同济大学 TONGJI UNIVERSITY
地址：中国上海市四平路1239号 邮编：200092
1239 SIPING ROAD SHANGHAI CHINA 200092
电话（TEL）：+86 21- 传真（FAX）：+86 21-
网址（WEB）：www.tongji.edu.cn

Grouping, adjusting and Pooling (GAP) Procedure.

1. 改进动机：① 本文是提出一种利用稀疏信息去进行两样本的多重检验. 其中第一步为构建协变量序列，第二步为利用辅助协变量进行推断（三步算法，grouping, adjusting, pooling）其中本文GAP的显著优点是其能够处理各种相关结构（高维情形）

② 对于两样本检验，$\beta_1=\beta_2$. 传统方法为构建统计量 $\{T_1,\cdots,T_m\}$，选取阈值去控制 FDP. 这种方法忽视了 β_1,β_2 本身稀疏的信息，导致了信息上的损失. 而类似的改进方法有 CARS procedure. 其通过构建辅助协变量 $\{S_1,\cdots,S_m\}$ 去利用稀疏信息，进而进行综合推断. 但 CARS 不能应用在相关结构的检验问题中. 其只能用在独立结构的检验问题中.

③ 思想分两步，一为构建 (T_i,S_i) 可以准确地获取样本信息，二是有效地整合 T_i 和 S_i 的信息，（类似 US 方法）. 算法中的三步 - GAP. grouping 步将根据 S_i 将假设分为 K 组，分成 K 稀疏水平不同的若干组别；adjusting 步通过调整 p 值将结构信息考虑进来. 结构信息由各组别反映；pooling 步综合考虑调整后的所有p值，选取一个阈值去控制FDR

④ 根据 $\{S_1,\cdots,S_m\}$ 去进行分组会造成信息损失. 组别越少，分割变量越离散，信息损失越多. 组别越多，计算越复杂，计算成本越高. 实际选择 $K=3$ 或 4.

⑤ 贡献："① 处理了带有结构的多重检验 FDR 问题. ② 存在的方法分组依赖先验. GAP 由原样本构造 (T_i,S_i) 进行分组. ③ 不恰当的分组会扭曲原假设下 p 值的分布. ④ GAP 通过条件独立性的准则确保合适的分组以及 FDR 的渐近控制. ⑤ 新颖的 p 值加权方法，使得其适用于各种相依结构

⑥ 已知组别的多重检验方法：pool 和 seperate. 前者是忽视了组别标签且直接应用 BH 方法在所有检验上，经证明 pool 是不有效的，甚至可能 invalid. 后者是在各组别内单独应用 BH 方法最后将所有拒绝的假设并到一起，如 US 方法.

⑦ 经证明，选为的权重可以控制 FDR，但 power 会受当权重信息的影响. GAP 的权重是适当的.

⑧ 对于 non-null 比例的估计：用 $\hat{\pi}_i^{\lambda}=\frac{\#\{p_j>\lambda\}}{m(1-\lambda)}$ 来估计 non-null 比例，进而使用 $\hat{\pi}_i=(\varepsilon\vee\hat{\pi}_i^{\lambda})\wedge(1-\varepsilon)$ 使得其比例被约束在 $(\varepsilon,1-\varepsilon)$. $\varepsilon=10^{-5}$. （不需要调参）.

2. 方法步骤：Step 1：(Grouping) ① 计算 (T_i,S_i)

② 定义 $\lambda_0=-\infty,-4\sqrt{\log m}\le\lambda_1<\lambda_2<\cdots<\lambda_{K-1}\le 4\sqrt{\log m},\lambda_K=+\infty$.

令 $\Lambda=\{\lambda_1,\cdots,\lambda_{K-1}$ 是 $\mathcal{I}=\{\frac{i}{N}\sqrt{\log m}, i=-4N,\cdots,4N\}$，N 充分大 的一个子集，根据 S_i 进行分组. $G_l=\{1\le i\le m:\lambda_{l-1}<S_i\le\lambda_l\}$, $1\le l\le K$.

Step 2：(Adjusting). 定义 $m_l=|G_l|$，计算调整 p 值 $p_i^w=\min\{\frac{p_i}{w_i},1\}$, $i\in G_l$.

权重的计算：① initial adjusting：估计各组 non-null 比例 $\hat{\pi}_l$，则各组权重为 $w_l=\{\frac{K}{\sum_{l=1}^K}\frac{m_l\hat{\pi}_l}{1-\hat{\pi}_l}\}^{-1}\frac{m_l\hat{\pi}_l}{(1-\hat{\pi}_l)}$, $1\le l\le K$. 有 $p_i^w=\min(\frac{p_i}{w_i},1)$, $i\in G_l$

② further refining：考虑所有 Λ 的可能性，去寻求最优分组，$\Lambda=\phi$ 即为不分组. 组合所有 p_i^w 进行 BH procedure. 记 $k=\max\{i:\hat{p}_{(i)}^w\le\frac{\alpha}{m}i\}$

$\Lambda^*=\operatorname{argmax}_{\Lambda}\{k\}$. → 最优权重 w_i 基于最优分组计算而来. 即使得拒绝数最多的分组

同济大学 TONGJI UNIVERSITY
地址：中国上海市四平路1239号 邮编：200092
1239 SIPING ROAD SHANGHAI CHINA 200092
电话（TEL）：+86 21- 传真（FAX）：+86 21-
网址（WEB）：www.tongji.edu.cn

Step 3: (Pooling) 基于最优分组计算出最优权重，得出调整后的 p 值 p_i^w.
对所有调整后的 p 值进行 BH procedure，~~对于所有~~计算 $k=\max\{i: p_{(i)}^w \le \frac{\alpha i}{m}\}$，其中 $p_{(i)}^w$ 为 $p_i^w=\min(\frac{p_i}{w_i},1)$，$i=1,\cdots,m$，$i\in G_l$ 的次序统计量，把绝 $H_{(1)},\cdots,H_{(k)}$.

3. 理论假设

NOTE: GAP的FDP定义为 $FDP_{GAP}=\frac{\sum_{i\in H_0} I(p_i^w \le p_{(\hat{k}^w)}^w)}{\sum_{i=1}^m I(p_i^w \le p_{(\hat{k}^w)}^w)\vee 1}$，其中 $\hat{k}^w=\max\{i: p_{(i)}^w \le \frac{\alpha i}{m}\}$
而 $FDR=E(FDP_{GAP})$. 以下为理论证明必要的假设. 令 A_τ 为 H_0 的子集，$|A_\tau|=o(m^\nu)$，$\forall \nu>0$.
定义 $\tilde{H}_0=H_0\setminus A_\tau$，$n=n_1+n_2$

(A1) 渐近正态性：对检验统计量 $T_i, i\in\tilde{H}_0$，存在两个和 iid 随机样本 $\{z_{k,i}, i=1,\cdots,n_1\}$，$\{z_{k,i}, i=n_1+1,\cdots,n_1+n_2\}$，且 $E(z_{k,i})=0$. $E(e^{K z_{k,i}})<\infty$ 对某个 $K>0$ 使得 $M>0$，$\exists b_m$，$b_m=o((\log m)^{-\frac{1}{2}})$

$$P_{H_{0,i}}\left(|T_i-\frac{\sum_{k=1}^n z_{k,i}}{Var(\sum_{k=1}^n z_{k,i})^{\frac{1}{2}}}|\ge b_m\right)=O(m^{-M})$$

(A2) 弱相依性：定义 $(\rho_{ij,1})=R_1=corr(z_k)$，$k\in\{1,\cdots n_1\}$. $(\rho_{ij,2})=R_2=Corr(z_k)$. $k\in(n_1+1,\cdots n)$
$z_k=(z_{k,1},\cdots,z_{k,m})$. 有 $\max_{1\le i\le j\le m}|\rho_{ij,d}|\le\rho_d<1$，$d=1,2$. 进一步地，存在 $r>0$ 使得 $\max_{i\le i\le m}|T_i(r)|\le$
$T_i(r)=\{j: 1\le j\le m, |\rho_{ij,d}|\ge(\log m)^{-2-r}\}$，$d=1$ 或 2.

(A3) 渐近独立性：T_i 和 S_i 渐近独立，在 null 下，i.e. $\forall$ 常数 $M>0$. 有

$$P_{H_{0,i}}(|T_i|\ge t, |S_i|\ge\lambda)=(1+o(1))G_i(t)P(|N(0,1)+S_i|\ge\lambda)+O(m^{-M})$$

一致地对于 $0\le t\le 4\sqrt{\log m}$，$0\le\lambda\le 4\sqrt{\log m}$. $i\in\tilde{H}_0$，$s_i=E(S_i)$，$\forall 0\le y\le 4N$，固定的 N，有

$$P_{H_{0,i}}(|T_i|\ge t, |S_i|<\lambda_y)=(1+o(1))G_i(t)P(|N(0,1)+s_i|<\lambda_y)+O(m^{-M})$$

一致地对于 $0\le t\le 4\sqrt{\log m}$，$0\le y\le N$，$i\in\tilde{H}_0$. 其中 $\lambda_y=\frac{y}{N}\sqrt{\log m}$.

NOTE: (A1) 比较弱，只要求了 $T_i\sim AN(0,1)$，(A2) 指明不是所有统计量都强相关，(A3) 由 (T_i,S_i) 的构造易知. 定义 $S_\rho=\{i: 1\le i\le m, |\beta_{i,1}-\beta_{i,2}|\ge(\frac{(\log m)^{H_0}}{m})^{\frac{1}{2}}\}$

4. 理论结果.

定理一：设 $\rho>0$，$\delta>0$. $|S_0|\ge(\frac{1}{\Phi\alpha}+\delta)(\log m)^{\frac{3}{2}}$，$n_1\asymp n_2$，$m_0=|H_0|\ge cm$，$\exists c>0$. 只要在 (A1)-(A
且 $\log m=o(n^{\frac{1}{5}})$，$c>5$. 有 $\overline{\lim}_{n_1,p\to\infty} FDR_{GAP}\le\alpha$，且 $\lim_{n,m\to\infty} P(FDR_{GAP}\le\alpha+\varepsilon)=1$

定理二：在定理一的条件下. 有 $\text{亚}_{GAP}\ge\text{亚}_{BH}+o(1)$，$m\to\infty$.

Pf: 由 Thm1. $P(|\frac{FDP_{GAP}}{c\alpha'}-1|\ge\varepsilon)\to 0$. $c\in(0,1]$. $P(|\frac{FDP_{BH}}{\alpha'}-1|\ge\varepsilon)\to 0$. $\alpha'=\frac{m_0}{m}\alpha$
即有 $FDP_{GAP}+o_p(1)\le FDP_{BH}$，即 $\frac{\sum_{H_0}(I(p_i\le p_{(k)}))}{\sum_H(I(p_i\le p_{(k)}))}\ge\frac{\sum_{H_0} I(p_i^w\le p_{(\hat{k})}^w)}{\sum_H I(p_i^w\le p_{(\hat{k})}^w)}$，$\because A=\emptyset$ 时.
GAP 和 BH 等价. 又由 GAP 的最优组的定义. $\sum_H I(p_i^w\le p_{(\hat{k})}^w)\ge\sum_H I(p_i\le p_{(k)})$ 当且仅当 $A\ge B$
要证 $\sum_{H_1} I(p_i^w\le p_{(\hat{k}^w)}^w)\ge\sum_{H_1} I(p_i\le p_{(k)})$，即证 $(1-\alpha')A-(1-\alpha)B\ge 0$. $c\in(0,1]$，此为 c 的减
函数，$c=1$ 时取 min. 为 $(1-\alpha')(A-B)\ge 0$. 成立. 进而有

$$\frac{\sum_{H_1} I(p_i^w\le p_{(\hat{k}^w)}^w)}{|H_1|}\ge\frac{\sum_{H_1} I(p_i\le p_{(k)})}{|H_1|}+o_p(1),\Rightarrow \text{亚}_{GAP}\ge\text{亚}_{BH}+o(1)$$