基于累积量（半不变量）的混合 RNA-seq 样本的协方差矩阵的假设检验量的构造

目录

■Part I 累积量（半不变量）理论回顾 1

■Part II 问题背景的回顾和转化 3

■Part III 回顾 CSnet Estimator 论文中一阶原点矩和二阶原点矩的估计 5

■Part IV 三阶和四阶高阶（混合）原点矩的估计思路 6

■Part V 三阶和四阶高阶（混合）原点矩估计的理论推导： 7

■Part VI：协方差估计量的方差 Var(**σjj′(k)**)的 MSE 一致估计方法： 10

■Part VII：Max-entry 假设检验量的构造 11

■参考文献 11

■Part I 累积量（半不变量）理论回顾

一、 随机变量：

1、设随机变量 ξ_1 的 n 阶矩 m_n 存在，把它的特征函数 $\varphi(t)$ 按 Taylor 级数展开而得重要公式：

$$\varphi(t)=1+\sum_{k=1}^{n}\frac{m_k}{k!}(it)^k+o(t^n). \qquad (1)$$

由特征函数性质知：$\log\varphi(t)$ 的前 n 级导数在 0 点存在，令

$$\chi_k=\frac{1}{i^k}[\frac{d^k}{dt^k}\log\varphi(t)]_{t=0} \quad (k\le n), \qquad (2)$$

称 χ_k 为随机变量 ξ 的 k 阶累积量（半不变量），显然有

$$\log\varphi(t)=\sum_{k=1}^{n}\frac{\chi_k}{k!}(it)^k+o(t^n). \qquad (3)$$

通过比较（1）和（3）可得到累积量 χ_k 与矩 m_n 之间的关系，详见[1][1].

2、累积量具有下列性质：

a. 如随机变量 ξ_1、ξ_2 独立，且它们的 k 阶累积量 $\chi_k^{(1)}$，$\chi_k^{(2)}$ 存在，则 $\xi=\xi_1+\xi_2$ 的 k 阶累积量 χ_k 为：

$$\chi_k=\chi_k^{(1)}+\chi_k^{(2)}. \qquad (4)$$

b．在变换 $\eta=\xi+b$ 下，累积量（除一阶外）不变，b 为常数.

二、随机向量：

1、设 $\xi=(\xi_1,\cdots,\xi_k)$ 是随机向量，函数 $\varphi_\xi(\mathbf{t})=\mathbf{E}e^{i(t,\xi)}$，$\mathbf{t}=(t_1,\cdots,t_k)$, 是 ξ 的特征函数.

令 $\mathbf{E}|\xi_i|^n<\infty, i=1,\cdots,k.$，则将 $\xi=(\xi_1,\cdots,\xi_k)$ 展开为 Taylor 级数，得：

$$\varphi_\xi(t_1,\cdots,t_k)=\sum_{v_1+\cdots+v_k\le n}\frac{i^{v_1+\cdots+v_k}}{v_1!\cdots v_k!}m_\xi^{(v_1,\cdots,v_k)}t_1^{v_1}\cdots t_k^{v_k}+o(|t|^n), \qquad (5)$$

其中 $|t|=|t_1|+\cdots+|t_k|$, $m_\xi^{(v_1,\cdots,v_k)}=\mathbf{E}\xi_1^{v_1}\cdots\xi_k^{v_k}$ 是 $v=(v_1,\cdots,v_k)$ 阶（混合）矩.

由特征函数性质知，$\ln\varphi_\xi(t_1,\cdots,t_k)$ 可以在 $(0,\cdots,0)$ 表示为 Taylor 公式：

$$\ln\varphi_\xi(t_1,\cdots,t_k)=\sum_{v_1+\cdots+v_k\le n}\frac{i^{v_1+\cdots+v_k}}{v_1!\cdots v_k!}s_\xi^{(v_1,\cdots,v_k)}t_1^{v_1}\cdots t_k^{v_k}+o(|t|^n), \qquad (6)$$

其中 $s_\xi^{(v_1,\cdots,v_k)}=\frac{1}{i^{v_1+\cdots+v_k}}[\frac{\partial^{v_1+\cdots+v_k}}{\partial t_1^{v_1}\cdots\partial t_k^{v_k}}\ln\varphi_\xi(t_1,\cdots,t_k)]_{t_1=\cdots=t_k=0}$.

我们称 $s_\xi^{(v_1,\cdots,v_k)}$ 为随机向量 $\xi=(\xi_1,\cdots,\xi_k)$ 的 $v=(v_1,\cdots,v_k)$ 阶（混合）累积量（半不变量）.

2、累积量的性质：

a．如随机向量 ξ、η 独立，且它们的 $(v_1,\cdots,v_k)$ 阶累积量 $s_\xi^{(v_1,\cdots,v_k)}$，$s_\eta^{(v_1,\cdots,v_k)}$ 存在，则 $\xi+\eta$ 的 $(v_1,\cdots,v_k)$ 阶累积量为：

[1] [3]王梓坤.概率论基础及其应用.第 3 版.北京师范大学出版社.2007

$$s_{\xi+\eta}^{(v_1,\cdots,v_k)} = s_{\xi}^{(v_1,\cdots,v_k)} + s_{\eta}^{(v_1,\cdots,v_k)}.$$

（7）

b.累积量与矩的关系　(实际应用主要是计算公式（10）)

为简化记号，对具有非负整数分量的向量$v=(v_1,\cdots,v_k)$，令

$$v!=v_1!\cdots v_k!,\ |v|=v_1+\cdots+v_k,\ t^v=t_1^{v_1}\cdots t_k^{v_k}, s_{\xi}^{(v)}=s_{\xi}^{(v_1,\cdots,v_k)}, m_{\xi}^{(v)}=m_{\xi}^{(v_1,\cdots,v_k)}. \quad (8)$$

设$\xi=(\xi_1,\cdots,\xi_k)$是随机向量，$\mathbf{E}|\xi_i|^n<\infty, i=1,\cdots,k.$ 则有：

$$m_{\xi}^{(v)} = m_{\xi}^{(v_1,\cdots,v_k)} = \sum_{\lambda(1)+\cdots+\lambda(q)=v} \frac{v_1!\cdots v_k!}{\lambda(1)!\cdots\lambda(q)!}\prod_{p=1}^{q} s_{\xi}^{(\lambda(p))},$$

（9）

$$s_{\xi}^{(v)} = s_{\xi}^{(v_1,\cdots,v_k)} = \sum_{\lambda(1)+\cdots+\lambda(q)=v} (-1)^{q-1}(q-1)!\frac{v_1!\cdots v_k!}{\lambda(1)!\cdots\lambda(q)!}\prod_{p=1}^{q} m_{\xi}^{(\lambda(p))},$$

（10）

其中$\sum_{\lambda(1)+\cdots+\lambda(q)=v}$表示对于一切满足$|\lambda(p)|>0, \lambda(1)+\cdots+\lambda(q)=v$的”有序非负整数分量的向量$\lambda(p)$”求和。

更一般的形式，

$$m_{\xi}^{(v)} = m_{\xi}^{(v_1,\cdots,v_k)} = \sum_{\{r_1\lambda(1)+\cdots+r_x\lambda(x)=v\}} \frac{1}{r_1!\cdots r_x!}\frac{v_1!\cdots v_k!}{(\lambda(1)!)^{r_1}\cdots(\lambda(x)!)^{r_x}}\prod_{j=1}^{x}[s_{\xi}^{(\lambda(j))}]^{r_j},$$

（11）

$$s_{\xi}^{(v)} = s_{\xi}^{(v_1,\cdots,v_k)} = \sum_{\{r_1\lambda(1)+\cdots+r_x\lambda(x)=v\}} \frac{(-1)^{q-1}(q-1)!}{r_1!\cdots r_x!}\frac{v_1!\cdots v_k!}{(\lambda(1)!)^{r_1}\cdots(\lambda(x)!)^{r_x}}\prod_{j=1}^{x}[m_{\xi}^{(\lambda(j))}]^{r_j},$$

（12）

其中$\sum_{\{r_1\lambda(1)+\cdots+r_x\lambda(x)=v\}}$表示对于一切满足$|\lambda(j)|>0, r_1\lambda(1)+\cdots+r_x\lambda(x)=v$的”有序正整数$r_j$的数组”求和.

■Part II 问题背景的回顾和转化

仍设 n 维混合 RNA-序列数据样本 across p genes: $x_1,\ldots,x_n \in R^p$.并考虑 K 种细胞的混合采样模型：

$$\boldsymbol{x}_i = \sum_{k=1}^{K} \pi_{ik} \boldsymbol{x}_i^{(k)}, \qquad (1)$$

其中$x_i^{(k)}(k \in K)$之间独立：

$$\mathbb{E}(\boldsymbol{x}_i) = \sum_{k=1}^{K} \pi_{ik} \boldsymbol{\mu}^{(k)}, \qquad \mathrm{Cov}(\boldsymbol{x}_i) = \sum_{k=1}^{K} \pi_{ik}^2 \boldsymbol{\Sigma}^{(k)}. \qquad (2)$$

累积量观点：我们换个观点看（2），其中（2）能够拆开的原因从累积量理论的观点看，累积量满足对独立随机变量和的可加性，且因为 1 阶单随机变量累积量恰好是数学期望，2 阶单随机变量的累积量恰好是方差，（1，1）阶双随机变量的累积量恰好是协方差，即如下三个定理：

Theorem 1.1 (将会涉及到的 4 阶以下累积量)

我们设$s_{\xi}^{(v_1,\cdots,v_k)}$为随机向量$\xi=(\xi_1,\cdots,\xi_k)$的$v=(v_1,\cdots,v_k)$阶（混合）累积量（半不变量），现考虑$\xi=(X,Y)$, 则：

（1） $s_{\xi}^{(1,0)}=EX$ (2) $s_{\xi}^{(0,1)}=EY$

(3) $s_{\xi}^{(1,1)}=Cov(X,Y)$ (4) $s_{\xi}^{(2,0)}=D(X)$，$s_{\xi}^{(0,2)}=D(Y)$

（5）$s_X^{(3)} = E(X-EX)^3$ (仍恰为 3 阶中心矩)= $m_3 - 3m_1m_2 + 2m_1^3$

（6）$s_X^{(4)} = m_4 - 4m_1m_3 - 3m_2^2 + 8m_1^2m_2 - 6m_1^4$ (不是 4 阶中心矩了)

（7）$s_{X,Y}^{(2,1)} = EX^2Y + 4（EX）^2EY - EX^2EY -$ (EXY) EX

(8)$s_{X,Y}^{(2,2)} = EX^2Y^2 - 2EXY^2EX - 2EX^2YEY -$

$EX^2EY^2 - 4（EXY）^2 + 2（EX）^2EY^2 + 2EX^2(EY)^2 - 24(EX)^2(EY)^2$

Theorem 1.2（累积量的独立可加性）

如随机向量ξ、η独立，且它们的$(v_1,\cdots,v_k)$阶累积量$s_{\xi}^{(v_1,\cdots,v_k)}$，$s_{\eta}^{(v_1,\cdots,v_k)}$存在，则$\xi+\eta$的$(v_1,\cdots,v_k)$阶累积量为：

$$s_{\xi+\eta}^{(v_1,\cdots,v_k)} = s_{\xi}^{(v_1,\cdots,v_k)} + s_{\eta}^{(v_1,\cdots,v_k)}.$$

Theorem 1.3（累积量的常数齐次性）

对随机向量ξ，且它的$(v_1,\cdots,v_k)$阶累积量$s_\xi^{(v_1,\cdots,v_k)}$存在，则 cξ的$(v_1,\cdots,v_k)$阶累积量为

$$\mathbf{c}^{\sum_i \upsilon_i} s_\xi^{(v_1,\cdots,v_k)}$$

(注：由于 K 种细胞采样的混合比例（$\pi_{i1},\pi_{i2}\ldots,\pi_{iK}$）不同，不同实验所得的$x_i \in R^p$间不一定 i.i.d, 仅仅满足独立)

然而，不幸的是，当累积量的阶高于 3 阶时，唯有单个随机变量的 3 阶中心矩仍保持恰好是一个累积量。

为简化表达，不失一般性，当 p>1 时，直接用不同的字母 X，Y 表示这 p genes 中的任意某两个不同的基因表达数据 X=($X_1, X_2 \ldots X_n$), Y=($Y_1, Y_2 \ldots Y_n$)。

即用 $X_1, X_2 \ldots X_n \in R$; $Y_1, Y_2 \ldots Y_n \in R$ 分别表示 X 基因数据分量和 Y 基因数据分量所分别对应采集到的 n 个混合数据.

■问题的形式总结：

依据混合原点矩的线性和不同实验次数间的独立性两条性质，对 $\mathrm{Var}(\bar{\sigma}_{jj'}^{(k)})$可以进行直接的拆分计算，对它的估计可以转化为以下两种情形：

（1）当$j \neq j'$时，不妨设j对应的基因为 X, j'对应的基因为 Y

归结为对广义二次型方差 $\mathrm{Var}(X^TAY)$的估计,最终归结为对 4 阶以下混合原点矩 $E(X_iX_jY_kY_l)$的估计和讨论，（其中脚标i, $j,k,l \in [n]$ 注意到脚标不同时,即不同次实验的数据），可以根据不同次实验间的独立性拆开。

（2）当$j = j'$时，不妨设j, j'对应的基因都为 X:

归结为对二次型方差 $\mathrm{Var}(X^TAX)$的估计,最终归结为对 4 阶以下混合原点矩 $E(X_iX_jX_kX_l)$的估计和讨论。

■最终转化为要去估计的各阶（混合）原点中心矩包括：

（I） 所有可能的一阶原点矩：$E(X_s)$（$s \in [n]$）；$E(Y_s)$ （$s \in [n]$）

（II） 所有可能的二阶(混合)原点矩：$E(X_s^2)$（$s \in [n]$）；$E(Y_s^2)$ （$s \in [n]$）$E(X_sY_s)$ （$s \in [n]$）

（III） 所有涉及到的三阶(混合)原点矩：$E(X_s^2Y_s)$（$s \in [n]$）；$E(X_sY_s^2)$（$s \in [n]$）；$E(X_s^3)$（$s \in [n]$）；$E(Y_s^3)$ ($s \in [n]$)

（IV） 所有涉及到的四阶(混合)原点矩：$E(X_s^2Y_s^2)$（$s \in [n]$）；$E(X_s^4)$（$s \in [n]$）；

■Part III 回顾 CSnet Estimator 论文中一阶原点矩和二阶原点矩的估计

■(I) 一阶原点矩的估计：

其中（I），（II）的估计已在 CSnet Estimator[1.SZZ21]论文中完成，对(I)中的一阶原点矩, 回顾：

Denote $\boldsymbol{y}_j = (x_{1j}, \ldots, x_{nj})$ and $\boldsymbol{D} = (\pi_{ik})_{n\times K}$. Equation (3) entails estimation of the cell-type-specific mean $\boldsymbol{\mu}^{(k)}$ via

$$\hat{\mu}_j^{(k)} = \left[\left(\boldsymbol{D}^\top \boldsymbol{D}\right)^{-1} \boldsymbol{D}^\top \boldsymbol{y}_j\right]_k, \quad j \in [p],\ k \in [K], \tag{5}$$

$$\mathbb{E}(\boldsymbol{x}_i) = \sum_{k=1}^{K} \pi_{ik}\boldsymbol{\mu}^{(k)}, \qquad \mathrm{Cov}(\boldsymbol{x}_i) = \sum_{k=1}^{K} \pi_{ik}^2 \Sigma^{(k)}. \tag{2}$$

设$\hat{\mu}_{ij}$为参数$E(x_{ij})$的估计，则由（5）和（2），直接有估计：

$$\hat{\mu}_{ij} = \sum_{k=1}^{K} \pi_{ik}\, \hat{\mu}_j^{(k)} \qquad （A）$$

■（II）二阶(混合)原点矩的估计：

首先回顾：

where $[\boldsymbol{x}]_k$ is the kth entry in $\boldsymbol{x} \in \mathbb{R}^K$. Let $\hat{z}_{ij} = x_{ij} - \sum_{k=1}^{K} \pi_{ik}\hat{\mu}_j^{(k)}$ and $\hat{\boldsymbol{z}}_j = (\hat{z}_{1j}, \ldots, \hat{z}_{nj}) \in \mathbb{R}^n$. Denoting $\boldsymbol{H} = (\pi_{ik}^2)_{n\times K}$, equation (4) entails estimation of the cell-type-specific covariance $\sigma_{jj'}^{(k)}$ via

$$\hat{\sigma}_{jj'}^{(k)} = \left[\left(\boldsymbol{H}^\top \boldsymbol{H}\right)^{-1} \boldsymbol{H}^\top (\hat{\boldsymbol{z}}_j \circ \hat{\boldsymbol{z}}_{j'})\right]_k, \quad j, j' \in [p],\ k \in [K], \tag{6}$$

首先设$\hat{\sigma}_{ijj'}$为参数$E(x_{ij} - Ex_{ij})(x_{ij'} - Ex_{ij'})$,即 $\mathrm{Cov}(x_{ij}, x_{ij'})$的估计量，则有：

$$\hat{\sigma}_{ijj'} = \sum_{k=1}^{K} \pi_{ik}{}^2\, \hat{\sigma}_{jj'}^{(k)} \qquad （B）$$

再设$\hat{\beta}_{ijj'}$为参数$Ex_{ij}x_{ij'}$的估计。由恒等式$\mathrm{Cov}(X, Y) = EXY - EXEY$和（A），则有：

$$\hat{\beta}_{ijj'} = \hat{\sigma}_{ijj'} + \hat{\mu}_{ij}\hat{\mu}_{ij'} \qquad （C）$$

■Part IV 三阶和四阶高阶（混合）原点矩的估计思路

■（III）和（IV）中所涉及高阶混合矩的估计：

（I），（II）的估计能够通过（2）和对应的回归方程（3），（4）得到，虽然是一阶矩和二阶矩的情况，但可以看成之前提过的 **Theorem1.1** 和 **Theorem1.2,** 以及 **Theorem 1.3** 的特殊情形， 这里$z_{ij} = xij - E(xij)$。

$$\mathbb{E}(\boldsymbol{x}_i) = \sum_{k=1}^{K} \pi_{ik}\boldsymbol{\mu}^{(k)}, \qquad \mathrm{Cov}(\boldsymbol{x}_i) = \sum_{k=1}^{K} \pi_{ik}^2 \boldsymbol{\Sigma}^{(k)}. \tag{2}$$

Letting $\boldsymbol{\mu}^{(k)} = (\mu_1^{(k)}, \ldots, \mu_p^{(k)})$ and $\boldsymbol{\Sigma}^{(k)} = (\sigma_{jj'}^{(k)})_{p\times p}$, (1) and (2) together imply

$$x_{ij} = \sum_{k=1}^{K} \pi_{ik}\mu_j^{(k)} + z_{ij}, \quad j \in [p], \tag{3}$$

$$z_{ij}z_{ij'} = \sum_{k=1}^{K} \pi_{ik}^2 \sigma_{jj'}^{(k)} + \epsilon_{ijj'}, \quad j, j' \in [p], \tag{4}$$

where $\mathbb{E}(z_{ij}) = 0$ and $\mathbb{E}(\epsilon_{ijj'}) = 0$. This formulation facilitates an efficient least squares

所以一个自然的想法是：

想估计（III），（IV）中所涉及的更高阶的（混合）原点矩的估计，可以考虑先用同样的回归方法估计相应的整体的高阶累积量，然后代入之前已估计出来的的低阶的（混合）原点矩，最终得到，所想要的高阶（混合）原点矩的估计。

■Part V 三阶和四阶原点矩估计的理论推导：

本节中，我们将以 $\mathbf{E}(x_{ij}^3)$（$i \in [n], j \in [p]$）和 $E(x_{ij}^4)$（$i \in [n]$, $j \in [p]$）为例子，推导它们的 MSE 一致估计。

首先进行随机变量 X 的（3）阶，（4）阶累积量$s_{X,Y}^{(3,0)}s_{X,Y}^{(4,0)}$,简写为$s_X^{(3)}$，$s_X^{(4)}$的计算，始终假设$m_i = E(X^i)$，且相应累积量存在，则：

（1） $s_X^{(3)} = E(X - EX)^3$ (仍恰为 3 阶中心矩)= $m_3 - 3m_1m_2 + 2m_1^3$

（2） $s_X^{(4)} = m_4 - 4m_1m_3 - 3m_2^2 + 8m_1^2m_2 - 6m_1^4$ (不是 4 阶中心矩了)

注意：

$$s_X^{(3)} = m_3 - 3m_1m_2 + 2m_1^3 = E(X^3 - 3m_1m_2 + 2m_1^3)$$

$$\begin{aligned} s_X^{(4)} &= m_4 - 4m_1m_3 - 3m_2^2 + 8m_1^2m_2 - 6m_1^4 \\ &= EX^4 - 4m_1m_3 - 3m_2^2 + 8m_1^2m_2 - 6m_1^4 \\ &= E(X^4 - 4m_1m_3 - 3m_2^2 + 8m_1^2m_2 - 6m_1^4) \end{aligned}$$

其中$(X^3 - 3m_1m_2 + 2m_1^3)$以及$(X^4 - 4m_1m_3 - 3m_2^2 + 8m_1^2m_2 - 6m_1^4)$地位上

对等于如下回归方程（3）和（4）中的$z_{ij}z_{ij'}=(xij-E(xij))(xij'-E(xij'))$部分.

Letting $\boldsymbol{\mu}^{(k)}=(\mu_1^{(k)},\ldots,\mu_p^{(k)})$ and $\Sigma^{(k)}=(\sigma_{jj'}^{(k)})_{p\times p}$, (1) and (2) together imply

$$x_{ij} = \sum_{k=1}^{K}\pi_{ik}\mu_j^{(k)}+z_{ij},\quad j\in[p], \tag{3}$$

$$z_{ij}z_{ij'} = \sum_{k=1}^{K}\pi_{ik}^2\sigma_{jj'}^{(k)}+\epsilon_{ijj'},\quad j,j'\in[p], \tag{4}$$

where $\mathbb{E}(z_{ij})=0$ and $\mathbb{E}(\epsilon_{ijj'})=0$. This formulation facilitates an efficient least squares

下面我们以 $E(x_{ij}^3)$（$i\in[n],j\in[p]$）和 $E(x_{ij}^4)$（$i\in[n],\ j\in[p]$）的估计为例推导其基于高阶累积量的回归估计，类似（3），（4）的回归顺序，要想先估计 4 阶矩，先要先估计 3 阶矩。

首先设$w_{X_{ij}}^{(3)}=x_{ij}{}^3-3m_{ij1}m_{ij2}+2m_{ij1}^3,(m_{ijk}=E(x_{ij}^k))$

而注意到在（3），（4）回归方程的估计中已经得到m_{ij1}，m_{ij2}的估计，即一阶原点矩和二阶（混合）原点矩的估计：

$$\hat{\mu}_{ij}=\sum_{k=1}^{K}\pi_{ik}\hat{\mu}_j^{(k)}\ldots(A)$$

$$\hat{\beta}_{ijj'}=\hat{\sigma}_{ijj'}+\hat{\mu}_{ij}\hat{\mu}_{ij'}\ldots(C)$$

这里重新记估计量$\hat{\mu}_{ij}$为符号$\hat{m}_{ij1}$，$\hat{\beta}_{ijj}$为符号$\hat{m}_{ij2}$，再设

$$h_{X_{ij}}^{(4)}=x_{ij}{}^4-4m_{ij1}m_{ij3}-3m_{ij2}^2+8m_{ij1}^2m_{ij2}-6m_{ij1}^4(m_{ijk}=E(X_{ij}^k))$$

则
$$w_{X_{ij}}^{(3)}=\sum_{k=1}^{K}\pi_{ik}{}^3s_{Xj}^{(k)(3)}+\varepsilon_{ij},\ j\in[p]\quad(5)$$

$$h_{X_{ij}}^{(4)}=\sum_{k=1}^{K}\pi_{ik}{}^4s_{Xj}^{(k)(4)}+\varepsilon'_{ij},\ j\in[p]\quad(6)$$

其中 $E(\varepsilon_{ij})=0,\ E(\varepsilon'_{ij})=0.$

记$\hat{w}_{X_{ij}}^{(3)}=x_{ij}{}^3-3\hat{m}_{ij1}\hat{m}_{ij2}+2\hat{m}_{ij1}^3$，类似有$\hat{h}_{X_{ij}}^{(4)}$,但注意$\hat{h}_{X_{ij}}^{(4)}$中的$\hat{m}_{ij3}$需要在（5）做完后才有相应的估计量，就如（3）和（4）之间的关系，先有期望才能估计方差。

设$\hat{\boldsymbol{w}}_{\boldsymbol{X_j}}^{(\boldsymbol{3})}=(\hat{w}_{X_{1j}}^{(3)},\hat{w}_{X_{2j}}^{(3)},\ldots,\hat{w}_{X_{nj}}^{(3)})$　（$n\times1$）

$\hat{\boldsymbol{h}}_{\boldsymbol{X_j}}^{(\boldsymbol{4})}=(\hat{h}_{X_{1j}}^{(3)},\hat{h}_{X_{2j}}^{(3)},\ldots,\hat{h}_{X_{nj}}^{(3)})$　（$n\times1$）

于是：

设 $W=(\pi_{ik}{}^{3})n\times K;\quad V=(\pi_{ik}{}^{4})n\times K.$

则$s_{Xj}^{(k)(3)}$的估计为：

$$\hat{s}_{Xj}^{(k)(3)}=[(W^{T}W)^{-1}W^{T}\hat{\mathbf{w}}_{\mathbf{X_j}}^{(\mathbf{3})}]_{\mathbf{k}}\ \ \mathbf{j}\in[\mathbf{p}],\mathbf{k}\in[\mathbf{K}]$$

从而：

$$\hat{s}_{Xij}^{(3)}=\sum_{k=1}^{K}\pi_{ik}{}^{3}\hat{s}_{Xj}^{(k)(3)}$$

Theorem 5.1

（1）最后首先得到 $E(X_{ij}^{3})$的一个 MSE 一致估计为：

$$\hat{m}_{ij3}=\hat{s}_{Xij}^{(3)}+3\hat{m}_{ij1}\hat{m}_{ij2}-2\hat{m}_{ij1}^{3}$$

同理可得到：

$$\hat{s}_{Xj}^{(k)(4)}=[(V^{T}V)^{-1}V^{T}\hat{\mathbf{h}}_{\mathbf{X_j}}^{(\mathbf{4})}]_{\mathbf{k}}\ \ \mathbf{j}\in[\mathbf{p}],\mathbf{k}\in[\mathbf{K}]$$

从而：

$$\hat{s}_{Xij}^{(4)}=\sum_{k=1}^{K}\pi_{ik}{}^{4}\hat{s}_{Xj}^{(k)(4)}$$

（2）进而得到 $E(X_{ij}^{4})$的一个 MSE 一致估计为：

$$\hat{m}_{ij4}=\hat{s}_{Xij}^{(4)}+4\hat{m}_{ij1}\hat{m}_{ij3}3\hat{m}_{ij2}^{2}-8\hat{m}_{ij1}^{2}\hat{m}_{ij2}+6\hat{m}_{ij1}^{4}$$

对于（III），（IV）中其他的混合 3 阶原点矩，4 阶原点矩也可类似的给出回归估计。

最后再计算两个要使用的高阶累积量作为本节的结束：

Example 5.1

随机变量 X, Y 的$s_{X,Y}^{(2,1)}$，$s_{X,Y}^{(2,2)}$的计算，设$m_{i}=E(X^{i})$，若相应累积量存在，则：

（1） $s_{X,Y}^{(2,1)}=EX^{2}Y+4（EX）^{2}EY-EX^{2}EY-（EXY）EX$

（2） $s_{X,Y}^{(2,2)}=EX^{2}Y^{2}-2EXY^{2}EX-2EX^{2}YEY-$

$$EX^{2}EY^{2}-4（EXY）^{2}+2（EX）^{2}EY^{2}+2EX^{2}(EY)^{2}-24(EX)^{2}(EY)^{2}$$

■Part VI：协方差估计量的方差 $Var(\bar{\sigma}_{jj'}^{(k)})$的 MSE 一致估计方法：

固定 k，对任意第 k 类细胞，仍用 X，Y 表示 p 个基因中的某两个基因，类似于 $E(x_{ij}^3)$和 $E(x_{ij}^4)$（$i \in [n], j \in [p]$）情形的估计，原始论文中通过一阶累积量数学期望$s_{X,Y}^{(1,0)(k)} = EX$ 和$s_{X,Y}^{(0,1)(k)} = EY$给出了EX 和 EY的估计，通过二阶累积量协方差$s_{X,Y}^{(1,1)(k)} = Cov(X,Y) = E(X-EX)(Y-EY)$给出了 EXY 的估计。方才我们用三阶累积量$s_{X,Y}^{(3,0)(k)} = E(X-EX)^3$ (仍恰为 3 阶中心矩)$= m_3 - 3m_1m_2 + 2m_1^3$给出了EX^3的估计，四阶累积量$s_{X,Y}^{(4,0)(k)} = m_4 - 4m_1m_3 - 3m_2^2 + 8m_1^2m_2 - 6m_1^4$ (不是 4 阶中心矩)给出了EX^4的估计

我们可以依次通过对$s_{X,Y}^{(2,1)(k)}$，$s_{X,Y}^{(2,2)(k)}$，以及对称的$s_{X,Y}^{(1,2)(k)}$的回归方程 $(k \in [K])$去估计相应的累积量，再代入前面已经被估计出的低阶累积量，最余给出其他所需要的混合中心矩 $E(X_s^2Y_s)$（$s \in [n]$）；$E(X_sY_s^2)$ （$s \in [n]$）$E(X_s^2Y_s^2)$（$s \in [n]$）的估计，用$\widehat{E}(X_s^mY_s^l)(m,l = 0,1,2\ldots)$表示相应的估计量，从而通过这些混合局估计可以如下构造出一种 $Var(\bar{\sigma}_{jj'}^{(k)})$的 MSE 一致估计：

又由于当 $j \neq j'$时，若 $Var(\bar{\sigma}_{jj'}^{(k)}) = F_{jj'}(E(X_{ij}^1X_{ij'}^2), E\left(X_{ij}^2X_{ij'}^2\right), E(X_{ij}^2X_{ij'}^1), E(X_{ij}^1), E\left(X_{ij'}^2\right), E\left(X_{ij}^2\right), E(X_{ij'}), , E\left(X_{ij}\,X_{ij'}\right), i \in [n])$其中$F_{jj'}$指的是将标准差估计量 $\bar{\sigma}_{jj'}^{(k)}$的方差直接依据混合矩的线性和不同实验间的独立性将其直接计算拆开，表示为以关于X_{ij} 和$X_{ij'}$ $(i \in [n])$高阶混合中心矩为自变量的函数。

当$j = j'$时，若 $Var(\bar{\sigma}_{jj'}^{(k)}) = Var\left(\bar{\sigma}_{jj}^{(k)}\right) = F_{jj}$ $(E(X_{ij'}^2), E\left(X_{ij}^4\right), E(X_{ij'}^1), E(X_{ij}^3), i \in [n])$ 其中 Function 指的是将标准差估计量 $\bar{\sigma}_{jj'}^{(k)}$的方差直接依据混合矩的线性和不同实验间的独立性将其直接计算拆开，表示为以关于X_{ij} （$i \in [n]$）的高阶中心矩为自变量的函数。

Theorem 6.1

若用$\widehat{E}(X_{ij}^sX_{ij'}^t)$表示$E(X_{ij}^sX_{ij'}^t)$的用累积量回归所得的 MSE 估计，用$\widehat{Var}(\bar{\sigma}_{jj'}^{(k)})$表

示 $\mathrm{Var}(\bar{\sigma}_{jj'}^{(k)})$的一个一致估计量，则：

$$\widehat{\mathrm{Var}}(\bar{\sigma}_{jj'}^{(k)}) = \begin{cases} F_{jj'}\left(\hat{E}\left(X_{ij}^{1}X_{ij'}^{2}\right), \hat{E}\left(X_{ij}^{2}X_{ij'}^{2}\right), \hat{E}\left(X_{ij}^{2}X_{ij'}^{1}\right), \hat{E}(X_{ij}^{1}), \hat{E}\left(X_{ij'}^{2}\right), \hat{E}(X_{ij}^{2}), \hat{E}\left(X_{ij'}^{1}\right), \hat{E}\left(X_{ij}^{1}X_{ij'}^{1}\right), i \in [n]\right) & \text{if } j \neq \\ F_{jj}\left(\hat{E}\left(X_{ij'}^{2}\right), \hat{E}(X_{ij}^{4}), \hat{E}\left(X_{ij'}^{1}\right), \hat{E}(X_{ij}^{3}), i \in [n]\right) & \text{if } j = \end{cases}$$

而对于$\widehat{\mathrm{Var}}(\bar{\sigma}_{jj'}^{(k)})$，$F_{jj'}$是一些 MSE 估计的直接加和，因为每一项加和的单项是至多两个混合矩的 MSE 的估计的乘积所以$\widehat{\mathrm{Var}}(\bar{\sigma}_{jj'}^{(k)})$仍是一个 MSE 一致估计。

■Part VII：Max-entry 假设检验量的构造

Theorem 7.1

从而第 k 类细胞，我们考虑两组 bulk RNA 数据所得的协方差矩阵的假设检验 $H_0: \sum_1^{(k)} = \sum_2^{(k)}$ ，可以用如下 Max-entry Testing 进行检验,定义

$$M_{jj'}^{(k)} := \frac{(\bar{\sigma}_{jj'(1)}^{(k)} - \bar{\sigma}_{jj'(2)}^{(k)})^2}{\widehat{\mathrm{Var}}\left(\bar{\sigma}_{jj'(1)}^{(k)}\right) + \widehat{\mathrm{Var}}\left(\bar{\sigma}_{jj'(2)}^{(k)}\right)} \quad (1 \le j \le j' \le p)$$

则我们考虑如下 Max-entry 检验量：

$$M_n^{(k)} := \max_{1 \le j \le j' \le p} M_{jj'}^{(k)} = \max_{1 \le j \le j' \le p} \frac{(\bar{\sigma}_{jj'(1)}^{(k)} - \bar{\sigma}_{jj'(2)}^{(k)})^2}{\widehat{\mathrm{Var}}\left(\bar{\sigma}_{jj'(1)}^{(k)}\right) + \widehat{\mathrm{Var}}\left(\bar{\sigma}_{jj'(2)}^{(k)}\right)}$$

从[2.JASA13]的工作知道它是一个呈现 cumulative 分布的假设检验量。

■参考文献

1. Estimating cell-type-specific gene co-expression networks from bulk gene expression data with an application to Alzheimer's disease, Chang Su, Jingfei Zhang, Hongyu Zhao, doi: https://doi.org/10.1101/2021.12.21.473558

2.Cai, Tony, et al. "Two-Sample Covariance Matrix Testing and Support Recovery in High-Dimensional and Sparse Settings." Journal of the American Statistical Association, vol. 108, no. 501, 2013, pp. 265–77. JSTOR, http://www.jstor.org/stable/23427527. Accessed 25 Sep. 2022.

3.王梓坤.概率论基础及其应用.第 3 版.北京师范大学出版社.2007