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

Covariate-assisted ranking and screening — CARS procedure.

1. 改进目标，本文介绍了一种两样本的多重检验方法，构造辅助统计量以包含更多信息，提高Power. 传统方法一般将样本提炼为一向量（如p值）进而选择cutoff进行比较，但第一步中的提炼样本信息可能会造成显著的信息损失从而导出非最优（suboptimal）的方法. CARS通过辅助协变量序列提取结构信息.（本文考虑均值两样本稀疏情形）

① CARS分两步，构造 primary test statistics 和 auxiliary covariate，进而将二者结合构造多重检验方法. 思想是通过辅助统计量改变各假设的重要性，不再exchangeable，故关键的一步为利用辅助统计量去反映结构信息

② 相关的两个技巧为 screening 以及 grouping，前者将问题视为 unequal，通过 screen-and-clean 的方法. 先将以和 u_Y 的支撑筛选出来，进而再做检验. 而后者为将检验分组. divide-and-test. 如 Liu(2014) 的 DS 方法. 而 Screen 方法可理解为 narrow 了我们本来要考虑的，以提高 test 的准确性. 但样本的分割往往导致 power loss. 这两个方法都涉及到将一个连续变量分成K部分这导致无法充分利用辅助统计量的信息（在一个离散点切割，本应概率为0）

③ 本文主要关注 Gaussian 情形，数值表现 non-Gaussian 亦可. 设 T_{i1}, T_{i2} 为统计量，包含以下 $\theta_{i1}=I(u_{xi}\neq u_{yi})$ 和 $\theta_{i2}=I(u_{xi}\neq 0 \text{ 或 } u_{yi}\neq 0)$ 的信息，显然有 $\theta_{i2}=0 \Rightarrow \theta_{i1}=0$.

(1) 构造统计量：首先，T_{i1} 要获得 θ_{i1} 的信息，形如 $\bar{x}_{i1}-\bar{y}_{i1}$. 其次，获得 θ_{i2} 的信息，形如 $\bar{x}_{i2}+K\bar{y}_{i2}$，$K=\frac{n_2\sigma_{xi}}{n_1\sigma_{yi}}$，使 T_{i1}, T_{i2} 零相关. 最后进行标准化即可

$$(T_{i1}, T_{i2})=\sqrt{\frac{n_1n_2}{n}}\left(\frac{\bar{x}_{i1}-\bar{y}_{i1}}{\sigma_{pi}}, \frac{\bar{x}_{i1}+K\bar{y}_{i2}}{\sqrt{K}\sigma_{pi}}\right), \quad \sigma_{pi}^2=\frac{n_2}{n_1}\sigma_{xi}^2+\frac{n_1}{n}\sigma_{yi}^2$$

(2) 构造二元模型，考虑 T_1, T_2 的 jpdf: $f(t_{i1}, t_{i2})=\sum_{(j,k)\in(0,1)^2}\pi_{jk}f(t_{i1}, t_{i2}|\theta_{i1}=j, \theta_{i2}=k)$，$\pi_{jk}=P(\theta_{i1}=j, \theta_{i2}=k)$，令 $\pi_j=P(\theta_{i1}=1)$，根据 $\pi_{10}=0$. 以及零相关（正态独立），可简化该式. 见 Prop 1.

(3) 多重检验，利用 mFDR=FDR+o(1). Power 定义有很多. ETP（本文）. average power. missed discovery rate. FNP, FNR 等. 令 $T_{OR}^i(t_1, t_2)=P(\theta_{i2}=0|T_{i1}=t_1, T_{i2}=t_2)=\frac{q^*(t_2)f_{i0}(t_1)}{f(t_1, t_2)}$

其中 $f_{i0}(t_1)=f(t_1|\theta_{i2}=0)$，$q^*(t_2)=(1-\pi_1)f(t_2|\theta_{i2}=0)$

④ 注：oracle statistic T_{OR}^i 可看成是后验概率，$H_{i,0}$ 为真，给定 T_1, T_2 的概率，这可以作为对 $H_{i,0}$ 是否为 null 的显著性水平. 这种将 T_1, T_2 通过 pooling 成 T_{OR} 的方法没有信息损失. 若考虑最坏的情况，T_2 无任何信息：$n_x=n_y$，$\sigma_{xi}^2=\sigma_{yi}^2$，$u_{xi}=-u_{yi}$. T_{OR}^i 化为 Lfdr 统计量（即单变量模型下的 Adaptive z-value 方法）. CARS 也可以分为两步：ranking 和 thresolding.

⑤ 统计量 $T_{OR}=\frac{q^*(t_2)f_{i0}(t_1)}{f(t_1, t_2)}$ 的估计. $f_{i0}(t_1)$ 是当 $\theta_i=0$ 时的 mpdf，假设是已知的. $f(t_1, t_2)$ 是 jpdf 结合前文所提到的化简方式，利用非参的核估计方法进行估计，重点是对 $q^*(t_2)$ 进行估计 $q^*(t_2)=(1-\pi_1)f(t_2|\theta_{i2}=0)$

(1) 方法一，计数 $Q^*(t_2, h)=\#\{i: T_{i2}\in[t_2-\frac{h}{2}, t_2+\frac{h}{2}], \theta_{i2}=0\}/m$，$q^*(t_2)=\lim_{h\to 0}\frac{E(Q^*(t_2, h))}{h}$，但由于 θ_{i2} 未知无法估计. 文章利用了 Screening 方法对 null 进行了筛选，设 P_i 为 T_{i2} 对应的 p 值假设 P 值很大的认定为 null，即 $P_i>\tau$，则 $\theta_{i2}=0$. 由此估计. $Q^\tau(t_2, h)=\frac{\#\{i: T_{i2}\in[t_2-\frac{h}{2}, t_2+\frac{h}{2}], P_i>\tau\}}{m(1-\tau)}$

估计 $q^\tau(t_2)=\lim_{h\to 0}\frac{E(Q^\tau(t_2, h))}{h}=\frac{\int_{\{t_1: P_i>\tau\}}f(t_1, t_2)dt_1}{1-\tau}$，综上，$\hat{f}(t_1, t_2)=\frac{1}{m}\sum_{i=1}^{m}K_{h_1}(t_1-t_{i1})K_{h_2}(t_2-t_{i2})$.

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

其中，$K(t)$为核函数，h_1, h_2为带宽，设$\Gamma(\tau)=\{i: P_i \geq \tau\}$，有 $\hat{q}^{\tau}(t_2)=\frac{\sum_{i\in\Gamma(\tau)} K_{h_2}(t_2-t_{2i})}{m(1-\tau)}$

则 $\hat{f}_{OR}(t_1,t_2)=\frac{\hat{q}^{\tau}(t_2) f_{10}(t_1)}{\hat{f}(t_1,t_2)} \wedge 1$

(2) 方法二：另一种 $q^*(t_2)$ 的估计，是一种相合估计，且由最优化理论中得到，如下：

$\hat{q}^{*\tau}(t_2)=\frac{1}{k}\sum_{\gamma=1}^{k}\frac{\hat{s}_2-\hat{s}_1(T_\gamma-T_k)}{\hat{s}_2\hat{s}_0-\hat{s}_1^2}K_{h_T}(T_\gamma-T_k)\hat{q}^{\tau}_{T_\gamma}$，其中，$\hat{s}_r=\frac{1}{k}\sum_{\gamma=1}^{k}(T_\gamma-T_k)^r K_{h_T}(T_\gamma-T_k)$

$T_0,\cdots,T_k$为$(0,1)$上的分点，$\hat{q}^{\tau}_{T_\gamma}$为方法一的估计。此法为数得最小的偏差，在不增加过多方差的前提下。

二、方法步骤：

Step 1：计算检验统计量 T_{i1} 和辅助统计量 T_{i2}，$i=1,\cdots,m$

Step 2：估计 pool 统计量 $\hat{T}^{\tau,i}_{OR}$，并排序 $\hat{T}^{\tau,(1)}_{OR}\leq\cdots\leq\hat{T}^{\tau,(m)}_{OR}$

Step 3：令 $k=\max\{j: \frac{1}{j}\sum_{i=1}^{j}\hat{T}^{\tau,(i)}_{OR}\leq\alpha\}$，拒绝 $H_{(1)},\cdots,H_{(k)}$.

NOTE：这里的 $\frac{1}{j}\sum_{i=1}^{j}\hat{T}^{\tau,(i)}_{OR}$ 是对 mFDR 的估计，思想与 Adaptive z-value 一样。

三、理论结果

定理一（ORACLE）(1) 对 $0<\lambda\leq 1$，设 $Q_{OR}(\lambda)$ 为 $\{I(T^i_{OR}<\lambda): 1\leq i\leq m\}$ 的 mFDR，则 $Q_{OR}(\lambda)<\lambda$ 且 $Q_{OR}(\lambda)$ 为 λ 的非降函数。(2) 设选取 $\alpha<\bar{\alpha}=Q_{OR}(1)$，则 oracle 阈值 $\lambda_{OR}=\sup\{\lambda: Q_{OR}(\lambda)\leq\alpha\}$ 存在且唯一，$Q_{OR}(\lambda_{OR})=\alpha$。进而，定义 $\delta_{OR}=(\delta^i_{OR}, i=1,\cdots,m)$，其中 $\delta^i_{OR}=I(T^i_{OR}<\lambda_{OR})$，则 δ_{OR} 是最优的，即对 $\forall$ FDR$\leq\alpha$ 的检验类 $\mathcal{D}_\alpha$，$\delta\in\mathcal{D}_\alpha$，$ETP_\delta\leq ETP_{\delta_{OR}}$，且 $\mathcal{D}_\alpha$ 是包含基于 T_1, T_2 自身 mFDR$_\delta\leq\alpha$ 的检验。

证明：首先 $T^i_{OR}(t_1,t_2)=\frac{P(\theta_{i2}=0)f(t_1,t_2|\theta_{i2}=0)}{f(t_1,t_2)}=\frac{(1-\pi_1)f(t_1|\theta_{i2}=0)f(t_2|\theta_{i2}=0)}{f(t_1,t_2)}=\frac{q^*(t_2)f_0(t_1)}{f(t_1,t_2)}$

(1) ①证 $\alpha_t<t$，在 Adaptive z-value 文首已经证明。$Q_{OR}(\lambda)=\frac{E(T_{OR}I(T_{OR}<\lambda))}{E(I(T_{OR}<\lambda))}$ → 由 Prop 1 条件独立性。

设 $Q_{OR}(t)=\alpha_t$，由 mFDR 定义（$Q_{OR}(\lambda)$ 定义）。$E\{\sum_{i=1}^{m}(T^i_{OR}-\alpha_t)I(T^i_{OR}<t)\}=\alpha$ (A.1) 对 (t_1,t_2) 积分。由 (A.1) 得 $\alpha_t<t$。②证 $Q_{OR}(t)$ 对 t 单调。设 $Q_{OR}(t_j)=\alpha_j, j=1,2$。记 $t_1<t_2$，有 $\alpha_1\leq\alpha_2$。反证：若 $\alpha_1>\alpha_2$，则 $(T^i_{OR}-\alpha_2)I(T^i_{OR}<t_2)=(T^i_{OR}-\alpha_2)I(T^i_{OR}<t_1)+(T^i_{OR}-\alpha_2)I(t_1\leq T^i_{OR}\leq t_2)$

$\geq(T^i_{OR}-\alpha_1)I(T^i_{OR}<t_1)+(\alpha_1-\alpha_2)I(T^i_{OR}<t_1)+(T^i_{OR}-\alpha_1)I(t_1\leq T^i_{OR}<t_2)$

两边取期望，有 $E(\sum_{i=1}^{m}(T^i_{OR}-\alpha_2)I(T^i_{OR}<t_2))>0$，这与 (A.1) 矛盾。（期望大于0，∵$\alpha_1<t_1$）

(2) oracle 阈值 $t_{OR}=\sup\{t\in(0,1), Q_{OR}(t)\leq\alpha\}$，要证明 t_{OR} 可取到。$Q_{OR}(\lambda)=\frac{(1-p)G_{OR}(\lambda)}{G_{OR}(\lambda)}$ 为连续函数，又为单调函数，故有 $Q_{OR}(t_{OR})=\alpha$，$\alpha<\bar{\alpha}$，定义 δ_{OR}。令 δ^* 为任意决策 mFDR$(\delta^*)\leq\alpha$，则有 $E(\sum_{i=1}^{m}(T^i_{OR}-\alpha)\delta^i_{OR})=0$，$E(\sum_{i=1}^{m}(T^i_{OR}-\alpha)\delta^i_{*})\leq\alpha$ (A.2) 由此有 $E(\sum_{i=1}^{m}(\delta^i_{OR}-\delta^*)(T^i_{OR}-\alpha))\geq 0$ (A.3)

考虑变换 $f(x)=\frac{x-\alpha}{1-x}$（↑），$\delta^i_{OR}\Leftrightarrow\delta^i_{OR}=I(\frac{T^i_{OR}-\alpha}{1-T^i_{OR}}<\lambda_{OR})$，$\lambda_{OR}=f(t_{OR})$。

∵ $\alpha<t_{OR}<1$，则有 $\lambda_{OR}>0$。注意有 $\begin{cases}\delta^i_{OR}>\delta^i_*, & T^i_{OR}-\alpha-\lambda_{OR}(1-T^i_{OR})<0\\ \delta^i_{OR}<\delta^i_*, & T^i_{OR}-\alpha-\lambda_{OR}(1-T^i_{OR})>0\end{cases}$ 分类讨论即可 $\delta_{OR}=1$ 和 $\delta_{OR}=0$。(A.4)

$\forall i$，$(\delta^i_{OR}-\delta^i_*)(T^i_{OR}-\alpha-\lambda_{OR}(1-T^i_{OR}))\leq 0$，对 i 求和取期望，$E(\sum(\delta^i_{OR}-\delta^i_*)(T^i_{OR}-\alpha-\lambda_{OR}(1-T^i_{OR}))\leq 0$

结合 (A.3)、(A.4)，有 $\lambda_{OR}E(\sum(\delta^i_{OR}-\delta^i_*)(1-T^i_{OR}))\geq E(\sum(\delta^i_{OR}-\delta^i_*)(T^i_{OR}-\alpha))\geq 0$，由 Adaptive z-value 知 $ETP=E(\sum\theta_{i1}\delta_i)=E(\sum(1-T^i_{OR})\delta_i)$，由此得，∵$\lambda_{OR}>0$，$ETP_{\delta_{OR}}\geq ETP(\delta^*)$

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

定理=(Data-Driven) 条件 (1) $E\|\hat{q}^{\tau}-q^{\tau}\|^2\to 0$. (1)' $E\|\hat{q}^{*}-q^{*}\|^2\to 0$

(2) $E\|\hat{f}-f\|^2=E[\iint(\hat{f}(t_1,t_2)-f(t_1,t_2))^2\,dt_1dt_2]\to 0$.

(1) 若条件(1),(2)成立. 则mFDR和FDR在CARS下控制在$\alpha+o(1)$.

2) 若条件(1)',(2)成立. 用$\hat{q}^{*}$估计q^{*}, 则CARS自动FDR水平为$\alpha+o(1)$, 进而, 用ETP_{CARS}和ETP_{OR}表示CARS和oracle的ETP. 则有 $ETP_{CARS}/ETP_{OR}=1+o(1)$