

Neyman-Pearson補題の多重検定への拡張

野間 久史 データ科学研究系 助教

研究目的:

多重検定における『総合的な検出力を最大化する検定方式』とは？
 ≪Neyman-Pearson補題の多重検定への拡張≫

Motivation:

- FDR (False Discovery Rate) [1, 2] に基づく多重検定の方法論は、マイクロアレイ実験などのゲノム研究で、広く利用されている。
- 与えられた情報から、可能な限り、多くの関連遺伝子(臨床アウトカムと関連のある遺伝子)を選択したい。
 ⇒ 古典的な検定論における『最強力検定』に対応する『総合的な検出力を最大化する検定方式』とは？

Storey [3]によるOptimal Discovery Procedure (ODP)

Storey [3]による最適性の基準

一定のEFP (Expected False Positive; Average Type-I Error Rate)のもとで、ETP (Expected True Positive; Average Power)を最大化する検定方式

⇒ 多重検定における最強力検定(平均検出力を最大化)

Storey [3]の補題(ODP)

m 個の検定を同時に行うものと考え、それぞれの検定におけるデータセットが $\mathbf{x}_k = (x_{k1}, \dots, x_{kn})^T$ で与えられ ($k = 1, \dots, m$)、共通の確率空間の上に定義される確率ベクトルであるとする。帰無仮説が正しいものを m_0 個、対立仮説が正しいものを m_1 個とし、一般性を失うことなく、それぞれの確率密度関数を $f_1(\mathbf{x}), \dots, f_{m_0}(\mathbf{x})$ と $g_{m_0+1}(\mathbf{x}), \dots, g_m(\mathbf{x})$ と表記する。

上記の設定のもとで、以下の検定統計量が、Storeyの最適性の基準を達成する検定方式を与える。

$$S_{\text{ODP}}(\mathbf{x}) = \frac{g_{m_0+1}(\mathbf{x}) + g_{m_0+2}(\mathbf{x}) + \dots + g_m(\mathbf{x})}{f_1(\mathbf{x}) + f_2(\mathbf{x}) + \dots + f_{m_0}(\mathbf{x})}$$

$S_{\text{ODP}}(\mathbf{x})$ による多重検定の実践的な問題点

(1) Neyman-Pearson補題と同じく、 $S_{\text{ODP}}(\mathbf{x})$ がそれぞれの検定における、真の確率密度関数によって構成されている (m 個の検定それぞれで、個別にパラメータを推定する必要がある;ゲノム研究では、 m は数万)

(2) $S_{\text{ODP}}(\mathbf{x})$ の分母・分子は、「帰無仮説が真」「対立仮説が真」のもので分類されている

⇒ 実際の推測では、これらを適当な方法で推定する必要がある。

Storey *et al.* [4] は、一般化尤度比検定の考え方を基にした方法を提案しているが、上記の2つの推定の問題は、本質的に難しく、現実的な設定のもとで、必ずしも理論的なETPの上限を達成しない。

参考文献

- [1] Benjamini Y, Hochberg Y. *J. Royal Statist. Soc. B* 1995; **57**: 289-300.
- [2] Storey JD. *J. Royal Statist. Soc. B* 2002; **64**: 479-498.
- [3] Storey JD. *J. Royal Statist. Soc. B* 2007; **69**: 347-368.
- [4] Storey JD, Dai JY, Leek JT. *Biostatistics* 2007; **8**: 414-432.
- [5] Noma H, Matsui S. *Statist. Med.* 2012; **31**: 165-176.
- [6] Cao J, Xie XJ, Zhang S, *et al. BMC Bioinformatics* 2009; **10**: 5.
- [7] Kirschner-Schwabe R, Lottaz C, Todling J, *et al. Clin. Cancer Res.* 2006; **12**: 4553-4561.
- [8] Desmedt C, Piette F, Loi S, *et al. Clin. Cancer Res.* 2007; **13**: 3207-3214.
- [9] Tusher VG, Tibshirani R, Chu G. *Proc. Natl. Acad. Sci. USA* 2001; **98**: 5116-5121.
- [10] Cui XG, Hwang JTG, Qiu J, *et al. Biostatistics* 2005; **6**: 59-75.
- [11] Efron B, Tibshirani R, Storey JD, Tusher V. *J. Am. Statist. Assoc.* 2001; **96**: 1151-1160.

経験ベイズ法によるODP (Noma and Matsui [5])

StoreyのODPは、Steinの縮小推定量と同じく、「検定間の情報を共有することによって、全体の検出力を改善している」

⇒ 経験ベイズ法としての定式化

ODP補題(Noma and Matsui [5])

m 個の検定のデータセット $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m$ が、それぞれ確率密度関数 $f(\mathbf{x} | \theta_k, \psi_k)$ から生成されるものとする (θ_k : 関心のあるパラメータ, ψ_k : 局外パラメータ; $k = 1, \dots, m$)。パラメータ (θ_k, ψ_k) は、帰無仮説・対立仮説が真であるものそれぞれで交換可能であるとして、以下の分布に従うものとする。

$$\begin{aligned} \text{Null: } (\theta_k, \psi_k) &\sim G_0(\theta, \psi | \xi_0) & (*) \\ \text{Alternative: } (\theta_k, \psi_k) &\sim G_1(\theta, \psi | \xi_1) \end{aligned}$$

ξ_0, ξ_1 はそれぞれの変量効果の分布の超パラメータを表す。これらの設定のもとで、Storey [3] のODPの基準を達成する方式は、以下のように導かれる。

補題(Noma and Matsui [5]):

変量効果モデル (*)のもとで、以下の検定統計量が、Storeyの最適性の基準を達成する検定方式を与える。

$$R_{\text{ODP}}(\mathbf{x}) = \frac{E_{G_1}[f(\mathbf{x} | \theta, \psi)]}{E_{G_0}[f(\mathbf{x} | \theta, \psi)]} = \frac{\int f(\mathbf{x} | \theta, \psi) dG_1(\theta, \psi | \xi_1)}{\int f(\mathbf{x} | \theta, \psi) dG_0(\theta, \psi | \xi_0)}$$

⇒ 最適な検定方式は『平均尤度比』『周辺尤度比』を基準にして構成される

$R_{\text{ODP}}(\mathbf{x})$ は、 $S_{\text{ODP}}(\mathbf{x})$ の推定の問題を本質的に回避することができる

- (1) 個々の検定ごとにパラメータを推定する必要はなく、 ξ_0, ξ_1 を正確に推定できればよい
- (2) 検定統計量の構成において、帰無仮説の真偽を推定する必要はない

経験ベイズ法

⇒ 超パラメータ ξ_0, ξ_1 を、最尤推定値に置き換えた方式

適用事例: 白血病, 乳がんの臨床研究

白血病・乳がんの2つの臨床研究におけるマイクロアレイ実験 [7, 8] で、FDR 5%, 10% のもとでの検定を行い、既存の代表的な方法との結果(有意になった遺伝子数)を比較した。FDRの制御の方法は、Storey [2] による方法をすべての方法で用いて、公平な比較になるようにしている。

	Kirschner-Schwabe <i>et al.</i> [7]		Desmedt <i>et al.</i> [8]	
	FDR 5%	FDR 10%	FDR 5%	FDR 10%
Noma and Matsui [5] 法	385	670	66	168
Storey <i>et al.</i> [4] 法	234	464	16	142
Cao <i>et al.</i> [6] 法	188	332	55	152
SAM [9]	324	517	18	124
Student <i>t</i> -test	246	419	13	148
Shrunken <i>t</i> -test [10]	267	446	13	133
Local fdr [11]	349	508	12	158