Comparison of Large-scale Multiple Testing Procedures with Two Examples of Microarrays

By Meng Du

Department of Statistics, University of Toronto

Joint Work with Muni Srivastava

May 11, 2006
Hamilton, Canada

Outline

1. Introduction
   - 1.1 The Problem
   - 1.2 Motivation
2. The Methods
   - 2.1 Notations
   - 2.2 Type I Error Rates
   - 2.3 Large-scale Multiple Testing Procedures
3. Examples
   - 3.1 The Datasets
   - 3.2 Results

1 Introduction

1.1 The Problem

- Let p-dimensional random vectors

$$\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \cdots, \boldsymbol{x}_{iN_i} \overset{iid}{\sim} N_p(\boldsymbol{\mu}_i, \Sigma), \quad i = 1, 2,$$

  - mean vectors $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \cdots, \mu_{ip})', \ i = 1, 2 : p \times 1$
  - covariance matrix $\Sigma = (\sigma_{ij}) : p \times p$, $\Sigma > 0$ (positive definite)

- Interested in testing the multiple hypotheses that

$$H_i : \ \mu_{1i} = \mu_{2i} \quad \text{vs.} \quad A_i : \mu_{1i} \neq \mu_{2i}, \quad i = 1, 2, \cdots, p,$$

- Assume p is very large, particularly $N = N_1 + N_2 < p$ (number of observtions smaller than the dimension)

1.2 Motivation

- DNA microarrays ($N < p$): thousands of gene expression values are measured on relatively fewer subjects.
  - Example: 6817 (p) genes vs. 72 (N) subjects in Dudoit, et. al. (2002)
- When p is large, classical testing procedures are TOO CONSERVATIVE.
  - Example: Bonferroni procedure rejects each H_j at significane level of α/p such that

$$\begin{aligned} FWER &= \Pr\{at\ least\ one\ H_j\ is\ falsely\ rejected\} \\ &\leq \sum_{j=1}^{p} \Pr\{H_j\ is\ falsely\ rejected\} = \alpha \end{aligned}$$

  - When p is large, FWER could be much smaller than α. Bonferroni procedure is too conservative.

2 Large-scale Multiple Testing Procedures.

2.1 Notations

- The sample mean vectors

$$\bar{\boldsymbol{x}}_i = N_i^{-1} \sum_{j=1}^{N_i} \boldsymbol{x}_{ij} = (\bar{x}_{i1}, \bar{x}_{i2}, \cdots, \bar{x}_{ip})', \ i = 1, 2$$

  - $\bar{x}_{ij}$: the sample mean of the jth component of the i group

- The pooled sample covariance matrix

$$S = n^{-1} \sum_{i=1}^{2} \sum_{j=1}^{N_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)' = (s_{ij}), \ n = N - 2$$

  - s_{ij}: the pooled sample covariance of the ith and the jth component;
  - s_{jj}: the common sample variance of the jth component.

- The test statistic for the single hypothesis $H_j : \mu_{1j} = \mu_{2j}$

$$T_j = \frac{N_1 N_2}{(N_1 + N_2)} \frac{(\bar{x}_{1j} - \bar{x}_{2j})^2}{s_{jj}} \sim F_{1,n}, \text{ under } H_j$$

  - $F_{1,n}$ is the F-distribution with 1 and n degrees of freedom
  - $n = N_1 + N_2 - 2 = N - 2$.

- The p-value of T_j when $T_j = t_j$:

$$p_j = \mathrm{P}\{T_F > t_j | H_j : \mu_{1j} = \mu_{2j}\}$$

  - T_F is an $F_{1,n}$ random variable

- Order the p-values $p_1, p_2, \cdots, p_p$ and the corresponding hypotheses $H_1, H_2, \cdots, H_p$ as

$$\begin{aligned} p_{(1)} &\leq p_{(2)} \leq \cdots \leq p_{(p)} \\ H_{(1)}, \ H_{(2)}, &\quad \cdots, \ H_{(p)} \end{aligned}$$

2.2 Type I Error Rates

Let V be the number of false rejections and R be the total number of rejections.

- Family-wise Error Rate (FWER)

$$\mathrm{P}\{V \geq 1\}.$$

  That is, the rate of at least one false rejection of a true null hypothesis.

- False Discovery Rate (FDR)

$$\mathrm{E}[V/R] = \mathrm{E}[V/R|R > 0]\mathrm{P}(R > 0).$$

  That is the expected proportion of false rejctions among all rejections.

- γFWER

$$\mathrm{P}\{V \geq \gamma \times p\},$$

  where $1/p \leq \gamma < 1$. That is the rate of at least $100\gamma\%$ false rejections among all the hypotheses.

- Connections between FWER, FDR, and γFWER

$$\begin{aligned} FWER &\geq FDR \\ FWER &= FDR,\ if\ all\ H_j,\ j = 1, 2, \cdots, p,\ are\ true. \\ FWER &\geq \gamma FWER \\ FWER &= \gamma FWER,\ if\ \gamma = 1/p \end{aligned}$$

  Therefore,

- Control of FWER is MORE STRINGENT than control of FDR or γFWER
- Control of FDR or γFWER without controlling FWER may potentially INCREASE the POWER of the testing procedure.

2.3 Large-scale Multiple Testing Procedures

- The Procedures

Method	Reject	Control*	Restriction
Bonf	$H_j, \quad if\ p_j \leq \alpha/p$	FWER	No
Holm	$H_{(j)},\ if \forall i \leq j,\ p_{(i)} \leq \frac{\alpha}{(p-i+1)}$	FWER	No
Hochberg	$H_{(j)},\ if \exists i \geq j,\ p_{(i)} \leq \frac{\alpha}{(p-i+1)}$	FWER	ind. or pos dep**
FDRp	$H_{(j)},\ if \exists i \geq j,\ p_{(i)} \leq \frac{j\alpha}{p}$	FDR	ind or pos dep**
cFDRp	$H_{(j)},\ if \exists i \geq j,\ p_{(i)} \leq \frac{j\alpha}{p\sum_{k=1}^{p} 1/k}$	FDR	No
γFWERp	$H_j, \quad if\ p_j \leq \gamma\alpha$ ***	γFWER	No

* controls FWER or FDR at nominal level α.

** independence or positive dependence b.w. test statistics $T_j, j = 1, \cdots, p$

*** $\frac{1}{p} \leq \gamma < 1$ is preselected, e.g. γ=0.05;$\gamma = 1/p$, γFWERp$\Longleftrightarrow$Bonf.

- Conservativeness

  $P1 \prec P2$ denotes procedure $P1$ is more conservative than $P2$.

$$
\begin{aligned}
Bonf &\prec Holm \prec Hochberg \prec FDR \\
cFDR &\prec FDR \\
Bonf &\prec \gamma FWER,\ unless\ \gamma = 1/p \\
FDR \text{ or } cFDR &\ ? \quad \gamma FWER\ \ (depends\ the\ value\ of\ \gamma)
\end{aligned}
$$

  – Less conservative procedure may potentially increase the power at cost of higher Type I error (false rejection rate).

3 Examples

3.1 The Datasets

- **Colon Data.** 2000 (p) gene expression levels are available on 22 (N_1) normal colon tissues and 40 (N_2) tumor colon tissues. ($N = 62 < p$) [http://microarray.princeton.edu/oncology/affydata/index.html; Alon et al. (1999)]

  Q: Are the tumor genes differentially expressed from the normal genes?

- **Leukemia Data.** 3571 genes expressions are avaible from 47 (N_1) patients suffering from acute lymphoblastic leukemia (ALL) and 25 (N_2) patients suffering from acute myeloid leukemia (AML) ($N = 72 < p$) [(http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi); Dudoit, Fridlyland and Speed (2002), Golub et al. (1999)]

  Q: Are the genes of these two types of cancers differentially expressed?

3.2 Results

Table 3.1 Number of Selected Genes
Which are Considered as Differentially Expressed

Colon Dataset ($p = 2000$)

α	0.1	0.05	0.025
Bonf[1]	70	53	45
Holm[1]	70	55	45
Hoch[1]	70	55	45
FDRp[2]	478	354	257
cFDRp[2]	188	143	95
γ_1FWERp[3]	292	243	193
γ_2FWERp[3]	369	292	243

Leukemia Dataset ($p = 3571$)

α	0.1	0.05	0.025
Bonf[1]	306	279	228
Holm[1]	311	284	233
Hoch[1]	311	284	233
FDRp[2]	1366	1105	912
cFDRp[2]	736	607	511
γ_1FWERp[3]	865	746	644
γ_2FWERp[3]	991	865	746

1: $FWER \leq \alpha$; 2: $FDR \leq \alpha$; 3: γ-$FWER \leq \alpha$; $\gamma_1 = 0.05$, $\gamma_2 = 0.1$.

Table 3.2 Number of Selected Genes
Which are Considered as Differentially Expressed
with Total Number of Genes Having Been Reduced

Colon Dataset

Reduced p	478	354	257	2000
Bon[1]	85	94	101	53
Holm[1]	92	102	139	55
cFDRp[2]	280	326	257	143
γ_1FWERp[3]	243	243	243	243
γ_2FWERp[3]	292	292	257	292

Leukemia Dataset

Reduced p	1366	1105	912	3571
Bon[1]	320	332	341	279
Holm[1]	336	356	387	284
cFDRp[2]	818	874	912	607
γ_1FWERp[3]	745	745	747	746
γ_2FWERp[3]	865	863	864	865

1: $FWER \leq 0.05$;

2: $FDR \leq 0.05$;

3: $\gamma FWER \leq 0.05$; $\gamma_1 = 0.05,\ \gamma_2 = 0.1$.

References

[1] Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. U.S.A.* **96** 6745-6750.

[2] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc.* **B 57** 289-300.

[3] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **4** 1165-1181.

[4] Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tremors using gene expression data. *J. Amer. Statist. Assoc.* **97** 77-87.

[5] Dudoit, S., Shaffer J. P. and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* **18** 71-103.

[6] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286** 531-537.

[7] Hommel, G. and Hoffman, T. (1988). Controlled uncertainty. In *Multiple Hypothesis Testing* (P. Bauer, G. Hommel and E. Sonnemann, eds) 154-161. Springer, Heidelberg.

[8] Lehmann, E. L. and Romano, J. P. (2005). Generalisions of the familywise error rate. *Ann. Statist.* **33** 1138-1154.

[9] Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73** 751-754.