

Testing multiple hypotheses using population information of samples (Supplementary materials)

Mingqi Wu and Faming Liang

Department of Statistics, Texas A&M University, College Station, TX 77843, USA

1 FDR CONTROL

Given the test scores, a multiple hypothesis testing procedure is still needed for identification of significant subjects. Here, we adopted the stochastic approximation-based FDR control method developed by Liang and Zhang (2008), which, hereafter, will be abbreviated as the SA-FDR method. The SA-FDR method falls into the class of empirical Bayes methods (Efron, 2004). Like other methods in this class, it works by fitting the test scores with a two-component mixture model

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z), \quad (1)$$

where π_0 is the prior probability that a null hypothesis is true, f_0 is the empirical null distribution, f_1 is the alternative distribution, and f_0 is stochastically smaller than f_1 . Given the estimators of π_0 and f_0 , the positive FDR (Storey, 2004) of a rejection rule $\Lambda = \{Z_i \geq z_0\}$ can be estimated by

$$\widehat{\text{Fdr}}(\Lambda) = \frac{N \widehat{\pi}_0 [1 - \widehat{F}_0(z_0)]}{\#\{z_i : z_i \geq z_0\}}, \quad (2)$$

where $\#\{z_i : z_i \geq z_0\}$ denotes the number of subjects with test scores greater than z_0 , $\widehat{\pi}_0$ denotes the estimator of π_0 , and \widehat{F}_0 denotes the CDF estimator of f_0 . Note that $\widehat{\text{Fdr}}(\Lambda)$ can be intuitively interpreted as the expected proportion of null subjects, i.e., the subjects with the null hypotheses being true, among those with the test score greater than z_0 . Following the suggestion by Storey (2002), the q -value defined below

$$q(z) \equiv \inf_{\{\Lambda: z \in \Lambda\}} \text{Fdr}(\Lambda), \quad (3)$$

is used in this paper as a reference quantity for the decision of multiple hypothesis testing.

In Liang and Zhang (2008), π_0 and f_0 are estimated using a two-step procedure:

- Fit the distribution of the test scores with a mixture of exponential power distributions using the stochastic approximation method (Robbins and Monro, 1951; Benveniste *et al.*, 1990).
- Clustering the components of the mixture exponential power distributions into two clusters, which correspond to f_0 and f_1 of the mixture (1) respectively, according to the mutual distance between the components.

Liang and Zhang (2008) showed theoretically that the method is valid under general dependence between test scores. We note that

for the population-based test proposed in this paper, the use of the SA-FDR method is not essential. Any other multiple comparison methods, e.g., the methods developed by Benjamini and Yekutieli (2001), Storey *et al.* (2004), and Efron (2004), can be equally used here. To use the method proposed by Benjamini and Yekutieli (2001) and Storey *et al.* (2004), one may need to transform the test scores to p -values via the transformation $P = 1 - \Phi^{-1}(Z)$.

2 COMPARISON: SCORE A VERSUS SCORE C

We compared Score A and Score C with various choices of the parameters $|C_k|$, n , ρ , and μ in terms of specificity and sensitivity of multiple hypothesis tests. The specificity is defined as the proportion of correctly identified non-significant subjects, and the sensitivity is defined as the proportion of correctly identified significant subjects. For the outcome of the multiple tests, let V and U denote, respectively, the numbers of true null hypotheses that are erroneously rejected and correctly accepted, and let T and S denote, respectively, the numbers of false null hypotheses that are erroneously accepted and correctly rejected. Then the specificity and sensitivity can be defined as

$$\text{specificity} = \frac{U}{U + V}, \quad \text{sensitivity} = \frac{S}{T + S}. \quad (4)$$

The specificity and sensitivity working together provide a good measure for the quality of multiple hypothesis tests. The average of sensitivity values over multiple datasets provides a natural estimate for the average power (Dudoit *et al.*, 2003) of the multiple hypothesis test. The average power has been used to assess the quality of multiple hypothesis tests by more and more authors, e.g., Rubin *et al.* (2006) and Storey (2007).

Effect of n In this comparison, we assessed the effect of n , the number of genes included in the dataset, on the effectiveness of score A. For this purpose, we fix $\mu = 3$, $\rho = 0.3$, $|C_k| = 10$, and considered three different pairs of (n, m) : (1250, 125), (2500, 250) and (5000, 250). For each pair of (n, m) , we generated 50 different datasets, calculated scores A and C , and applied the SA-FDR method to each of the datasets. The results were reported in Table 1, which show that Score A outperforms Score C consistently as long as n is reasonably large. It is remarkable that, even with only three-fourths of control and treatment samples, the tests based on Score A still have higher sensitivity (power) than those based on Score C. The tests based on these two types of scores have about the same specificity values.

Effect of ρ In this comparison, we assessed the effect of gene dependency on the effectiveness of score A. For this purpose, we fix

Table 1. Computational results for the data generated with $\mu = 3$, $\rho = 0.3$, $|C_k| = 10$, and different choices of (n, m) . The values of Spec. (specificity), Sens. (sensitivity/power) and their standard deviations (the numbers in the parentheses) were calculated by averaging over 50 datasets. $\Lambda(q)$ denotes a rejection region with the nominal value of FDR being q .

Setting	Score	Measure	$\Lambda(0.2)$	$\Lambda(0.1)$	$\Lambda(0.05)$
n=1250	A	Spec.	.975(.001)	.987(.001)	.993(.001)
		Sens.	.958(.004)	.931(.006)	.889(.019)
m=125	C	Spec.	.975(.002)	.990(.001)	.996(.000)
		Sens.	.885(.008)	.772(.011)	.625(.017)
n=2500	A	Spec.	.978(.001)	.989(.001)	.994(.000)
		Sens.	.950(.003)	.922(.005)	.882(.007)
m=250	C	Spec.	.976(.001)	.990(.001)	.996(.000)
		Sens.	.889(.005)	.774(.009)	.634(.013)
n=5000	A	Spec.	.990(.001)	.995(.000)	.997(.000)
		Sens.	.911(.005)	.867(.006)	.787(.018)
m=250	C	Spec.	.990(.001)	.997(.000)	.999(.000)
		Sens.	.781(.007)	.618(.010)	.445(.013)

$n = 2500$, $m = 250$, $\mu = 3$, and $|C_k| = 10$, and considered two different values of ρ , 0 and 0.6. For each value of ρ , we generated 50 different datasets, calculated scores A and C , and applied the SA-FDR method to each of the datasets. The results were summarized in Table 2. Combining with the results shown in Table 1 for the case $\rho = 0.3$, it can be seen that the performance of Score A is almost independently of the value of ρ . At all different values of ρ , the tests based on Score A have higher powers than those based on Score C.

Table 2. Computational results for the datasets generated with $n = 2500$, $m = 250$, $\mu = 3$, $|C_k| = 10$, and different values of ρ . Refer to Table 1 for the notations used in this table.

Setting	Score	Measure	$\Lambda(0.2)$	$\Lambda(0.1)$	$\Lambda(0.05)$
$\rho = 0.0$	A	Spec.	.978(.001)	.988(.001)	.994(.000)
		Sens.	.949(.002)	.918(.004)	.873(.006)
	C	Spec.	.974(.001)	.990(.001)	.996(.000)
		Sens.	.894(.005)	.775(.009)	.619(.015)
$\rho = 0.6$	A	Spec.	.979(.001)	.990(.001)	.994(.001)
		Sens.	.948(.003)	.913(.005)	.859(.008)
	C	Spec.	.976(.001)	.990(.001)	.996(.000)
		Sens.	.881(.006)	.767(.010)	.612(.015)

Effect of μ In this comparison, we assessed the effect of the expression levels of differentially expressed genes on the effectiveness of score A. For this purpose, we fix $n = 2500$, $m = 250$, $\rho = 0.3$, and $|C_k| = 10$, and considered two different values of

μ , 2 and 4. For each value of μ , we generated 50 different datasets, calculated scores A and C , and applied the SA-FDR method to each of the datasets. The computational results were summarized in Table 3. Combining with the results reported in Table 1 for the case $\mu = 3$, it can be seen that score A outperforms score C irrespective of the value of μ . The improvement is especially significant when the value of μ is small.

Table 3. Computational results for the datasets generated with $n = 2500$, $m = 250$, $\rho = 0.3$, $|C_k| = 10$, and different values of μ . Refer to Table 1 for the notations used in this table.

Setting	Score	Measure	$\Lambda(0.2)$	$\Lambda(0.1)$	$\Lambda(0.05)$
$\mu = 2$	A	Spec.	.972(.002)	.985(.001)	.993(.001)
		Sens.	.738(.012)	.620(.023)	.445(.036)
	C	Spec.	.984(.001)	.995(.001)	.999(.000)
		Sens.	.578(.014)	.359(.016)	.181(.017)
$\mu = 4$	A	Spec.	.982(.001)	.992(.001)	.996(.000)
		Sens.	.993(.001)	.988(.001)	.975(.003)
	C	Spec.	.974(.001)	.989(.001)	.995(.000)
		Sens.	.978(.002)	.931(.004)	.851(.007)

In summary, the population-based test can outperform significantly the two-sample t -test in almost all scenarios. The comparisons also indicate that for this example, to achieve the same or even higher testing power while maintaining the same level of specificity, the population-based test requires less than 3/4 of the control and treatment samples than does the two-sample t -test. This implies that use of the population-based test can potentially lead to a great saving of experiment cost.

REFERENCES

Benjamini, Y. and Yekutieli, D. (2001) On the control of false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**, 1165-1188.

Benveniste, A., Métivier, M., and Priouret, P. (1990) *Adaptive Algorithms and Stochastic Approximations*. Springer-Verlag.

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

Efron, B. (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, **99**, 96-104.

Liang, F. and Zhang, J. (2008) Estimating FDR under general dependence using stochastic approximation. *Biometrika*, **95**(4), 961-977.

Robbins, H. and Monro, S. (1951) A stochastic approximation method. *Annals of Mathematical Statistics*, **22**, 400-407.

Rubin, D., Dudoit, S. and van der Laan, M.J. (2006) A method to increase the power of multiple testing procedures through sample splitting. *Statist. Appl. in Genetics & Mol. Bio.*, **5**, article 19.

Storey, J.D. (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society B*, **64**, 479-498.

Storey, J.D. (2007) The optimal discovery procedure: A new approach to simultaneous significance testing. *J. R. Statist. Soc. B*, **69** 347-368.

Storey, J.D., Taylor, J.E. and Siegmund, D. (2004) Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society B*, **66**, 187-205.