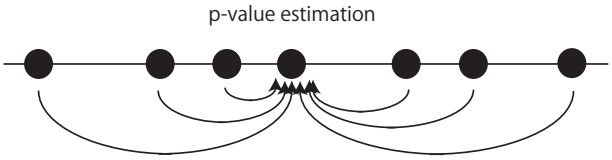
遺伝子発現差解析における並べ替え技法による P値の推定

予測発見戦略研究センター　遺伝子多様性解析グループ
数理・推論研究系　学習推論グループ
准教授　　藤澤　洋徳

1　はじめに

遺伝子発現データの特徴の一つは繰り返し実験回数が少ないことである．そこで，ある遺伝子の遺伝子発現差解析を行うのに，他の遺伝子のデータを援用するという試みがアドホックに提案されている．本研究では，アドホックにではなく，**他の遺伝子のデータを「うまく」援用できるための本質的な条件は何であるか**を適切に捉え，このタイプの研究に統一的な観点を与えたい．そして，そのような条件の下で，どのような検定が最適であるかを考えたい．最適な検定は，アドホックに提案された検定よりも，検出力が高いことを期待できる．



2　並べ替え技法の援用

二つのグループを表す確率変数を X と Y とする．対応する平均を μ_X と μ_Y で表す．遺伝子発現差がない（平均に差がない）という仮説は $H : \mu_X = \mu_Y$ で表される．二つのグループからの標本を $X_1, \ldots, X_n$ と $Y_1, \ldots, Y_m$ で表す．次のようなベクトル記号も用意しておく：$\boldsymbol{X} = (X_1, \ldots, X_n)'$．$\boldsymbol{Y} = (Y_1, \ldots, Y_m)'$．$\boldsymbol{Z} = (\boldsymbol{X}', \boldsymbol{Y}')'$．いま，検定統計量 $T(\boldsymbol{Z})$ が十分に大きければ，帰無仮説 H が棄却されると考えよう．実現値が $\boldsymbol{Z} = \boldsymbol{z}$ $(t = T(\boldsymbol{z}))$ であったときの p 値を次のように考えよう：

$$p = \Pr(T(\boldsymbol{Z}) > t \,|\, H).$$

標本 $\boldsymbol{Z}$ の並べ替え標本を $\boldsymbol{Z}_b^\dagger$ $(b = 1, \ldots, B)$ で表すことにする．ここで p 値を次のように単純に推定することにしよう：$(1/B)\sum_{b=1}^{B} I(T(\boldsymbol{Z}_b^\dagger) > t)$．ここで $I(A)$ は定義関数である．いま X と Y の標本平均と標本分散を $\bar{X}$, $\bar{Y}$, S_X^2, S_Y^2 で表すことにする．プールされた標本分散を $S^2 = \{(n-1)S_X^2 + (m-1)S_Y^2\}/(n+m-2)$ とおく．たとえば，検定統計量として，$T_a(\boldsymbol{Z}) = |\bar{X} - \bar{Y}|/\sqrt{(1/n + 1/m)S^2}$ を使った検定は，適当な想定の下では，一様最強力不偏検定であり，この検定統計量がしばしば使われている．

しかし，遺伝子発現差解析においては，少し問題が生じる．遺伝子発現データは繰り返し実験回数が少ないので，十分な数の並べ替え標本を得にくい．遺伝子発現データで典型的な $n = m = 4$ のとき，異なる $\boldsymbol{Z}_b^\dagger$ の個数は高々$B = 8!/(4!4!) = 56$ である．そこで，ある遺伝子における p 値を推定するときに，他の遺伝子のデータも援用するアイデアがしばしば使われている．なぜなら，遺伝子発現データでは，繰り返し実験回数は少ないけれども，遺伝子数 G は非常に大きいこと（たとえば数千）が一般的だからである．

第 g 番目の遺伝子における第 b 番目の並べ替え標本を $\boldsymbol{Z}^{\dagger}_{bg}$ とおく．そして，第 g 番目の遺伝子における p 値を次で推定する：

$$\hat{p}\left(T(\boldsymbol{Z}^{\dagger}); t_g\right) = \frac{1}{BG}\sum_{b=1}^{B}\sum_{g'=1}^{G} I(T(\boldsymbol{Z}^{\dagger}_{bg'}) > t_g),$$

ここで $t_g = T(\boldsymbol{z}_g)$ は第 g 番目の遺伝子に対応する実現値である．ただし，このような単純な援用は，数理的に問題がある．それは，他の遺伝子のデータを援用するときに，帰無仮説が正しい遺伝子と正しくない遺伝子が混在していることであり，さらに，p 値は帰無仮説が正しいところで定義されている点である．

3 Pan (2003) のアイデア

Pan (2003) は，標本を分割してうまく使うことで，前節の問題点を克服している．まずは標本を次のように分割する：$\boldsymbol{X}_{(1)} = (X_1, \ldots, X_{n_1})'$, $\boldsymbol{X}_{(2)} = (X_{n_1+1}, \ldots, X_n)'$, $n_2 = n - n_1$, $\boldsymbol{Y}_{(1)} = (Y_1, \ldots, Y_{m_1})'$, $\boldsymbol{Y}_{(2)} = (Y_{m_1+1}, \ldots, Y_m)'$, $m_2 = m - m_1$．さらに，対応する標本平均と標本分散を次で表すことにする：$\bar{X}_{(1)}$, $\bar{X}_{(2)}$, $\bar{Y}_{(1)}$, $\bar{Y}_{(2)}$, $S^2_{X(1)}$, $S^2_{X(2)}$, $S^2_{Y(1)}$, $S^2_{Y(2)}$．検定統計量としてはアドホックに次を用意した：

$$T_{Pan}(\boldsymbol{Z}) = \left|\frac{\bar{X}_{(1)} + \bar{X}_{(2)}}{2} - \frac{\bar{Y}_{(1)} + \bar{Y}_{(2)}}{2}\right| \Bigg/ \sqrt{\frac{1}{4}\left(\frac{S^2_{X(1)}}{n_1} + \frac{S^2_{X(2)}}{n_2} + \frac{S^2_{Y(1)}}{m_1} + \frac{S^2_{Y(2)}}{m_2}\right)}.$$

帰無統計量としては次を用意する：$T^{null}_{Pan}(\boldsymbol{Z}) = T_{Pan}\left(\boldsymbol{X}_{(1)}, -\boldsymbol{X}_{(2)}, -\boldsymbol{Y}_{(1)}, \boldsymbol{Y}_{(2)}\right)$．いま X と Y が平均パラメータに関して対称な分布をもつと仮定しよう．**このとき，帰無統計量 $\boldsymbol{T^{null}_{Pan}}$ は，帰無仮説が正しいかどうかに関係なく，検定統計量 $\boldsymbol{T_{Pan}}$ の帰無分布（帰無仮説が正しいと仮定したときの分布）と同じ分布をもつ．**これは次のように表現できる：

$$p_{Pan} = \Pr(T_{Pan}(\boldsymbol{Z}) > t \,|\, H) = \Pr(T^{null}_{Pan}(\boldsymbol{Z}) > t).$$

ここで**並べ替え標本を各グループ毎に作ることにしよう**．いま，$\boldsymbol{X}^*$ と $\boldsymbol{Y}^*$ を，それぞれ，$\boldsymbol{X}$ と $\boldsymbol{Y}$ の並べ替え標本とする．このとき，$\boldsymbol{Z}^* = (\boldsymbol{X}^{*\prime}, \boldsymbol{Y}^{*\prime})'$ に対して，**たとえ帰無仮説が正しくなくても，$\boldsymbol{T(Z^*)}$ は $\boldsymbol{T(Z)}$ と同じ分布をもつ．**この点が通常の並べ替え標本を使った場合と違う現象である．結果的に，第 g 番目の遺伝子の p 値の推定量として，$\hat{p}\left(T^{null}_{Pan}(\boldsymbol{Z}^*); t_g\right)$ を提案できる．

4 最適な検定

前節の Pan (2003) における性質を一般的に整理して，p 値の推定のときに，他の遺伝子のデータを利用しやすくするために，次のような条件を満たす検定統計量 T のクラスを考えることにしよう：

$$p = \Pr(T(\boldsymbol{Z}) > t \,|\, H) = \Pr(h(\boldsymbol{Z}) > t) \text{ を満たす帰無統計量 } h \text{ が存在する．}$$

さらに，母集団分布として適当なクラスを考えることで，一様最強力不偏検定を導出することができる．たとえば，あるクラスの下では，Pan (2003) に似た統計量が導出される．このとき p 値は簡単に $\hat{p}\,(h(\boldsymbol{Z}^*); t_g)$ で推定できる．また，別のクラスでは別のタイプの検定統計量を考えることができる．