i 覆盖的 H_j，则 H_i 是极小元。我们知道，H_i 至多覆盖两个元素。我们分三种情况讨论：

如果 H_i 不覆盖 $\bigcup_{i=1}^{l}T_i$ 中的任何元素，那么它在所有包含它的下集中都是极小元 .

$$\sum_{A\subseteq\{1,\ldots,n\}}(-1)^{|A|}1_{H_i\in\mathrm{Min}\left(\bigcap_{j\in A}T_j\right)}=\sum_{A\subseteq\{1,\ldots,l\}}(-1)^{|A|}=0$$

如果 H_i 只覆盖一个 $H_j\in T_1\cup\cdots\cup T_l$，那么 H_i 在所有不包含 H_j 的交集中都是极小元。设 $B=\{k\in\{1,\ldots,l\}|H_j\notin T_k\}$。那么，$H_i$ 是 $\bigcap_{k\in A}T_k$ 的极小元，当且仅当 A 包含 B 中的一个元素。因此

$$\sum_{A\subseteq\{1,\ldots,n\}}(-1)^{|A|}1_{H_i\in\mathrm{Min}\left(\bigcap_{j\in A}T_j\right)}=\sum_{\substack{A\subseteq\{1,\ldots,l\}\\A\cap B\neq\emptyset}}(-1)^{|A|}=\sum_{A\subseteq\{1,\ldots,l\}}(-1)^{|A|}-\sum_{A\subseteq\{1,\ldots,l\}\setminus B}(-1)^{|A|}=0$$

（根据我们的假设，我们知道 $\{1,\ldots,l\}\setminus B\neq\emptyset$ $H_j\in T_1\cup\cdots\cup T_l$。）

如果 H_i 覆盖 $T_1\cup\cdots\cup T_l$ 中的两个元素 H_{i_1} 和 H_{i_2}，那么如果 H_i 不包含 H_{i_1} 或 H_{i_2} 中的任一个，它在一个交集中就是极小的。令 $B_1=\{k\in\{1,\ldots,l\}|H_{i_1}\notin T_k\}$ 和 $B_2=\{k\in\{1,\ldots,l\}|H_{i_2}\notin T_k\}$。当且仅当 A 包含 B_1 中的一个元素和 B_2 中的一个元素时，H_i 在 $\bigcap_{k\in A}T_k$ 中是极小的。因此

$$\begin{aligned}\sum_{A\subseteq\{1,\ldots,n\}}(-1)^{|A|}1_{H_i\in\mathrm{Min}\left(\bigcap_{j\in A}T_j\right)}&=\sum_{\substack{A\subseteq\{1,\ldots,l\}\\A\cap B_1\neq\emptyset,A\cap B_2\neq\emptyset}}(-1)^{|A|}\\&=\sum_{A\subseteq\{1,\ldots,l\}}(-1)^{|A|}-\sum_{A\subseteq\{1,\ldots,l\}\setminus B_1}(-1)^{|A|}-\sum_{A\subseteq\{1,\ldots,l\}\setminus B_2}(-1)^{|A|}+\sum_{A\subseteq\{1,\ldots,l\}\setminus(B_1\cup B_2)}(-1)^{|A|}\\&=\begin{cases}1 & \text{if } B_1\cup B_2=\{1,\ldots,l\}\\0 & \text{otherwise}\end{cases}\end{aligned}$$

事件 $B_1\cup B_2=\{1,\ldots,l\}$ 等价于不存在同时包含 H_{i_1} 和 H_{i_2} 的 T_j。

□

注 2。上述证明乍看之下可能表明，$w_U=0$ 只要 U 是一个仅有一个极小元素的上集。然而，如果 U 只有一个极小元素，它就不能表示为上集的非平凡并集，因此我们不使用容斥不等式，而是利用 σ 是单调递增函数这一事实，并且有 $w_U=\sigma(U)-\sigma(U\setminus\{H_i\})$，其中 H_i 是 U 的唯一极小元素。

A.4. SHRED 与 SHREDDER 的截断值

在本节中，我们将 A.2节中的定理应用于 SHRED 和 SHREDDER 方法所产生的错误发现率控制问题，为其中的截断值提供具体取值。

推论 *1.* 设零假设 $H_1, \ldots, H_m$ 按逆蕴含关系构成偏序，并设拒绝假设子集 $J \subseteq \{1, \ldots, m\}$ 所产生的发现数量由规模函数 $\sigma(J) \leqslant \sum_{i \in Min(J)} \phi_i$ 表示，其中 $\phi_i \geqslant 0$ 是假设的权重，$Min(J)$ 表示 J 中按逆蕴含关系定义的极小假设集合（即不蕴含 J 中任何其他假设的零假设）。设 $I_N \subseteq \{1, \ldots, m\}$ 为真实零假设的指标集合。令 $p = \sum_{i \in Min(\{1,\ldots,m\})} \phi_i$ 且 $r_0 = \sum_{i \in I_N} \phi_i$。对于任意 $q > 0$，采用 $\alpha_{SHRED} = \frac{r_0(1+log(p)) - \sum_{i \in I_N} \phi_i log(\phi_i)}{q}$ 的广义线性逐步向上程序，可将广义错误发现率控制在水平 q。

证明。 由定理1，当我们使用截断值

$$\alpha_{\text{SHRED}} = \frac{\left(\sum_{i \in I_N} \phi_i\right)(1 + \log(\sigma(\{1, \ldots, m\}))) - \sum_{i \in I_N} \phi_i \log(\sigma_i)}{q} = \frac{r_0(1 + \log(p)) - \sum_{i \in I_N} \phi_i \log(\phi_i)}{q}$$

□

推论 2。 设零假设 $H_1, \ldots, H_m$ 按逆蕴含关系构成偏序，并设拒绝一组假设 $J \subseteq \{1, \ldots, m\}$ 后所得发现的数量由规模函数 $\sigma(J) \leqslant \sum_{i \in \text{Min}(J)} \phi_i$ 给出——其中 $\phi_i \geqslant 0$ 是赋予假设的权重，且 $\text{Min}(J)$ 表示在逆蕴含关系下 J 中的极小假设集合。设 $I_N \subseteq \{1, \ldots, m\}$ 为真实零假设的指标集合。假设保守的p值在真实零假设集合上满足 *PRDS* 条件。对任意 $q > 0$，采用 $\alpha_{SHRED_{PRDS}} = \frac{\sum_{i=1}^{m} \phi_i}{q}$ 的广义线性逐步向上程序可将 *gFDR* 控制在水平 q。

证明。 这可直接由定理 2 得出。 □

推论 3。 设零假设 $H_1, \ldots, H_m$ 在分叉结构中按逆蕴含关系构成偏序（即每个假设至多覆盖另外两个假设），并设拒绝假设子集 $J \subseteq \{1, \ldots, m\}$ 所产生的发现数量由规模函数 $\sigma(J) = \sum_{i \in \text{Min}(J)} \phi_i$ 给出，其中 $\phi_i \geqslant 0$ 是赋予假设的权重，而 $\text{Min}(J)$表示在逆蕴含关系下 J中的极小假设集合。设 $I_N \subseteq \{1, \ldots, m\}$ 为真实零假设的指标集合。假设经修正的 p值 $\{P_i' | i = 1, \ldots, m\}$ 在真实零假设的指标集合上满足 *PPRDS* 条件。对于任意 $q > 0$，采用 $\alpha_{SHREDDER} = \sigma_{(1,\ldots,m)}\, q$ 的广义线性逐步向上程序，可将广义错误发现率控制在水平 q。

证明。 这直接由定理 3 和命题 2 得出。 □

B. SHRED 的经验广义错误发现率控制

逐步上升法的错误发现率控制证明考虑的是最坏情形。在实践中，错误发现率控制通常要严格得多。此外，理论截断值依赖于未知的真实零假设，而在实践中，为了选择截断值，我们使用全部零假设以获得一个上界。我们开展一项实证研究，以确定典型情形下的实际错误发现率。为此，我们生成了 100 个模拟数据集，维度固定为$p = 300$，样本量为$n = 1000$。每个数据集抽取自 $X \sim N(\mu, \Sigma)$，其中 $\mu = (0, \ldots, 0)^\top$，$\Sigma_{ij} = \begin{cases} 1 & \text{若 } 0.5 \\ & \text{否则} \end{cases}$。高斯响应由随机选取的 $T = 100$个真实预测变量构造，相应的回归系数 β_i 针对 $i = 1, \ldots, T$ 独立地从标准正态分布中抽样。在该模拟设计下，我们评估了 SHRED 在多种阈值斜率 α取值下的经验FDR。

SE gPower	MSE	SE MSE
0.5%	1.14	0.08
0.5%	1.14	0.08
0.5%	1.14	0.09
0.5%	1.15	0.08
0.4%	1.15	0.08
0.5%	1.15	0.09
0.4%	1.16	0.09
0.4%	1.16	0.09
0.4%	1.17	0.08

tern. Values in the table are an

从表 1 可以看出，在这次模拟中，经验FDR与根据真实零假设数量计算的名义启发式FDR非常接近，也接近 PRDS FDR，但略低于后者。这表明，对于这些模拟，PRDS假设可能成立，或近似成立；经验FDR略低于根据真实零假设计算的名义FDR。由于我们不知道真实零假设的数量，而是用零假设总数作为上界，因此经验FDR远低于名义水平。这次模拟中，基于真实零假设计算的启发式方法和 PRDS 截断值之所以接近，是因为在随机选取的 100 个真实变量中，大小大于 1 的集合绝大多数都至少包含一个真实变量，因此并非真实零假设。由于几乎所有真实零假设都位于层次树的最低层，我们有 $\sum_{i\in I_N}\phi_i \approx p_0$。

C. 模拟表

模拟图的补充详情。

C.1. 高斯回归

表2： 高斯回归， $n=1000, p=300, T=100$，独立预测变量

方法	FDR 截止值	检验效能	检验效能标准误	FDR	假发现率标准误	MSE	均方误差标准误
套索回归	无 FDR 截断值	97.7%	0.2%	55.6%	0.5%	1.1016	0.0723
镜像	MDS 数据驱动	90.6%	0.9%	3.2%	0.2%	1.2097	0.0922
镜像	数据驱动 数据驱动	87.2%	0.9%	2.3%	0.3%	1.3336	0.0713
Knockoff	数据驱动	92.0%	0.3%	3.8%	0.3%	1.2155	0.0694
BH	PRDS	92.9%	0.3%	3.4%	0.2%	1.1159	0.0634
SHRED	PRDS	92.7%	0.3%	3.3%	0.2%	1.1261	0.0621
SHRED启发式算法	启发式方法	93.2%	0.4%	4.1%	0.3%	1.1201	0.0594
SHREDDER	PPRDS	92.8%	0.3%	3.3%	0.2%	1.1260	0.0687
SHRED	任意	90.8%	0.3%	0.4%	0.1%	1.1545	0.0701
BY	任意	90.7%	0.3%	0.6%	0.1%	1.1595	0.0713

表3： 高斯回归， $n=1000, p=300, T=100$,单一公共成分相关性 $\rho=0.5$

方法	FDR 截止值	检验效能	检验效能标准误	FDR	假发现率标准误	MSE	均方误差标准误
套索回归	无 FDR 截断值	96.3%	0.4%	53.1%	0.3%	1.1822	0.0311
镜像	MDS（数据驱动）	85.9%	0.6%	3.8%	0.4%	1.1009	0.0623
镜像	DS（数据驱动）	81.6%	0.4%	3.8%	0.6%	1.1343	0.0496
Knockoff	数据驱动	84.9%	0.8%	3.5%	0.5%	1.1138	0.0430
BH	PRDS	86.7%	0.6%	3.6%	0.5%	1.1251	0.0297
SHRED	PRDS	88.6%	0.4%	1.2%	0.1%	1.1273	0.0187
SHRED	启发式方法	90.0%	0.5%	4.4%	0.3%	1.1202	0.0118
SHREDDER	PPRDS	88.4%	0.5%	1.9%	0.2%	1.1214	0.0114
SHRED	任意	85.5%	0.5%	0.1%	0.09%	1.1337	0.0079
BY	任意	83.2%	0.4%	1.1%	0.3%	1.1396	0.0897

表 4： 高斯回归， $n = 1000$、 $p = 300$、 $T = 100$、单一公共成分相关性 $\rho = 0.9$

方法	FDR 截止值	检验效能	检验效能标准误	FDR	假发现率标准误	MSE	均方误差标准误
套索回归	无 FDR 截断值	94.8%	0.06%	50.7%	2.8%	1.1701	0.0302
镜像	MDS（数据驱动）	70.5%	0.8%	3.3%	0.4%	1.1225	0.0413
镜像	DS（数据驱动）	60.1%	1.5%	2.6%	0.4%	1.3498	0.0524
Knockoff	数据驱动	67.3%	1.1%	2.1%	0.4%	1.2525	0.0407
BH	PRDS	74.6%	0.9%	3.8%	0.9%	1.1313	0.0093
SHRED	PRDS	75.9%	0.7%	1.5%	0.1%	1.1276	0.0119
SHRED	启发式方法	78.5%	0.7%	3.7%	0.3%	1.1101	0.0095
SHREDDER	PPRDS	76.9%	0.7%	2.3%	0.2%	1.1274	0.0113
SHRED	任意	69.8%	0.6%	0.1%	0.05%	1.1522	0.0214
BY	任意	68.4%	0.7%	0.9%	0.4%	1.1449	0.0103

表 5： 高斯回归， $n = 1000, p = 300, T = 100$， 聚类相关 $\rho = 0.3$ to$\rho = 0.6$

方法	FDR 截止值	检验效能	检验效能标准误	FDR	假发现率标准误	MSE	均方误差标准误
套索回归	无 FDR 截断值	84.4%	0.7%	65.2%	0.3%	1.3334	0.0287
镜像	MDS（数据驱动）	61.2%	1.0%	6.7%	0.9%	2.1191	0.0745
镜像	DS（数据驱动）	20.8%	1.2%	7.5%	1.0%	4.2313	0.2349
Knockoff	数据驱动	39.9%	1.5%	12.6%	1.2%	2.1899	0.1075
BH	PRDS	74.1%	0.5%	2.5%	0.5%	1.5908	0.0615
SHRED	PRDS	75.2%	0.5%	1.4%	0.5%	1.4875	0.0724
SHRED	启发式方法	78.3%	0.6%	3.1%	0.5%	1.4678	0.0828
SHREDDER	PPRDS	76.2%	0.7%	2.3%	0.4%	1.5199	0.0915
SHRED	任意	72.7%	0.6%	0.2%	0.5%	1.5308	0.0610
BY	任意	71.8%	0.7%	0.7%	0.1%	1.5646	0.0871

表 6： 高斯回归， $n = 1000$、 $p = 300$、 $T = 100$， 聚类相关相关性$\rho = 0.6$ 至 $\rho = 0.9$

方法	FDR 截止值	检验效能	检验效能标准误	FDR	假发现率标准误	MSE	均方误差标准误
套索回归	无 FDR 截断值	80.8%	0.4%	57.7%	0.4%	1.3087	0.0168
镜像	MDS（数据驱动）	52.3%	1.7%	14.3%	1.2%	3.6486	0.0823
镜像	DS（数据驱动）	14.1%	0.7%	15.0%	1.3%	6.5350	0.1427
Knockoff	数据驱动	39.4%	1.3%	12.3%	1.2%	3.3186	0.0909
BH	PRDS	68.4%	0.7%	2.1%	0.3%	1.6703	0.0525
SHRED	PRDS	73.9%	0.8%	2.3%	0.4%	1.5845	0.0848
SHRED	启发式方法	78.1%	0.8%	3.2%	0.5%	1.5703	0.0892
SHREDDER	PPRDS	74.8%	0.7%	2.9%	0.5%	1.5735	0.0738
SHRED	任意	67.8%	0.9%	0.3%	0.05%	1.6816	0.0498
BY	任意	65.4%	0.9%	1.3%	0.4%	1.7291	0.0627

表 7： 高斯回归， $n = 1000, p = 300, T = 100$， AR(1)相关性 $\rho = 0.6$

方法	FDR 截止值	检验效能	检验效能标准误	FDR	假发现率标准误	MSE	均方误差标准误
套索回归	无 FDR 截断值	96.1%	0.4%	52.1%	0.1%	1.2100	0.0311
镜像	MDS（数据驱动）	85.1%	0.3%	5.1%	0.5%	1.2278	0.0318
镜像	DS（数据驱动）	82.5%	0.6%	3.6%	0.4%	1.2831	0.0300
Knockoff	数据驱动	82.4%	0.6%	3.3%	0.5%	1.2980	0.0414
BH	PRDS	84.7%	0.9%	3.4%	0.6%	1.1922	0.0223
SHRED	PRDS	87.9%	0.4%	1.6%	0.1%	1.1583	0.0191
SHRED	启发式方法	89.9%	0.5%	3.2%	0.3%	1.1498	0.0175
SHREDDER	PPRDS	87.9%	0.4%	1.8%	0.3%	1.1815	0.0195
SHRED	任意	84.5%	0.5%	0.4%	0.05%	1.1648	0.0571
BY	任意	82.4%	0.8%	1.2%	0.3%	1.2131	0.0591

表 8： 高斯回归， $n = 1000, p = 300, T = 100$， AR(1)相关性 $\rho = 0.9$

方法	FDR 截止值	检验效能	检验效能标准误	FDR	假发现率标准误	MSE	均方误差标准误
套索回归	无 FDR 截断值	92.7%	0.4%	51.7%	0.4%	1.1550	0.0170
镜像	MDS（数据驱动）	74.3%	0.8%	5.8%	0.5%	1.1832	0.0287
镜像	DS（数据驱动）	68.6%	1.4%	2.5%	0.7%	1.3509	0.0622
Knockoff	数据驱动	58.0%	1.3%	1.1%	0.2%	2.2691	0.0714
BH	PRDS	73.7%	0.9%	3.7%	0.6%	1.2407	0.0204
SHRED	PRDS	74.9%	0.6%	1.6%	0.2%	1.2105	0.0213
SHRED	启发式方法	77.8%	0.6%	4.3%	0.4%	1.2017	0.0187
SHREDDER	PPRDS	75.5%	0.6%	2.8%	0.3%	1.2051	0.0117
SHRED	任意	72.6%	0.6%	0.1%	0.05%	1.2770	0.0213
BY	任意	69.4%	0.7%	0.9%	0.3%	1.4048	0.0198

C.2. 逻辑回归

表9： 逻辑回归， $n = 5000, p = 200, T = 100$,单一公共分量相关性$\rho = 0.5$

方法	FDR 截止值	检验效能	检验效能标准误	FDR	假发现率标准误	准确率	准确率标准误	LL
套索回归	无 FDR 截断值	96.8%	0.3%	42.9%	0.4%	92.1562%	0.3008%	-727.7
镜像	MDS（数据驱动）	76.6%	0.7%	3.0%	0.4%	90.8892%	0.3186%	-780.1
镜像	DS（数据驱动）	70.9%	0.7%	2.9%	0.5%	90.4663%	0.2590%	-840.5
Knockoff	数据驱动	37.6%	0.2%	1.4%	0.1%	87.9016%	0.1897%	-893.7
BH	PRDS	76.5%	0.6%	3.1%	0.3%	90.6461%	0.2985%	-769.8
SHRED	PRDS	77.3%	0.6%	1.3%	0.2%	91.1502%	0.1916%	-770.2
SHRED	启发式方法	81.6%	0.6%	4.7%	0.2%	91.6991%	0.3007%	-748.3
SHREDDER	PPRDS	78.1%	0.6%	3.3%	0.2%	91.2882%	0.1677%	-757.5
SHRED	任意	73.9%	0.6%	0.2%	0.07%	90.8568%	0.1481%	-788.0
BY	任意	73.9%	0.7%	0.7%	0.1%	90.6332%	0.2845%	-814.9

表10： 逻辑回归， $n = 5000, p = 200, T = 100$,单一公共成分相关性 $\rho = 0.9$

方法	错误发现率截止值	检验效能	检验效能标准误	FDR	假发现率标准误	准确率	准确率标准误	LL
套索回归	无 FDR 截断值	88.2%	0.7%	38.2%	0.4%	90.9191%	0.5907%	-927.9
镜像	MDS（数据驱动）	51.6%	0.9%	3.0%	0.5%	88.3088%	0.7118%	-1014.4
镜像	DS（数据驱动）	46.4%	0.9%	2.0%	0.3%	88.0213%	0.6993%	-1023.7
Knockoff	数据驱动	19.5%	0.2%	0.8%	0.1%	87.0981%	0.6033%	-1175.2
BH	PRDS	52.7%	0.6%	3.5%	0.4%	88.8882%	0.5762%	-1009.7
SHRED	PRDS	53.1%	0.5%	1.2%	0.2%	88.9727%	0.5944%	-1013.8
SHRED	启发式方法	58.9%	0.4%	4.5%	0.2%	89.5810%	0.6211%	-1003.4
SHREDDER	PPRDS	54.7%	0.6%	2.9%	0.3%	89.2822%	0.5677%	-1005.6
SHRED	任意	49.2%	0.6%	0.3%	0.1%	88.5779%	0.6115%	-1048.4
BY	任意	48.2%	0.7%	0.7%	0.2%	88.5112%	0.6869%	-1045.3

表 11： 逻辑回归， $n = 5000, p = 200, T = 100$， 聚类相关性 $\rho = 0.3$ 至$\rho = 0.6$

方法	FDR 截止值	检验效能	检验效能标准误	FDR	假发现率标准误	准确率	准确率标准误	LL
套索回归	无 FDR 截断值	91.5%	0.6%	39.7%	0.4%	91.8013%	0.2876%	-611.9
镜像	MDS（数据驱动）	60.3%	0.9%	4.7%	0.4%	90.9192%	0.1810%	-706.4
镜像	DS（数据驱动）	46.2%	0.8%	4.8%	0.2%	88.3909%	0.3223%	-765.3
Knockoff	数据驱动	13.1%	0.8%	0.2%	0.1%	81.1001%	0.5257%	-935.5
BH	PRDS	61.0%	0.6%	4.3%	0.4%	90.8461%	0.2223%	-703.6
SHRED	PRDS	63.2%	0.6%	1.4%	0.3%	91.2971%	0.2352%	-669.3
SHRED	启发式方法	65.8%	0.5%	4.2%	0.4%	91.7951%	0.1365%	-617.5
SHREDDER	PPRDS	65.2%	0.5%	3.4%	0.4%	91.4110%	0.2059%	-666.3
SHRED	任意	59.9%	0.5%	0.4%	0.03%	90.6456%	0.1986%	-712.5
BY	任意	58.3%	0.5%	0.7%	0.3%	89.5713%	0.3076%	-725.5

表12： 逻辑回归， $n = 5000$、 $p = 200$、 $T = 100$， 聚类相关 $\rho = 0.6$ 至$\rho = 0.9$

方法	FDR 截止值	检验效能	检验效能标准误	FDR	假发现率标准误	准确率	准确率标准误	LL
套索回归	无 FDR 截断值	77.3%	0.6%	38.8%	0.7%	91.4874%	0.3097%	-778.1
镜像	MDS（数据驱动）	51.0%	0.6%	6.9%	0.2%	88.3108%	0.3244%	-905.3
镜像	DS（数据驱动）	20.3%	0.6%	5.3%	0.3%	86.4228%	0.1798%	-1086.4
Knockoff	数据驱动	10.5%	1.0%	0.01%	0.01%	77.5961%	0.5015%	-1231.2
BH	PRDS	54.2%	0.7%	3.8%	0.3%	88.5001%	0.3119%	-898.3
SHRED	PRDS	56.8%	0.7%	1.3%	0.3%	89.7150%	0.1991%	-893.4
SHRED	启发式方法	60.2%	0.6%	4.7%	0.3%	90.0983%	0.1892%	-886.5
SHREDDER	PPRDS	58.8%	0.7%	1.7%	0.4%	89.6112%	0.3024%	-891.6
SHRED	任意	51.8%	1.1%	0.1%	0.05%	88.9037%	0.4003%	-903.1
BY	任意	49.3%	0.9%	0.6%	0.4%	87.8916%	0.4101%	-916.5

表 13： 逻辑回归， $n = 5000, p = 200, T = 100$， AR(1)相关性 $\rho = 0.6$

方法	FDR 截止值	检验效能	检验效能标准误	FDR	假发现率标准误	准确率	准确率标准误	LL
套索回归	无 FDR 截断值	91.6%	0.6%	42.6%	0.5%	91.7359%	0.2021%	-641.6
镜像	MDS（数据驱动）	77.1%	0.6%	2.0%	0.5%	91.7118%	0.1420%	-721.7
镜像	DS（数据驱动）	75.8%	0.7%	2.1%	0.5%	91.3000%	0.1358%	-725.8
Knockoff	数据驱动	33.9%	0.4%	1.3%	0.2%	89.1149%	0.1182%	-806.2
BH	PRDS	76.5%	0.6%	2.2%	0.3%	91.5112%	0.0953%	-724.2
SHRED	PRDS	77.0%	0.5%	1.2%	0.2%	91.8896%	0.0893%	-728.8
SHRED	启发式方法	80.4%	0.5%	3.8%	0.4%	92.1404%	0.1355%	-711.8
SHREDDER	PPRDS	79.3%	0.6%	3.7%	0.3%	91.9986%	0.0887%	-717.3
SHRED	任意	73.3%	0.6%	0.2%	0.01%	91.6456%	0.1003%	-757.9
BY	任意	72.5%	0.6%	0.8%	0.3%	91.2325%	0.1134%	-758.8

表 14： 逻辑回归， $n = 5000, p = 200, T = 100$， AR(1)相关性 $\rho = 0.9$

方法	错误发现率 截止值	检验效能	检验效能标准误	FDR	假发现率标准误	准确率	准确率标准误	LL
套索回归	无 FDR 截断值	89.0%	0.7%	38.7%	0.4%	91.3000%	0.2000%	-664.4
镜像	MDS（数据驱动）	53.5%	0.8%	5.5%	0.4%	90.6000%	0.2000%	-822.6
镜像	DS（数据驱动）	49.7%	0.7%	5.0%	0.5%	88.7522%	0.2158%	-834.9
Knockoff	数据驱动	15.7%	0.5%	0.6%	0.1%	85.3180%	0.1589%	-1027.6
BH	PRDS	52.3%	0.7%	3.3%	0.3%	91.1372%	0.1301%	-827.8
SHRED	PRDS	53.9%	0.6%	1.6%	0.2%	91.4110%	0.1152%	-819.6
SHRED	启发式方法	57.7%	0.6%	4.3%	0.4%	91.6380%	0.1355%	-773.4
SHREDDER	PPRDS	55.6%	0.7%	3.7%	0.3%	91.4732%	0.1356%	-784.9
SHRED	任意	45.1%	0.6%	0.3%	0.1%	88.9065%	0.1509%	-948.3
BY	任意	43.8%	0.6%	0.9%	0.2%	88.5072%	0.2137%	-951.3

C.3. 泊松回归

表 15: 泊松回归， $n = 5000, p = 200, T = 100$， 独立预测变量

方法	FDR 截止值	检验效能	检验效能标准误	FDR	假发现率标准误	MSE	均方误差标准误	LL
套索回归	无 FDR 截断值	89.9%	1.2%	43.6%	0.7%	1.5200	0.1521	-1023.4
镜像	MDS（数据驱动）	52.8%	1.1%	4.5%	0.6%	1.6845	0.1410	-1280.6
镜像	DS（数据驱动）	48.7%	1.0%	3.3%	0.6%	1.8189	0.1528	-1312.3
BH	PRDS	58.7%	0.4%	1.6%	0.2%	1.5690	0.1211	-1222.8
SHRED	PRDS	58.6%	0.6%	1.8%	0.3%	1.5603	0.1136	-1221.9
SHRED	启发式方法	60.1%	0.5%	3.1%	0.2%	1.5249	0.1108	-1220.4
SHREDDER	PPRDS	58.8%	0.5%	2.1%	0.3%	1.5595	0.1009	-1221.7
SHRED	任意	51.3%	0.4%	0.6%	0.2%	1.6808	0.1189	-1253.4
BY	任意	51.2%	0.3%	0.5%	0.2%	1.6872	0.1289	-1254.0

表 16： 泊松回归，$n = 5000, p = 200, T = 100$, 单一公共成分相关性 $\rho = 0.5$

方法	错误发现率 截止值	检验效能	检验效能标准误	FDR	假发现率标准误	MSE	均方误差标准误	LL
套索回归	无 FDR 截断值	88.3%	1.4%	40.2%	0.9%	1.5331	0.1580	-1064.3
镜像	MDS（数据驱动）	40.1%	1.6%	2.1%	0.4%	1.6742	0.1287	-1519.1
镜像	DS（数据驱动）	18.4%	1.6%	0.8%	0.4%	1.8560	0.1519	-1763.9
Knockoff	数据驱动	0%	NA	0%	NA	NA	NA	NA
BH	PRDS	55.0%	0.4%	2.7%	0.9%	1.6544	0.1645	-934.9
SHRED	PRDS	56.5%	0.7%	1.8%	0.4%	1.5224	0.1509	-913.7
SHRED	启发式方法	60.8%	0.4%	3.0%	0.6%	1.3308	0.0895	-879.9
SHREDDER	PPRDS	57.7%	0.4%	2.3%	0.5%	1.4196	0.0649	-912.6
SHRED	任意	50.2%	0.7%	0.1%	0.01%	1.6635	0.1267	-944.9
BY	任意	49.3%	0.8%	1.1%	0.9%	1.6821	0.1424	-1173.6

表 17： 泊松回归，$n = 5000, p = 200, T = 100$,单一公共成分相关 $\rho = 0.9$

方法	FDR 截止值	检验效能	检验效能标准误	FDR	假发现率标准误	MSE	均方误差标准误	LL
套索回归	无 FDR 截断值	69.3%	1.7%	37.9%	1.1%	1.5119	0.1215	-1348.7
镜像	MDS（数据驱动）	21.0%	0.6%	5.8%	0.8%	1.7272	0.1109	-1531.1
镜像	DS（数据驱动）	14.0%	0.8%	4.0%	0.5%	1.8631	0.1213	-1839.9
Knockoff	数据驱动	0%	NA	0%	NA	NA	NA	NA
BH	PRDS	38.8%	0.7%	3.3%	1.0%	1.5769	0.0919	-1532.6
SHRED	PRDS	41.8%	0.8%	1.4%	0.1%	1.5046	0.0819	-1250.6
SHRED	启发式方法	45.4%	0.7%	3.2%	0.1%	1.4161	0.0800	-1250.4
SHREDDER	PPRDS	42.8%	0.3%	1.8%	0.8%	1.4446	0.0799	-1257.5
SHRED	任意	34.9%	0.9%	0.2%	0.02%	1.6211	0.0831	-1535.8
BY	任意	32.0%	0.7%	1.3%	0.7%	1.6659	0.1459	-1552.7

表 18： 泊松回归，$n = 5000, p = 200, T = 100$，聚类相关 $\rho = 0.3$ 至$\rho = 0.6$

方法	FDR 截止值	检验效能	检验效能标准误	FDR	假发现率标准误	MSE	均方误差标准误	LL
套索回归	无 FDR 截断值	76.2%	2.4%	37.9%	1.4%	1.6099	0.1814	-1305.8
镜像	MDS（数据驱动）	30.4%	1.2%	7.5%	0.3%	2.6833	0.1795	-1989.3
镜像	DS（数据驱动）	8.9%	1.7%	6.0%	0.6%	4.9602	0.3427	-3816.7
Knockoff	数据驱动	0%	NA	0%	NA	NA	NA	NA
BH	PRDS	35.3%	1.4%	2.5%	0.5%	2.4304	0.2109	-1798.4
SHRED	PRDS	36.4%	0.6%	1.6%	0.5%	2.3435	0.1499	-1696.6
SHRED	启发式方法	39.3%	0.9%	1.7%	0.4%	2.2038	0.1916	-1658.8
SHREDDER	PPRDS	37.0%	0.8%	1.2%	0.2%	2.2548	0.1908	-1687.5
SHRED	任意	34.1%	0.9%	0.2%	0.03%	2.5448	0.1876	-1801.8
BY	任意	33.9%	1.1%	0.3%	0.01%	2.5932	0.2005	-1801.9

表 19： 泊松回归，$n = 5000$、$p = 200$、$T = 100$，聚类相关性 $\rho = 0.6$ 至$\rho = 0.9$

方法	FDR 截止值	检验效能	检验效能标准误	FDR	假发现率标准误	MSE	均方误差标准误	LL
套索回归	无 FDR 截断值	74.8%	1.6%	39.9%	1.0%	2.1270	0.2229	-2075.5
镜像	MDS（数据驱动）	28.4%	1.7%	9.5%	0.4%	3.4712	0.2340	-2362.3
镜像	DS（数据驱动）	8.1%	2.1%	6.0%	0.8%	5.3871	0.3202	-4353.1
Knockoff	数据驱动	0%	NA	0%	NA	NA	NA	NA
BH	PRDS	31.3%	1.5%	1.7%	0.5%	3.2335	0.2483	-2288.8
SHRED	PRDS	33.7%	1.0%	1.5%	0.5%	3.2223	0.2012	-2252.5
SHRED	启发式方法	34.8%	1.4%	2.3%	0.6%	3.2891	0.2163	-2239.6
SHREDDER	PPRDS	33.4%	1.3%	1.5%	0.5%	3.2214	0.2016	-2243.7
SHRED	任意	30.1%	1.4%	0.3%	0.02%	3.4893	0.2315	-2324.4
BY	任意	29.3%	1.4%	0.4%	0.01%	3.5064	0.2328	-2359.7

表20： 泊松回归，$n = 5000, p = 200, T = 100$，一阶自回归相关性 $\rho = 0.6$

方法	FDR 截止值	检验效能	检验效能标准误	FDR	假发现率标准误	MSE	均方误差标准误	LL
套索回归	无 FDR 截断值	98.0%	0.04%	40.9%	0.8%	1.5649	0.1180	-1012.6
镜像	MDS（数据驱动）	73.1%	1.3%	3.0%	0.2%	1.8737	0.1699	-1284.5
镜像	DS（数据驱动）	66.6%	1.6%	1.3%	0.6%	1.9498	0.1907	-1440.3
Knockoff	数据驱动	0%	NA	0%	NA	NA	NA	NA
BH	PRDS	83.5%	0.2%	2.9%	0.6%	1.6109	0.0835	-1092.4
SHRED	PRDS	83.8%	0.6%	1.5%	0.4%	1.5097	0.0949	-1073.9
SHRED	启发式方法	86.6%	0.5%	3.6%	0.6%	1.5126	0.0849	-1061.3
SHREDDER	PPRDS	85.3%	0.6%	2.5%	0.6%	1.5246	0.1218	-1072.4
SHRED	任意	82.5%	0.8%	0.2%	0.02%	1.6357	0.0811	-1096.2
BY	任意	81.1%	0.6%	1.4%	0.6%	1.6886	0.0635	-1094.7

表21： 泊松回归，$n = 5000$、$p = 200$、$T = 100$，一阶自回归相关性 $\rho = 0.9$

方法	FDR 截止值	检验效能	检验效能标准误	FDR	假发现率标准误	MSE	均方误差标准误	LL
套索回归	无 FDR 截断值	87.5%	1.8%	33.4%	2.0%	2.0505	0.2099	-1556.2
镜像	MDS（数据驱动）	41.1%	1.5%	1.9%	1.0%	2.5636	0.1897	-2219.9
镜像	DS（数据驱动）	20.3%	2.0%	0.0%	0.0%	2.6331	0.2193	-3262.8
Knockoff	数据驱动	0%	NA	0%	NA	NA	NA	NA
BH	PRDS	75.9%	0.3%	1.8%	0.6%	1.9089	0.0808	-1596.7
SHRED	PRDS	78.7%	0.8%	1.3%	0.6%	1.8848	0.0811	-1563.9
SHRED	启发式方法	80.8%	1.1%	3.5%	0.7%	1.8792	0.1017	-1550.8
SHREDDER	PPRDS	79.2%	0.9%	2.1%	0.5%	1.9101	0.1190	-1583.5
SHRED	任意	73.8%	0.7%	0.2%	0.01%	1.9934	0.0889	-1599.7
BY	任意	72.4%	0.9%	0.6%	0.07%	2.0531	0.1295	-1613.7

C.4. 比较 SHRED 与 BH方法的模型性能

表22： H_A 的单样本t检验 p值：$\mu_{\text{MSE}_{\text{SHRED}}} < \mu_{\text{MSE}_{\text{BH}}}$ 和 H_A ： $\mu_{\text{Accuracy}_{\text{SHRED}}} > \mu_{\text{Accuracy}_{\text{BH}}}$，或用于其他比较的类似假设。

模拟	SHRED PRDS 与 Benjamini–Hochberg 方法	SHREDDER 与 Benjamini–Hochberg 方法	SHRED 与 由
高斯，独立	0.98	0.98	0.98
高斯，单一公共成分 $\rho = 0.5$	0.51	0.19	0.16
高斯，单一公共成分 $\rho = 0.9$	0.15	0.14	< 0.01
高斯 ，低聚类	< 0.01	< 0.01	< 0.01
高斯，高聚类	0.21	0.12	0.31
高斯，一阶自回归 $\rho = 0.6$	0.35	0.13	0.25
高斯，一阶自回归, $\rho = 0.9$	0.044	0.023	< 0.01
逻辑斯蒂，独立	0.25	0.37	0.35
逻辑斯蒂，单一公共成分 $\rho = 0.5$	0.038	0.022	0.24
逻辑斯蒂，单一公共成分 $\rho = 0.9$	0.41	0.37	0.47
逻辑斯蒂，低聚类	0.18	0.08	< 0.01
逻辑斯蒂，高聚类	0.031	0.08	0.039
逻辑斯蒂，一阶自回归 $\rho = 0.6$	< 0.01	< 0.01	< 0.01
逻辑斯蒂，一阶自回归， $\rho = 0.9$	< 0.01	< 0.01	< 0.01
泊松，独立	0.48	0.47	0.48
泊松，单一公共成分 $\rho = 0.5$	0.28	0.043	0.58
泊松，单一公共成分 $\rho = 0.9$	0.38	0.14	0.38
泊松 ，低聚类	0.39	0.26	0.35
泊松，高聚类	0.59	0.59	0.43
泊松，一阶自回归 $\rho = 0.6$	0.30	0.27	0.16
泊松，一阶自回归， $\rho = 0.9$	0.32	0.57	0.46

参考文献

Benjamini, Y. 和 Yekutieli, D.（2001年）。依赖关系下多重检验中错误发现率的控制。《统计年鉴》，29(4)。