arXiv:2304.05261v1 [math.ST] 11 Apr 2023

关于相关正态分布均值的双侧备择假设多重检验中错误发现率的控制

萨纳特·K. 萨卡尔*

2023年4月12日

摘要

本文重新审视了关于在双侧备择假设下同时检验多元正态均值的以下未决问题：Benjamini和Hochberg 的方法（BH，1995）能否在不对相关系数施加任何依赖结构的情况下控制错误发现率（错误发现率）？人们普遍认为这个问题的答案是肯定的；自从有关该问题的数值研究结果见于大量论文并强有力地支持这一结论以来，文献中一直对此作此猜想。据我们所知，迄今为止文献中尚未提出这一答案的理论依据。本文对这一猜想给出了部分证明。更具体地说，我们考虑以下两种情形——（i）协方差矩阵已知；（ii）协方差矩阵是某个已知矩阵的未知标量倍数——并证明在这两种情形下，基于原始 z-或 t-检验统计量的某些加权版本的 BH 型逐步递增法能够控制错误发现率。

1 引言

Benjamini 与 Hochberg（1995）提出的错误发现率（FDR）是多重检验中衡量总体第一类错误的一种有力概念。随着多重检验成为现代科学研究中统计调查所广泛采用的推断工具，Benjamini 与 Hochberg（1995）设计的用于控制这一错误率的方法，即广为人知的 BH 方法，如今已成为最常用的多重检验方法之一。尽管 BH 方法广受欢迎，但在许多实际应用中，将其作为有效的错误发现率控制方法使用却常常受到质疑，因为确保其有效性的条件在这些应用所遇到的多重检验情境中往往并不满足。其中一种情境涉及对具有正定相关矩阵的相关正态随机变量的均值进行多重检验。当备择假设均为单侧时，只要相关系数非负，BH 方法在这一情境下即可被证明能够控制错误发现率 [例如参见 Benjamini 与 Yekutieli（2001）、Blanchard 与 Roquain（2008）、Finner 等（2007）、Sarkar（2002）]。然而，没有这样的结果

*美国天普大学统计学、运筹学与数据科学系。电子邮箱：sanat@temple.edu。本研究由美国国家科学基金会（NSF）资助，项目编号 DMS 2210687

关于在双侧备择假设下对均值进行多重检验时其错误发现率控制的文献研究尚不充分。针对双侧备择假设进行检验，在科学意义上通常比针对单侧备择假设进行检验更为重要。因此，在许多现代统计应用中，如果不从理论上验证 BH方法在上述双侧检验情形下确实是一种有效的错误发现率控制方法，其适用性将受到限制。这也是促使近期研究热潮兴起的动机之一：这些研究绕开 BH方法的使用，并开发替代方法[Barber & Cand`es (2015)、Fithin& Lei(2022)、Sarkar&Tang (2022)]，从而能够在相同的多重检验情形下从理论上控制错误发现率。

当然，多重检验研究者普遍坚信，而且常常有人猜想：无论相关矩阵为何，BH 方法都确实能够控制针对双侧备择假设的相关正态均值多重检验中的错误发现率。正如 Benjamini (2010) 所言：

> 通常无需针对一般依赖关系进行修正：令人信服的模拟和理论证据表明，对于具有任意相关结构的双侧 z-检验，情况也相同[Reiner-Benaim (2007)]，，但这一理论尚待完整证明。

许多研究 BH 方法在此类检验问题中的错误发现率控制的论文所开展的数值研究结果，也为这一信念提供了有力支持。因此，鉴于上述替代方法近期取得的进展，证明这一猜想，从而使 BH 方法成为一种理论上有效的错误发现率控制方法，并能与这些其他方法相竞争，看来是一项紧迫而重要的工作。

本文通过 BH 方法的一个加权版本，对已知但任意的相关矩阵情形下的这一猜想给出部分证明。具体而言，我们考虑一种 BH 型逐步递增法，其中，Benjamini (2010) 上述评论中提到的双侧 z-检验的 z^2 值 []，根据每个 z 与其他项的相关程度进行加权，从而证明这一猜想。随后，我们拓展证明中所用的论证，证明了涉及双侧 t-检验中采用类似加权的 t^2 值的 BH 型逐步递增法。这些证明作为特殊情形，确立了涉及上述加权 z-和t-检验统计量的 Simes 全局检验的有效性。关于涉及未加权 z-和 t-检验统计量的 Simes 全局检验猜想，参见 Finner等（2017）。

本文组织如下。第 2 节通过公式介绍逐步升序检验的错误发现率以及错误发现率可以得到控制的条件。第3节给出我们提出的方法及其相关的主要结果。第 4 节在等相关多元正态分布和线性回归模型下的变量选择背景中阐述这些方法，并提出一种用于变量选择的新型BH 型错误发现率控制程序。第 5 节以一些补充说明作为全文的结尾。

2 预备知识

本节先介绍与 BH 方法或一种密切相关的逐步递增方法有关的一些基本符号、公式和假设，然后进一步讨论，为下一节的主要

结果做好铺垫。

给定一组 d 原假设 $H_i,\ i = 1,\ldots,d$，需要同时检验，并使用各自的 p 值或其某些递增函数 $P_i,\ i = 1,\ldots,d$，BH 方法是一种应用于 P_i 的逐步升序检验，其临界常数为 $\alpha_i = i\alpha_1,\ i = 1,\ldots,d$；也就是说，它求出$R = \max_{1\le i\le d}\{i:\ P_{(i)} \le \alpha_i\}$，并在最大值存在时，拒绝所有满足 $P_i \le P_{(R)}$ 的 i 对应的 H_i；否则，不拒绝任何假设。该方法旨在控制

$$\mathrm{FDR} = \mathbb{E}(\mathrm{FDP}),\ \text{where FDP (False Discovery Proportion)} = \frac{V}{R \vee 1}, \tag{2.1}$$

$(a \vee b = \max(a,b))$，其中 V 和 R 分别是在显著性水平 $d\alpha_1$ 下被错误拒绝和被拒绝的原假设数量。

以下引理给出了具有任意一组临界常数的逐步升序检验的错误发现率的显式表达式：

引理 1 设 $I_0 = \{j:\ H_j \text{ 为真}\}$。则将临界常数 $0 < \alpha_1 \le \cdots \le \alpha_d < 1$ 应用于 P_i 时，逐步升序检验的错误发现率为

$$\begin{aligned} FDR &= \sum_{i\in I_0} \mathbb{E}\left\{ \frac{\mathbb{I}(P_i \le \alpha_{R_{-i}+1})}{R_{-i}+1} \right\} \\ &= \sum_{i\in I_0}\sum_{r=0}^{d-1} \mathbb{E}\left\{ \frac{\mathbb{b}^{\mathbb{I}}(P_i \le \alpha_{r+1} \mid \boldsymbol{P}_{-i})\, \mathbb{I}(R(\boldsymbol{P}_{-i}) = r)}{r+1} \right\}, \end{aligned} \tag{2.2}$$

其中 $\boldsymbol{P}_{-i} = (P_1,\ldots,P_d) \setminus \{P_i\}$ 且 $R_{-i} \equiv R(\boldsymbol{P}_{-i})$ = 最大值$_{1\le j\le d-1}\{j:\ P_{(j)\setminus\{i\}} \le \alpha_{j+1}\}$，其中$P_{(1)\setminus\{i\}} \le \cdots \le P_{(d-1)\setminus\{i\}}$是 $\boldsymbol{P}_{-i}$的有序分量。

关于这一公式，参见 Sarkar（2002，2008）；不过，其他论文中也出现了类似公式 [，例如 Benjamini 与 Yekutieli（2001）、Blanchard 与 Roquain（2008）、Finner 等（2007）]。

当 P_i彼此独立时，由该公式可立即得出，基于这些 P_i的 BH 方法的错误发现率等于 $\sum_{i\in I_0} \mathbb{b}^{\mathbb{I}}(P_i \le \alpha_1)$，因此在以下假设下，该方法的错误发现率受控于 $|I_0|\alpha_1 \le d\alpha_1$，其中 $|I_0|$ 是 I_0的基数：

假设1对于每个 $i \in I_0$，P_i均随机大于服从 $U(0,1)$ 分布的随机变量。

当 P_i 彼此不独立时，要找出使 BH 方法或一种密切相关的逐步上升法仍能控制错误发现率的条件，一种方法是将上述公式改写为另一种形式，从而揭示为证明错误发现率控制而需要对 P_i 具备的依赖类型。其中一个这样的公式是

$$\begin{aligned} \mathrm{FDR} &= \sum_{i\in I_0} \mathbb{b}^{\mathbb{I}}(P_i \le \alpha_1) + \\ &\quad \sum_{i\in I_0}\sum_{r=1}^{d-1} \mathbb{E}\left\{ \mathbb{b}^{\mathbb{I}}(R(\boldsymbol{P}_{-i}) \ge r \mid P_i) \left[\frac{\mathbb{I}(P_i \le \alpha_{r+1})}{r+1} - \frac{\mathbb{I}(P_i \le \alpha_r)}{r} \right] \right\}, \end{aligned} \tag{2.3}$$

（Sarkar（2002））。集合 $\{\boldsymbol{P}_{-i}:\ R(\boldsymbol{P}_{-i}) \geq r\}$ 关于 $\boldsymbol{P}_{-i}$ 是递减的，其中 $r = 1$ 固定，....., $d-1$。这是在假设 1 下证明错误发现率控制的关键，并促使 Benjamini 和 Yekutieli（2001）等人考虑对 P_i 提出以下依赖结构假设，以确保上述递减性质：

假设2 P_i 关于对应于原假设的 P_i 的子集具有正回归依赖性（PRDS），即，

$$\mathbb{E}\{\phi(\boldsymbol{P}_{-i}) \mid P_i\} \uparrow (or \downarrow),\ P_i\ for\ each\ i \in I_0, \tag{2.4}$$

对于 $\boldsymbol{P}_{-i}$ 的任意逐坐标递增（或递减）函数。

PRDS 是一种正依赖条件，由于在许多实际场景中都能满足，包括 p 值由具有非负相关系数的正态检验统计量生成的场景，因此如今已被接受为唯一一种能够使 BH 方法在非渐近设定下控制错误发现率的正依赖条件。然而遗憾的是，它无法捕捉检验统计量所表现出的正依赖结构，因而也无法捕捉在许多其他且相对更重要的场景中由此产生的相应 p 值。其中包括检验统计量服从折叠多元正态分布或折叠多元 t 分布的情形。这些统计量出现在已知或未知方差的相关正态分布的均值针对双侧备择假设进行多重检验的背景下。在单侧检验问题中用于检验PRDS条件的多元正态分布工具或分布性质，对于折叠多元正态分布或用于双侧检验问题的 t 检验统计量已不再适用。例如，当检验统计量服从多元正态分布时，在相关系数为非负的情况下，给定 P_i 时 $\boldsymbol{P}_{-i}$的条件分布会随 P_i 的增加而随机增大，由此可以针对单侧检验问题验证PRDS条件。类似地，当偏相关系数全部为正时，已知多元正态分布是二阶全正的（MTP_2），这意味着 PRDS 条件成立；参见 Karlin& Rinott（1980）和Sarkar（2002）。对于基于折叠多元正态分布或折叠 t 检验统计量的 BH 方法，这些工具和分布性质并不适用。

因此，在非渐近情形下，对于具有任意但已知的正定相关矩阵的多元正态分布的均值进行多重检验时，无论涉及 z^2-值或 t^2-值的 BH 方法，还是涉及它们的某些递增函数的 BH 方法，能否被证明控制错误发现率，始终是多重检验中的重要开放问题之一。下一节将通过我们所称的加权 BH 方法来解决这一问题。

注 1 这里有必要明确说明，当给定权重 $w_i > 0, i = 1,..., d$ 的加权方案涉及底层检验统计量，而不是文献中通常假定的 p-值时，我们所说的加权 BH方法究竟是什么意思。设 T_i 是生成 p-值 $P_i = \bar{F}_0(T_i)$的检验统计量，其中 $i = 1,..., d$ 使用其原假设生存函数 $\bar{F}_0$；并设 $w_i^{-1}T_i$（其中 $i = 1, ..., d$）是 T_i's 的加权版本。那么，对于 $\widetilde{P}_i = \bar{F}_0(w_i^{-1}\bar{F}_0^{-1}(P_i)), i = 1,..., d$ 应用 BH 方法，并使用临界常数 $i\alpha_1$（其中 $i = 1, \ldots d$），我们称之为基于 P_i's的加权 BH方法。或者，如果我们关注的是用检验统计量而非 p-值来表述的 BH 方法；

也就是说，在将 T_i 按 $T_{(1)} \leq \cdots \leq T_{(d)}$排序后，找出 $R = \min\{i\colon\ T_{(i)} \geq \bar{F}_0^{-1}((d-i+1)\alpha_1)\}$，并对满足 $T_i \geq T_{(R)}$ 的 H_i for i 作出拒绝，前提是该最小值存在；否则不拒绝任何对象。随后，我们将 T_i 替换为其加权版本（使用分配给它们的一些权重定义）的方法称为加权 BH 方法。

3 主要结果

本节介绍本文的主要结果，即定理 1 和定理 2，它们与上述针对多元正态分布均值进行多重检验的双侧检验有关。在这两个定理中，假设相关矩阵已知，但不呈现任何特定的依赖结构。定理 1 提出了我们基于加权 z^2 值的错误发现率控制 BH 型逐步递增法；定理 2 则提出了基于加权 t^2 值的同一方法。

3.1 加权 z^2 值

设有一个 d 维随机向量 $\boldsymbol{X} = (X_1, \ldots, X_d)' \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，其均值向量 $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)'$ 未知，且协方差矩阵 $\boldsymbol{\Sigma} = ((\sigma_{ij}))$ 已知并为正定。我们的检验问题是同时对 $i = 1, \ldots, d$ 检验 $H_i\colon\ \mu_i = 0$ 对 $\mu_i \neq 0$，并将错误发现率控制在 α。

设 $\bar{\Psi}_n = 1 - \Psi_n$，其中 Ψ_n 表示 χ_n^2 的累积分布函数；χ_n^2 是具有 n 个自由度的中心卡方随机变量。则 $P_i = \bar{\Psi}_1(Z_i^2)$、$i = 1$、.....、$d$，其中 $Z_i = X_i/\sqrt{\sigma_{ii}}$，是双侧 z 检验中的 p 值。我们不将 BH 方法应用于这些 p 值，而是考虑使用临界常数 $i\alpha_1$、$i = 1$、.....、d 将其变换为以下递增函数，并将 BH 方法应用于这些函数；其中 α_1 的选取需适当，并取决于要控制错误发现率的水平 α：

$$\widetilde{P}_i = \bar{\Psi}_1\left(w_i^{-1}\bar{\Psi}_1^{-1}(P_i)\right), \text{ where } w_i = 1 - R_i^2,\ i = 1, \ldots, d, \tag{3.1}$$

且 R_i^2 是 Z_i 与 $(Z_1$，.....，$Z_d) \setminus \{Z_i\}$ 之间的复相关系数平方。换言之，我们考虑为每个 Z_i^2 指定权重 w_i，以反映 Z_i 与其他变量之间的相关程度[1]，并定义 $Y_i = w_i^{-2} Z_i$ 后，对于 $i = 1$、.....、d，我们对 Y_i^2 运行 BH 方法，即对加权 z^2 值使用临界常数 $\bar{\Psi}_1^{-1}((d-i+1)\alpha_1)$、$i = 1$、.....、$d$（如注 1 所述）。这就是我们提出的涉及双侧 z 检验的加权 BH 方法。以下定理说明该方法能够控制错误发现率。

~定理1将BH 方法应用于 (3.1) 中的 P_i 时，使用临界常数$i\alpha_1$, $i = 1, \ldots, d$，其中 α_1 满足

$$\sum_{i=1}^{d} \bar{\Psi}_1\left(w_i\bar{\Psi}_1^{-1}(\alpha_1)\right) = \alpha, \tag{3.2}$$

控制在 α。

在证明该定理之前，我们先陈述两个引理，它们将在我们证明该定理的过程中发挥关键作用。

第一个引理（引理 2）给出了一个与 $\boldsymbol{Y}=(Y_1,\ldots,Y_d)'$ 的分布有关的有用结果。在推导该分布时，我们首先注意到 $\boldsymbol{Z}=(Z_1,\ldots,Z_d)'\sim N_d(\boldsymbol{\nu},\boldsymbol{\Sigma})$，其中 $\boldsymbol{\nu}=(\nu_1,\ldots,\nu_d)$，且对于 $i=1,\ldots,d$，有 $\nu_i=\mu_i/\sqrt{\sigma_{ii}}$，并且可以假设 $\boldsymbol{\Sigma}$ 是相关矩阵。此外，$\boldsymbol{\Sigma}^{-1}$ 的第 i 个对角元素等于$(1-R_i^2)^{-1}=w_i^{-1}$。因此，$\boldsymbol{Y}=\mathrm{diag}\{w_i^{-\frac{1}{2}}\}\boldsymbol{Z}\sim N_d(\boldsymbol{\delta},\boldsymbol{\Gamma}^{-1})$，其中 $\boldsymbol{\delta}=(\delta_1,\ldots,\delta_d)'$，且对于 $i=1,\ldots,d$，有 $\delta_i=\nu_i/\sqrt{w_i}=\mu_i/\sqrt{\sigma_{ii}(1-R_i^2)}$，并且 $\boldsymbol{\Gamma}=\mathrm{diag}\{w_i^{\frac{1}{2}}\}\boldsymbol{\Sigma}^{-1}\mathrm{diag}\{w_i^{\frac{1}{2}}\}$ 的第 i 个对角元素为 1。

采用以下记号：

- $\boldsymbol{Y}_{-i}$：$\boldsymbol{Y}$去掉第 i个元素 Y_i后得到的（$d-1$）维子向量，
- $\boldsymbol{\delta}_{-i}$：$\boldsymbol{\delta}$ 去掉第 i个元素 δ_i后得到的（$d-1$）维子向量，
- $\boldsymbol{\gamma}_{-i,i}$：$\boldsymbol{\Gamma}$ 的第 i列去掉第 i个元素1后得到的（$d-1$）维子向量，以及
- $\boldsymbol{\Gamma}_{-i,-i}$：$\boldsymbol{\Gamma}$ 去掉第 i行和第 i列后得到的（$d-1$）$\times$（$d-1$）主子矩阵，

由此可见

$$\begin{aligned} Y_i-\delta_i \mid \boldsymbol{Y}_{-i} &\sim N(-\boldsymbol{\gamma}'_{-i,i}(\boldsymbol{Y}_{-i}-\boldsymbol{\delta}_{-i}),1)\\ \boldsymbol{Y}_{-i} &\sim N_{d-1}\left(\boldsymbol{\delta}_{-i},[\boldsymbol{\Gamma}_{-i,-i}-\boldsymbol{\gamma}_{-i,i}\boldsymbol{\gamma}'_{-i,i}]^{-1}\right), \end{aligned}$$

从而得到以下引理。

引理 2 当 $\mu_i=0$，

$$Y_i^2 \mid \boldsymbol{Y}_{-i}\sim {\chi'_1}^2(\lambda_i(\boldsymbol{Y}_{-i})), \tag{3.3}$$

自由度为 1、非中心参数为 $\lambda_i(\boldsymbol{Y}_{-i})=(\boldsymbol{\gamma}'_{-i,i}(\boldsymbol{Y}_{-i}-\boldsymbol{\delta}_{-i}))^2$ 的非中心卡方分布。

下一个引理给出了一个与非中心卡方分布相关的有用结果。

引理 3

$$\frac{\mathbb{P}[{\chi'_n}^2(\lambda)\geq\bar{\Psi}_n^{-1}(u)]}{u}\downarrow u\in(0,1),$$

对任意固定的 n、$\lambda>0$。

证明。利用 ${\chi'_n}^2(\lambda)\overset{d}{=}\mathbb{E}_J(\chi^2_{n+2J})$ 这一事实，其中期望是相对于$J\sim$泊松分布 $\left(\frac{\lambda}{2}\right)$取的，我们首先注意到

$$\mathbb{P}[{\chi'_n}^2(\lambda)\geq\bar{\Psi}_n^{-1}(u)]=\mathbb{E}_J\left[\bar{\Psi}_{n+2J}(\bar{\Psi}_n^{-1}(u)\right]. \tag{3.4}$$

考虑函数 $g(u)=\bar{\Psi}_{n+h}(\bar{\Psi}_n^{-1}(u))$ $u\in(0,1)$，对于任意固定的 $n,h>0$，并注意到 $\Psi_n(y)$ 的密度 $\psi_n(y)$ 为 $\propto e^{-\frac{1}{2}y}y^{\frac{n}{2}-1}\mathbb{I}(y>0)$；$n>0$，于是我们得到

$$\frac{dg(u)}{du}=\frac{\psi_{n+h}(\bar{\Psi}_n^{-1}(u))}{\psi_n(\bar{\Psi}_n^{-1}(u))}\propto\left(\bar{\Psi}_n^{-1}(u)\right)^{\frac{h}{2}}\downarrow u\in(0,1);$$

也就是说， $g(u)$ 在 $u \in (0,1)$ 上是凹函数。由于 $g(0) = 0$， g 的凹性意味着 $\frac{g(u)}{u} \downarrow u \in (0,1)$。将此结果代入 (3.4)，即可证明引理 3。

定理 1 的证明。我们使用 (2.2) 中的公式，其中 P_i 和 $\boldsymbol{P}_{-i}$ 分别用 Y_i 和 $\boldsymbol{Y}_{-i}$ 表示。由引理 2 可知，对 (3.1) 中的 P_i 应用逐步升序检验，若采用任意一组临界常数 $\alpha_i, i = 1, \ldots, d$，则其错误发现率为

$$\begin{aligned}
\text{FDR} &= \sum_{i \in I_0} \sum_{r=0}^{d-1} \mathbb{E}\left\{ \frac{\mathbb{P}[Y_i^2 \geq \bar{\Psi}_1^{-1}(\alpha_{r+1}) \mid \boldsymbol{Y}_{-i}]\, \mathbb{I}(R(\boldsymbol{Y}_{-i}) = r)}{r+1} \right\} \\
&= \sum_{i \in I_0} \sum_{r=0}^{d-1} \mathbb{E}\left\{ \frac{\mathbb{P}[\chi_1'^2(\lambda_i(\boldsymbol{Y}_{-i})) \geq \bar{\Psi}_1^{-1}(\alpha_{r+1})]}{r+1} \mathbb{I}(R(\boldsymbol{Y}_{-i}) = r) \right\} \\
&\leq \sum_{i \in I_0} \sum_{r=0}^{d-1} \mathbb{E}\left\{ \frac{\alpha_{r+1}}{r+1} \frac{\mathbb{P}[\chi_1'^2(\lambda_i(\boldsymbol{Y}_{-i})) \geq \bar{\Psi}_1^{-1}(\alpha_1)]}{\alpha_1} \mathbb{I}(R(\boldsymbol{Y}_{-i}) = r) \right\},
\end{aligned}$$

其中不等式由引理 3 得出。

因此，对于满足 $\alpha_r = r\alpha_1$ 的 BH 方法，我们最终得到

$$\begin{aligned}
\text{FDR} &\leq \sum_{i \in I_0} \mathbb{E}\left\{ \mathbb{P}[\chi_1'^2(\lambda_i(\boldsymbol{Y}_{-i})) \geq \bar{\Psi}_1^{-1}(\alpha_1)] \sum_{r=0}^{d-1} \mathbb{I}(R(\boldsymbol{Y}_{-i}) = r) \right\} \\
&= \sum_{i \in I_0} \mathbb{E}\left\{ \mathbb{P}[\chi_1'^2(\lambda_i(\boldsymbol{Y}_{-i})) \geq \bar{\Psi}_1^{-1}(\alpha_1)] \right\} = \sum_{i \in I_0} \mathbb{P}[Y_i^2 \geq \bar{\Psi}_1^{-1}(\alpha_1)] \\
&= \sum_{i \in I_0} \mathbb{P}[Z_i^2 \geq (1 - R_i^2)\bar{\Psi}_1^{-1}(\alpha_1)] \leq \sum_{i=1}^{d} \bar{\Psi}_1\left(w_i \bar{\Psi}_1^{-1}(\alpha_1)\right) \\
&= \alpha. \qquad (3.5)
\end{aligned}$$

定理得证。

3.2 加权 t^2 值

涉及双侧 t 检验的多重检验问题设置如下：给定 $\boldsymbol{X} = (X_1, \ldots, X_d)' \sim N_d(\boldsymbol{\mu}, \tau^2\boldsymbol{\Sigma})$，其中 $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)'$ 是未知向量， $\boldsymbol{\Sigma} = ((\sigma_{ij}))$ 是已知的正定矩阵；以及 $V \sim \tau^2\chi_m^2$，其与 $\boldsymbol{X}$ 相互独立，其中 $\tau^2 > 0$ 是未知标量， $m > 0$ 是已知的某个自由度。我们要同时检验 H_i： $\mu_i = 0$ 对 $\mu_i \neq 0$，其中 $i = 1, \ldots, d$，并将错误发现率控制在 α。

令 $\Psi_{m,\,n}$ 表示 $F_{m,\,n}$ 的累积分布函数，其中该随机变量服从具有 m 和 n 个自由度的中心 F 分布，而 $\bar{\Psi}_{m,\,n} = 1 - \Psi_{m,\,n}$。那么 $P_i = \bar{\Psi}_{1,\,m}^{-1}(T_i^2)$，其中 $T_i = \frac{\sqrt{m}Z_i}{\sqrt{V}}$（$i = 1, \ldots, d$）是双侧 t 检验中的 p 值。与双侧 z 检验的情形一样，我们考虑使用这些 p 值的以下递增函数：

$$\widetilde{P}_i = \bar{\Psi}_{1,m}\left(w_i^{-1}\bar{\Psi}_{1,m}^{-1}(P_i)\right), \text{ where } w_i = 1 - R_i^2, i = 1, \ldots, d, \qquad (3.6)$$

根据临界常数 $i\alpha_1$, $i = 1, \ldots, d$ 运行 BH 方法，其中 α_1 根据错误发现率控制水平进行适当选择。换言之，我们考虑加权 t^2

值 $w_i^{-1}T_i^2 = \frac{mY_i^2}{V}$、 $i = 1$、…、 d，而不是 t^2 值 T_i^2's，在 BH 方法中使用临界常数 $\bar{\Psi}_1^{-1}$, m(($d-i+1)\alpha_1$)、 $i = 1$、…、 d（如注 1 所述）。这是我们针对双侧 t-检验提出的加权 BH 方法。以下定理阐述了它能够控制错误发现率这一事实。

~定理 2：将 BH 方法应用于 (3.6) 中的 P_i，并使用临界常数$i\alpha_1, i = 1, \ldots, d$，其中 α_1 满足

$$\sum_{i=1}^{d} \bar{\Psi}_{1,m}\left(w_i \bar{\Psi}_{1,m}^{-1}(\alpha_1)\right) = \alpha, \tag{3.7}$$

控制错误发现率不超过 α。

以下两个引理将在我们对该定理的证明中发挥关键作用。

引理 4

$$\frac{\bar{\Psi}_n(\theta w)}{\bar{\Psi}_n(\theta w')} \uparrow \theta > 0, \ \textit{for any fixed}\, 0 < w < w' < \infty. \tag{3.8}$$

证明。设 $W \sim \frac{1}{\theta}\chi_n^2$。由于 W 在 w 处的概率密度函数，即 $\theta^{\frac{n}{2}} e^{-\frac{1}{2}\theta w} w^{\frac{n}{2}-1}/2^{\frac{n}{2}}\Gamma(\frac{n}{2})$，在 (w, θ^{-1}) 中是TP_2(二阶全正)，因此 W的生存函数， $\bar{\Psi}_n(\theta w) = \mathbb{P}(W \geq w)$，在$(w, \theta^{-1})$ 中也是 TP_2。所以，引理中所要求的单调性结果成立。（关于 TP_2 的相关结果，参见 Karlin (1968)）。

引理 5

$$\frac{\bar{\Psi}_{m+h,n}(\bar{\Psi}_{m,n}^{-1}(u))}{u} \downarrow u \in (0, 1),$$

对于任意固定的 m、 n、 $h > 0$。

证明。该引理可通过与引理 3 相同的方法证明，即证明函数 $g(u) = \bar{\Psi}_{m+h,\ n}(\bar{\Psi}_{m}^{-1},\ _n(u))$ $u \in (0, 1)$ 对任意固定的 m、 n、 $h > 0$ 在 $u \in (0, 1)$ 上是凹函数。

定理 2 的证明。我们可以不失一般性地假设 $\tau^2 = 1$ 且 $\mathbf{\Sigma}$ 是相关矩阵。与定理 1 的证明一样，首先利用 (2.2) 中的公式可知，基于 (3.6) 中的 p 值以及任意一组临界常数 α_i、 $i = 1$、.....、 d 的逐步递增检验，其错误发现率为

$$\begin{aligned}
\mathrm{FDR} &= \sum_{i\in I_0}\sum_{r=0}^{d-1} \mathbb{E}\left\{ \frac{\mathbb{P}[Y_i^2 \geq \frac{V}{m}\bar{\Psi}_{1,m}^{-1}(\alpha_{r+1}) |\ \boldsymbol{Y}_{-i}, V]_{\mathbb{I}}\,(R(\boldsymbol{Y}_{-i}, V) = r)}{r+1} \right\} \\
&= \sum_{i\in I_0}\sum_{r=0}^{d-1} \mathbb{E}\left\{ \frac{\mathbb{P}[{\chi'_1}^2(\lambda_i(\boldsymbol{Y}_{-i})) \geq \frac{V}{m}\bar{\Psi}_{1,m}^{-1}(\alpha_{r+1})]}{r+1} \ \middle|\ V \right)_{\mathbb{I}}(R(\boldsymbol{Y}_{-i}, V) = r) \Bigg\} \\
&= \sum_{i\in I_0}\sum_{r=0}^{d-1} \mathbb{E}\left[\mathbb{E}_{J_i}\left\{ \frac{\bar{\Psi}_{1+2J_i}\left(\frac{V}{m}\bar{\Psi}_{1,m}^{-1}(\alpha_{r+1})\right)}{r+1} \ \middle|\ \boldsymbol{Y}_{-i}, V \right\}_{\mathbb{I}}(R(\boldsymbol{Y}_{-i}, V) = r) \right].
\end{aligned} \tag{3.9}$$

(3.9) 的右侧可以像 (2.3) 中那样重新改写为如下形式：

$$
\begin{aligned}
& \mathrm{FDR} - \sum_{i\in I_0} \bar{\Psi}_{1,m}((1-R_i^2)\bar{\Psi}_{1,m}^{-1}(\alpha_1)) \\
\leq\ & \sum_{i\in I_0}\sum_{r=1}^{d-1} \mathbb{E}\left[\mathbb{E}_{J_i}\left\{\frac{\bar{\Psi}_{1+2J_i}(\frac{V}{m}\bar{\Psi}_{1,m}^{-1}(\alpha_{r+1}))}{r+1} - \frac{\bar{\Psi}_{1+2J_i}(\frac{V}{m}\bar{\Psi}_{1,m}^{-1}(\alpha_r))}{r} \;\middle|\; \boldsymbol{Y}_{-i}, V\right\} \times \right. \\
& \left. \mathbb{I}\,(R(\boldsymbol{Y}_{-i},V)\geq r)\right] \\
=\ & \sum_{i\in I_0}\sum_{r=1}^{d-1} \mathbb{E}\left(\mathbb{E}_V\left[\left\{\frac{\bar{\Psi}_{1+2J_i}(\frac{V}{m}\bar{\Psi}_{1,m}^{-1}(\alpha_{r+1}))}{(r+1)\bar{\Psi}_{1+2J_i}(\frac{V}{m}\bar{\Psi}_{1,m}^{-1}(\alpha_r))} - \frac{1}{r}\right\} \times \right.\right. \\
& \left.\left. \mathbb{I}\,(R(\boldsymbol{Y}_{-i},V)\geq r)\,\bar{\Psi}_{1+2J_i}\left(\frac{V}{m}\bar{\Psi}_{1,m}^{-1}(\alpha_r)\right) \mid \boldsymbol{Y}_{-i}, J_i\right]\right).
\end{aligned}
$$

根据引理 4，我们注意到，给定 $(\boldsymbol{Y}_{-i}, J_i)$，

$$
\frac{\bar{\Psi}_{1+2J_i}(\frac{V}{m}\bar{\Psi}_{1,m}^{-1}(\alpha_{r+1}))}{(r+1)\bar{\Psi}_{1+2J_i}(\frac{V}{m}\bar{\Psi}_{1,m}^{-1}(\alpha_r))} - \frac{1}{r} \uparrow V > 0,
$$

此外，$\mathbb{I}\,(R(\boldsymbol{Y}_{-i}, V) \geq r) \downarrow V$，给定 $(\boldsymbol{Y}_{-i}, J_i)$。因此，根据 Kimball 不等式，我们得到

$$
\begin{aligned}
& \mathbb{E}_V\left[\left\{\frac{\bar{\Psi}_{1+2J_i}(\frac{V}{m}\bar{\Psi}_{1,m}^{-1}(\alpha_{r+1}))}{(r+1)\bar{\Psi}_{1+2J_i}(\frac{V}{m}\bar{\Psi}_{1,m}^{-1}(\alpha_r))} - \frac{1}{r}\right\} \times \right. \\
& \qquad \left. \mathbb{I}\,(R(\boldsymbol{Y}_{-i},V)\geq r)\,\bar{\Psi}_{1+2J_i}(\frac{V}{m}\bar{\Psi}_{1,m}^{-1}(\alpha_r)) \;\middle|\; \boldsymbol{Y}_{-i}, J_i\right] \\
\leq\ & \left\{\frac{\mathbb{E}_V\left[\bar{\Psi}_{1+2J_i}(\frac{V}{m}\bar{\Psi}_{1,m}^{-1}(\alpha_{r+1}))\right]}{r+1} - \frac{\mathbb{E}_V\left[\bar{\Psi}_{1+2J_i}(\frac{V}{m}\bar{\Psi}_{1,m}^{-1}(\alpha_r))\right]}{r}\right\} \times \\
& \qquad \frac{\mathbb{E}_V\left[\mathbb{I}\,(R(\boldsymbol{Y}_{-i},V)\geq r)\,\bar{\Psi}_{1+2J_i}(\frac{V}{m}\bar{\Psi}_{1,m}^{-1}(\alpha_r)) \;\middle|\; \boldsymbol{Y}_{-i}, J_i\right]}{\mathbb{E}_V\left[\bar{\Psi}_{1+2J_i}(\frac{V}{m}\bar{\Psi}_{1,m}^{-1}(\alpha_r)) \;\middle|\; \boldsymbol{Y}_{-i}, J_i\right]}.
\end{aligned}
$$

对于 BH 方法，其中 $\alpha_r = r\alpha_1$，对每个 $r = 1, \ldots, d$，该值小于或等于零，因为

$$
\begin{aligned}
& \frac{\mathbb{E}_V\left[\bar{\Psi}_{1+2J_i}(\frac{V}{m}\bar{\Psi}_{1,m}^{-1}(\alpha_{r+1}))\right]}{r+1} - \frac{\mathbb{E}_V\left[\bar{\Psi}_{1+2J_i}(\frac{V}{m}\bar{\Psi}_{1,m}^{-1}(\alpha_r))\right]}{r} \\
=\ & \alpha_1\left[\frac{\bar{\Psi}_{1+2J_i,m}(\bar{\Psi}_{1,m}^{-1}(\alpha_{r+1}))}{\alpha_{r+1}} - \frac{\bar{\Psi}_{1+2J_i,m}(\bar{\Psi}_{1,m}^{-1}(\alpha_r))}{\alpha_r}\right] \\
\leq\ & 0,
\end{aligned}
$$

根据引理 5。这就证明了 BH 方法所需的不等式：

$$
\mathrm{FDR} \leq \sum_{i\in I_0} \bar{\Psi}_{1,m}(w_i\bar{\Psi}_{1,m}^{-1}(\alpha_1)) \leq \alpha
$$

通过考察满足 (3.7) 的 α_1。定理 2 的证明至此完成。

4 进一步讨论

这里，我们讨论上述主要结果在一些特殊情形下的表现。

显然，当基础多元正态分布的协方差矩阵为对角矩阵，且其对角元素为已知数或已知元素的未知标量倍数时，我们提出的加权 BH 方法会退化为相应的原始 BH 方法，后者使用通常的、未加权的 z^2 或 t^2 值。对于其他情形，例如等相关多元正态分布以及线性回归中变量选择所产生的多元正态分布，所提出的方法如下所示。

4.1 等相关多元正态分布

设第 3.1 节和第 3.2 节中的 $\boldsymbol{X}$ 具有如下相关矩阵：$(1-\rho)I_d + \rho \mathbf{1}_d \mathbf{1}_d'$ [，其中$\mathbf{1}_d = (1, \ldots, 1)'$：$d \times 1$]，对于某个 $-1/(p-1) < \rho < 1$。在这种情况下，所有 d 复相关均与 $R^2 = (p-1)\rho^2/[1 + (p-2)\rho]$ 相同，因此定理 1 和定理 2 中的加权 BH 方法~基于相应的 P_i，其中

$$w_i = \frac{(1-\rho)[1+(p-1)\rho]}{1+(p-2)\rho}, \ \forall\, i = 1, \ldots, d.$$

4.2 错误发现率控制下的变量选择

考虑以下线性回归模型下的变量/特征选择问题：

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4.1}$$

其中，$\boldsymbol{Y}$是一个 n维响应向量，$\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d)$ 是一个秩为$d \le n$ 的 $n \times d$设计矩阵，其各列表示关于 d 个变量/特征$X_1, \ldots, X_d$的已知观测向量；$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$ 是与这些变量/特征对应的未知回归系数向量；$\boldsymbol{\epsilon} \sim N_d(0, \tau^2 \boldsymbol{I}_d)$是高斯噪声。

变量选择可表述为一个多重检验问题：同时对 $i = 1,\ldots,d$检验原假设H_i：$\beta_i = 0$及其备择假设 $\beta_i \neq 0$；根据基于回归系数某些估计值的多重检验程序，被拒绝的原假设所对应的变量将被选出/发现为重要变量。该多重检验程序的错误发现率是衡量选择中潜在错误的一项有力指标。

给定 $\boldsymbol{\beta}$ 的普通最小二乘估计为 $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_d)' = \boldsymbol{A}^{-1}\boldsymbol{X}'\boldsymbol{Y}$，其中 $\boldsymbol{A} = \boldsymbol{X}'\boldsymbol{X}$，其分布为 $N_d(\boldsymbol{\beta}, \tau^2 \boldsymbol{A}^{-1})$，并且独立于 $\widehat{\tau}^2 = [\|\boldsymbol{Y}\|^2 - \widehat{\boldsymbol{\beta}}' \boldsymbol{A} \widehat{\boldsymbol{\beta}}]/(n-d) \sim \tau^2 \chi^2_{n-d}/(n-d)$。因此，对于可用于变量选择的强大错误发现率控制程序，一个自然的选择是基于以下 t^2 值的 BH 方法：

$$T_i^2 = \frac{Z_i^2}{\widehat{\tau}^2}, \text{ where } Z_i^2 = \frac{\widehat{\beta}_i^2}{a^{ii}}, \text{ for } i = 1, \ldots, d, \tag{4.2}$$

其中 a^{ii} 是 $\boldsymbol{A}^{-1}$ 的第 i 个对角线元素。遗憾的是，正如引言中所指出的，无法从理论上保证 BH 方法（其原始形式）能够控制错误发现率。

在第3.2 节提出的方法基础上，我们现在可以提出一种用于变量选择且已证明能够进行错误发现率控制的新型 BH 型逐步上升方法。为此，我们首先注意到， R_i^2（ $\widehat{\beta}_i$ 与 $\widehat{\boldsymbol{\beta}}_{-i}$ 之间的复相关系数平方）等于 $1 - a_{ii}^{-1}/a^{ii}$。因此，我们可以正式描述所提出的变量选择错误发现率控制程序如下：

定义 1（用于变量选择的加权 BH 方法）令 $P_i = \bar{\Psi}_{1,n-d}(T_i^2)$ 为对应于 H_i的原始 p 值。使用 $\widetilde{P}_i = \bar{\Psi}_{1,n-d}(w_i^{-1}\bar{\Psi}_{1,n-d}^{-1}(P_i))$ 运行 BH 型逐步递增法，其中 $w_i = a_{ii}^{-1}/a^{ii}$, $i = 1, \ldots, d$ 和临界常数 $i\alpha_i$, $i = 1, \ldots, d$，且 α_1 满足条件：

$$\sum_{i=1}^{d} \bar{\Psi}_{1,n-d}(w_i \bar{\Psi}_{1,n-d}^{-1}(\alpha_1)) = \alpha.$$

5 总结

本文回答了以下问题，这是对摘要中所述内容的转述：对于多元正态均值的多重检验，BH 方法能否控制双侧 z 检验或 t 检验中的错误发现率，或者能否开发一种 BH 型逐步递增法来实现这一点？当然，假设相关系数已知，第 3 节提出的加权 BH 方法对此给出了肯定的答案。当相关系数和方差均未知时，如何针对基于 Hotelling's T^2 检验中边际 t^2 统计量的双侧 t 检验回答这一问题，是一个重要但颇具挑战性的开放问题。

所提出的加权 BH 方法可视为对 BH 方法依据底层相关结构进行的某种调整，类似于 Fithin & Lei (2022) 中的依赖性调整 BH （ DBH ）方法。不过，值得指出的是，尽管 DBH 完全考虑了底层相关结构，且检验效能相当高，但在实际应用中可能不够便于使用，因为它无法表示为闭式形式。该方法通过计算机辅助算法实现，需要进行大量计算。关于我们提出的、在错误发现率控制下进行变量选择时涉及双侧 t-检验的加权 BH 方法（第 4.2 节）的创新性，我们必须指出，尽管 Barber & Cand`es (2015) 和 Sarkar & Tang (2022) 提出的基于 knockoff 的错误发现率控制方法提供了有力的替代方案，但当 $d \leq n < 2d$ 时，这些方法并不适用；它们要求 n 大于或等于 $2d$。

参考文献

Barber, R. F. & Candès, E. J.(2015). 通过 knockoff 控制错误发现率。《统计年鉴》43, 2055–2085。

Benjamini, Y. (2010).发现错误发现率。《英国皇家统计学会杂志：B辑（方法论）》72, 405–416。

Benjamini, Y. & Hochberg, Y. (1995). 控制错误发现率：一种实用而有效的多重检验方法。《皇家统计学会学报：B辑（方法论）》57, 289–300。

Benjamini, Y. & Yekutieli, D. (2001). 相关正态分布双侧备择下多重检验中的错误发现率控制。《统计年鉴》29, 1165–1188.Blanchard, G. & Roquain, E. (2008). 错误发现率控制的两个简单充分条件。《电子统计学杂志》2, 963–992.Finner, H., Dickhaus, T. & Roters, M(2007). 依赖性与错误发现率：渐近性。《统计年鉴》35,1432–-1455.Finner, H., Roters, M & Strassburger, K.(2017). 依赖性下的Simes检验。《统计论文》58,775–-789.Fithian, W.& Lei, L. (2022)依赖性下错误发现率控制的条件校准。《统计年鉴》50,3091–3118.Karlin, S. (1968). 总正性。斯坦福大学出版社。Karlin, S. & Rinott, Y. (1980). 测度排序的类及相关的相关不等式：多元全正分布。《多元分析杂志》10, 467–498.Reiner-Benaim, A. (2007)通过BH程序控制双侧相关检验的错误发现率及其对基因表达数据分析的启示。《生物统计学杂志》,49,107–126.Sarkar, S. K. (2002). 逐步多重检验程序中的错误发现率若干结果。《统计年鉴》30,239–257.Sarkar, S. K. (2008).错误发现率控制方法。Sankhy ˜a：《印度统计学杂志》70, 135–168.Sarkar, S. K.& Tang C. (2022). 调整Benjamini-Hochberg方法，以控制敲除辅助变量选择中的错误发现率。《生物计量学》109, 1149–1155。