

# A Note on Controlling the Number of False Positives

Edward L. Korn\* and Boris Freidlin

Biometric Research Branch, National Cancer Institute, Bethesda, Maryland 20852-7434, U.S.A.

\*email: korne@ctep.nci.nih.gov

**SUMMARY.** Lehmann and Romano (2005, *Annals of Statistics* **33**, 1138–1154) discuss a Bonferroni-type procedure that bounds the probability that the number of false positives is larger than a specified number. We note that this procedure will have poor power as compared to a multivariate permutation test type procedure when the experimental design accommodates a permutation test. An example is given involving gene expression microarray data of breast cancer tumors.

**KEY WORDS:** Bonferroni; False discovery proportion; Multiple testing; Multivariate permutation test; Permutation test.

## 1. Introduction

In many applications involving high-dimensional data, there may be thousands of null hypotheses being considered (e.g., one for each of 10,000 genes), and there may be interest in identifying which null hypotheses, if any, can be rejected. Controlling the family-wise error rate is typically too stringent a criterion, as it may lead to limited power to detect real differences. In these cases it may be useful to limit the number of false positives or the proportion of false positives (false discovery proportion) with high probability (Korn et al., 2004). (An alternative is to estimate the expected false discovery proportion, which is sometimes known as the false discovery rate [Benjamini and Hochberg, 1995]; see Ge, Dudoit, and Speed [2003] and Li et al. [2005] for extensive reviews of the false discovery rate and Korn et al. [2004] for a discussion of the differences between estimating the false discovery proportion and the false discovery rate.) Lehmann and Romano (2005) recently investigated a simple Bonferroni-type procedure (Hommel and Hoffmann, 1988) to limit the number of false positives, which can then also be used to design procedures to limit the proportion of false positives. The Bonferroni-type procedure is based on univariate p-values associated with the hypotheses that are valid in the sense that

$$P(p_i \leq u) \leq u \quad \text{for all } u \in (0, 1) \quad (1)$$

when the  $i$ th null hypothesis is true. When the null hypotheses of interest can be tested with a permutation test, an alternative approach is based on the reference distribution of an appropriate order statistic of the univariate p-values using the multivariate permutation distribution; see Westfall and Young (1993), Korn et al. (2004), Romano and Wolf (2007), and Section 2 below.

In this note we demonstrate that in applications where a multivariate permutation test (MPT)-type procedure is appropriate, the Bonferroni-type procedure would be expected to have poor power as compared to the MPT-type procedure. We focus on the two-class comparison, a simple situation in which the MPT-type procedure is performed by permuting

the class labels between the two classes. For simplicity of presentation, we consider only single-step procedures in this article; in applications where only a very small proportion of nonnull hypotheses are expected, the gains of using a multi-step procedure are expected to be small. For example, with  $s$  hypotheses, the nominal p-value cut-offs for the Bonferroni procedure ( $\alpha/s$ ) and the Holm (Holm, 1979) procedure ( $\alpha/(s - i + 1)$  for the  $i$ th hypothesis tested) are close when  $s \gg i$ . Also for simplicity, we only consider controlling the number of false positives, as procedures that have poor power for this objective would be expected to have poor power for controlling the false discovery proportion. In particular, a procedure for controlling the number of false positives can be modified to control approximately the false discovery proportion by estimating the false discovery proportion as the number of allowable false positives divided by the number of rejected null hypothesis; see Korn et al. (2007).

## 2. Two-Class Comparison

Let  $x_1, \dots, x_n$  and  $y_1, \dots, y_m$  be  $s$ -dimensional vectors for the observations in class 1 and 2, respectively. The Bonferroni-type procedure uses a univariate p-value (associated with testing the null hypothesis for that variable) for each of the  $s$  variables. For example, the null hypothesis for the  $i$ th variable may be that the two class means are the same, and the p-value may be from a two-sample  $t$ -statistic. To ensure that  $k$  or more false positives occur with  $\leq \alpha$  probability, the Bonferroni-type procedure rejects all hypotheses whose p-values are  $\leq k\alpha/s$ . Letting  $s_0$  be the number of the  $s$  hypotheses that are null, and  $N$  be the number of these  $s_0$  null hypotheses that are rejected, the desired probability constraint follows from (Hommel and Hoffmann, 1988; Lehmann and Romano, 2005)

$$\begin{aligned} P(N \geq k) &= \frac{E(N)}{k} - \frac{\sum_{i=1}^{k-1} iP(N = i) + \sum_{i=1}^{s-k} iP(N = k + i)}{k}, \\ &\leq \frac{E(N)}{k} \leq \frac{\alpha_k^* s_0}{k} \leq \alpha, \end{aligned} \quad (2)$$

where  $\alpha_k^* \equiv k\alpha/s$  and the validity (1) of the p-values is used for the penultimate inequality.

For the MPT procedure to ensure that  $k$  or more false positives occur with  $\leq \alpha$  probability, let  $W_i$  be a univariate statistic associated with the  $i$ th variable such that smaller values of  $W_i$  suggest the  $i$ th null hypothesis is not true,  $i = 1, \dots, s$ . (For the two-class comparison, we assume that the  $s_0$ -dimensional multivariate distribution of the variables associated with null hypotheses is the same regardless of class label.) For example,  $W_i$  could be a p-value associated with a hypothesis test, but need not be. Consider constructing permuted datasets by permuting the class labels, and, for each permuted dataset, calculating the  $W$ 's on all the variables, say,  $W_1^*, \dots, W_s^*$ . Let the ordered  $W^*$ 's be denoted  $W_{(1)}^* \leq \dots \leq W_{(s)}^*$ , and let

$$c_{\alpha,k} = \text{MAX}\{c \mid P(W_{(k)}^* < c \mid \{x_1, \dots, x_n, y_1, \dots, y_m\}) \leq \alpha\}, \tag{3}$$

where  $P(\bullet \mid \{x_1, \dots, x_n, y_1, \dots, y_m\})$  refers to the probability under the permutation distribution, and the MAX is the maximum over  $c$ . (The quantity  $c_{\alpha,k}$  is essentially the  $\alpha$ th quantile of  $W_{(k)}^*$ .) The MPT-based procedure rejects all null hypotheses associated with variables  $i$  such that  $W_i < c_{\alpha,k}$ .

To see why this procedure satisfies the probability error constraints, let  $I \subseteq \{1, \dots, s\}$  be the set of indices corresponding to true null hypotheses, let  $W_{(1)}^0 \leq \dots \leq W_{(s_0)}^0$  be the ordered  $W$  statistics on the original (unpermuted) dataset restricted to  $i \in I$ , let  $W_{(1)}^{*0} \leq \dots \leq W_{(s_0)}^{*0}$  be the ordered  $W$  statistics on a permuted dataset restricted to  $i \in I$ , and let

$$c_{\alpha,k}^0 = \text{MAX}\{c \mid P(W_{(k)}^{*0} < c \mid \{x_1, \dots, x_n, y_1, \dots, y_m\}) \leq \alpha\}.$$

Note that although  $c_{\alpha,k}^0$  is unknown to us, we do know that  $c_{\alpha,k}^0 \geq c_{\alpha,k}$  because  $W_{(k)}^{*0} \geq W_{(k)}^*$  (the  $\{W^{*0}\}$ 's being a subset of the  $\{W^*\}$ 's). The proof that the probability that  $k$  or more null hypothesis are rejected is  $\leq \alpha$  is as follows:

$$\begin{aligned} &P(k \text{ or more null hypotheses rejected}) \\ &= P(W_{(k)}^0 < c_{\alpha,k}) \\ &= E\left[P(W_{(k)}^0 < c_{\alpha,k} \mid \{X_1, \dots, X_n, Y_1, \dots, Y_m\})\right] \\ &\leq E\left[P(W_{(k)}^0 < c_{\alpha,k}^0 \mid \{X_1, \dots, X_n, Y_1, \dots, Y_m\})\right] \\ &\leq E(\alpha) = \alpha. \end{aligned}$$

For our comparison of the Bonferroni-type procedure with the MPT-based procedure, we will assume that a p-value from a two-sample  $t$ -test is used for each variable for the Bonferroni-type procedure, say  $p_i$ , and these same p-values as the test statistics for the MPT-based procedure, i.e.,  $W_i = p_i$ ,  $i = 1, \dots, s$ . First, consider the case when  $n, m \rightarrow \infty$  (asymptotic sample size case) and the global null hypothesis that  $\{X_1, \dots, X_n, Y_1, \dots, Y_m\}$  independent and identically distributed from an  $s$ -dimensional continuous distribution that has correlation matrix  $R$  and whose components have finite absolute third moments. Let  $F_{n,m}(u_1, \dots, u_s \mid \{X_1, \dots, X_n, Y_1, \dots, Y_m\})$  be the cumulative distribution function of the permutation distribution of the vector of  $t$ -statistics. Following Hoeffding (1952),  $F_{n,m}(u_1, \dots, u_s \mid \{X_1, \dots, X_n, Y_1, \dots, Y_m\})$  converges in probability to the cumulative distribution function of a multivariate normal distribution with

mean 0 and covariance (and correlation) matrix  $R$ . The following theorem, whose proof is given in the Appendix, shows that (i) we can use a multivariate normal distribution to evaluate the asymptotic properties of the MPT-based procedure and that (ii) the asymptotic p-value cut-off for rejection for the MPT-based procedure is greater than or equal Bonferroni-type procedure.

**THEOREM 1:** *Let  $\{X_1, \dots, X_n, Y_1, \dots, Y_m\}$  be independent and identically distributed from an  $s$ -dimensional continuous distribution that has correlation matrix  $R$  and whose components have finite absolute third moments. Let  $c_{\alpha,k}^{(n,m),\{X_1, \dots, X_n, Y_1, \dots, Y_m\}}$  be the  $c_{\alpha,k}$  defined by (3) with the explicit notation for the sample sizes and conditioning order statistics.*

$$(i) \lim_{n,m \rightarrow \infty} c_{\alpha,k}^{(n,m),\{X_1, \dots, X_n, Y_1, \dots, Y_m\}} \xrightarrow{P} c_{\alpha,k}^{(\infty)} \equiv 2(1 - \Phi(\gamma_{\alpha,k})),$$

where  $\gamma_{\alpha,k}$  is the  $1 - \alpha$ th quantile of the  $k$ th largest of  $|Z_1|, \dots, |Z_s|$ , and where  $Z = (Z_1, \dots, Z_s)$  has a multivariate normal distribution with mean 0 and covariance matrix  $R$ , and  $\Phi(\bullet)$  is a standard normal cumulative distribution function.

$$(ii) c_{\alpha,k}^\infty \geq \frac{k\alpha}{s}.$$

To see how much better the power is for the MPT-based procedure as compared to the Bonferroni-type procedure, we conducted a simulation that would be relevant to microarray experiments: 10,000 variables with a block diagonal correlation structure with block size of 100 and correlation  $\rho$  within the blocks. For the asymptotic sample size case, Table 1 displays the cut-offs required for the two procedures ( $\alpha = 0.05$ ) to reject the null hypothesis for any given variable in terms of the nominal p-value for that variable. For the Bonferroni-type

**Table 1**

*Rejection cut-offs ( $\times 10^{-5}$ ) for individual variables (in terms of nominal p-values) for two-class comparison (two-sided level = 0.05) for 10,000 variables (block correlation structure, 100 blocks with within-block correlation of  $\rho$ ) allowing for 0–10 errors using a Bonferroni-type procedure and MPT-based procedure (asymptotic results). For the MPT-based procedure, the cut-offs were obtained by simulating  $10^7$  multivariate normal vectors.*

Number of allowable errors ( $k - 1$ )	Bonferroni-type procedure	MPT-based procedure ( $c_{0.05,k}^\infty$ )	
		$\rho = 0$	$\rho = 0.5$
0	0.5	0.51	0.61
1	1.0	3.53	2.90
2	1.5	8.14	5.86
3	2.0	13.61	9.18
4	2.5	19.70	12.72
5	3.0	26.10	16.45
6	3.5	32.76	20.34
7	4.0	39.71	24.36
8	4.5	46.99	28.53
9	5.0	54.28	32.77
10	5.5	61.73	37.13

**Table 2**

Power to reject the null hypothesis (level = 0.05) for a variable in the two-class comparison for which  $E(\bar{X}_1 - \bar{X}_2) = 5 \times SD(\bar{X}_1 - \bar{X}_2)$  when there are 10,000 variables (block correlation structure, 100 blocks with within-block correlation of  $\rho$ ) allowing for 0–10 errors using a Bonferroni-type procedure and MPT-based procedure (asymptotic results).

Number of allowable errors ( $k - 1$ )	Bonferroni-type procedure (%)	MPT-based procedure	
		$\rho = 0$ (%)	$\rho = 0.5$ (%)
0	67	67	68
1	72	81	79
2	75	86	84
3	77	88	86
4	78	90	88
5	80	91	89
6	81	92	90
7	81	93	91
8	82	93	91
9	83	94	92
10	83	94	93

procedure, this cut-off is  $0.05k/10,000$ , where the procedure ensures that there are  $\geq k$  true null hypotheses rejected with at most  $\alpha$  probability. For the MPT-based procedure, the cut-offs ( $c_{0.05,k}^\infty$ ) were determined by simulating multivariate normal distributions, although they could have been obtained for the  $\rho = 0$  case by solving  $0.95 = \sum_{j=0}^{k-1} \binom{10,000}{j} (1-c)^{10,000-j} c^j$  for  $c$ . Table 2 shows the power to reject a null hypothesis for a variable whose true mean difference is five standard errors, i.e., the power equals  $1 - \Phi(\Phi^{-1}(1 - c_{0.05,k}^\infty/2) - 5)$ . (If one knew a priori that the correlation  $\rho$  were 0, then one could improve on the Bonferroni-type procedure [Guo and Romano, 2007; Sarkar, 2007].)

Tables 1 and 2 demonstrate that the Bonferroni procedure has reasonable power as compared to the MPT-based procedure allowing for 0 errors ( $k = 1$ ), but the Bonferroni-type procedure has a large loss of power when allowing for more than 0 errors ( $k \geq 1$ ). (Of course, the Bonferroni procedure for 0 errors will have a substantial loss of power if the correlation is unrealistically high. For example, if  $\rho = 0.9$ , the power in Table 2 for  $k = 1$  would be 81% for the MPT-based procedure [not shown] as compared to 67% for the Bonferroni procedure.) A heuristic explanation of why the Bonferroni-type procedure works well (for reasonable correlation) for  $k = 1$  and poorly for  $k > 1$  is as follows for  $k = 1$  versus  $k = 2$ . For simplicity, we assume (i) the global null hypothesis (i.e., all null hypotheses are true), (ii) complete independence of the variables ( $\rho = 0$ ), and (iii) the p-values have uniform distributions (i.e., a stronger assumption than (1)). From (2), we have for  $k = 1$

$$\begin{aligned}
 P(N \geq 1) &= \alpha_1^* s - \sum_{i=1}^{s-1} iP(N = 1 + i) \\
 &= \alpha - \sum_{i=1}^{s-1} iP(N = 1 + i),
 \end{aligned}$$

and for  $k = 2$

$$\begin{aligned}
 P(N \geq 2) &= \frac{\alpha_2^* s}{2} - \frac{P(N = 1) + \sum_{i=1}^{s-2} iP(N = 2 + i)}{2} \\
 &= \alpha - \frac{P(N = 1) + \sum_{i=1}^{s-2} iP(N = 2 + i)}{2}.
 \end{aligned}$$

The difference between  $\alpha$  and the type 1 error of the Bonferroni procedure ( $k = 1$ ) is approximately

$$P(N = 2) = \frac{s(s-1)}{2} (\alpha_1^*)^2 (1 - \alpha_1^*)^{s-2} \cong \alpha^2/2,$$

whereas for the Bonferroni-type procedure with  $k = 2$  it is approximately

$$\frac{1}{2}P(N = 1) = \frac{1}{2}s\alpha_2^*(1 - \alpha_2^*)^{s-1} \cong \alpha - 2\alpha^2.$$

Compared to the target level of  $\alpha$ , we see that the conservativeness of the Bonferroni procedure ( $k = 1$ ) is small ( $\alpha^2/2$ ), but the conservativeness of the Bonferroni-type procedure ( $k = 2$ ) is large ( $\alpha - 2\alpha^2$ ). This conservativeness of the level under the global null hypothesis translates into reduced power when there are nonnull hypotheses.

As the results in Tables 1 and 2 are asymptotic, we also consider a small-sample situation with sample sizes  $n = 10$  and  $m = 5$  in the two groups. A block correlation structure with  $\rho = 0.5$  is again used, with 190 of the normally distributed 10,000 variables nonnull with a true mean difference of 7.9 standard errors (a difference of 7.9 standard errors corresponds to approximately 65% power for a Bonferroni-adjusted  $t$ -test with 13 degrees of freedom); 10 of the blocks have 10 nonnull variables each, and the other 90 blocks each have one nonnull variable. Table 3 displays power to reject the null hypothesis for a nonnull variable and the quantiles of the distribution of the number of the 190 nonnull hypotheses rejected by the Bonferroni-type and MPT-based procedures. Allowing for zero errors ( $k = 1$ ), the Bonferroni and MPT-based procedures have similar characteristics. When more than zero errors are allowed, the Bonferroni-type procedure results in a substantial loss of power relative to the MPT-based procedure. This loss of power seen in Table 3 is even more dramatic than the asymptotic results given in Table 2, e.g., 75% versus 88% for  $k = 2$  (Table 3) and 72% versus 79% for  $k = 2$  (Table 2). The advantage of the MPT-based procedure is also seen by the marked increase in the number of the rejected nonnull hypothesis, e.g., median 167 versus 142 for  $k = 2$ . Small-sample properties of the methods are discussed additionally below.

### 3. An Example

Sotiriou et al. (2003) analyzed cDNA gene expression profiles from 99 tumor specimens from breast cancer patients. In addition to gene expression values for 7650 genes (probes) preprocessed as described in Sotiriou et al. (2003), there was standard prognostic variable information available for each patient. (The data are publicly available at <http://linus.nci.nih.gov/~brb/DataArchive.html>.) Here we consider

**Table 3**

*Simulated distributional properties of the number of nonnull hypotheses rejected ( $\alpha = 0.05$ ) in a two class comparison ( $n = 10$ ,  $m = 5$ ) when there are 10,000 variables (block correlation structure, 100 blocks with within-block correlation of  $\rho = 0.5$ ) with 190 nonnull variables (see text) allowing for 0–6 errors using a Bonferroni-type procedure and MPT-based procedure (2000 simulated datasets)*

Number of allowable errors ( $k - 1$ )	Bonferroni-type procedure				MPT-based procedure			
	Power <sup>a</sup>	First quartile	Median	Third quartile	Power <sup>a</sup>	First quartile	Median	Third quartile
0	0.655	117	123	129	0.665	118	125	131
1	0.747	137	142	148	0.879	163	167	171
2	0.800	147	152	157	0.932	175	178	181
3	0.839	153	158	163	0.959	180	183	185
4	0.858	158	163	167	0.970	183	185	187
5	0.875	162	166	170	0.977	185	186	188
6	0.886	165	169	173	0.983	186	187	188

<sup>a</sup>Power to reject the null hypothesis for a nonnull variable.

two two-class comparisons based on parametric two-sample  $t$ -tests and control for the number of false positive at the  $\alpha \leq 0.05$  level. For each comparison, we restrict attention to genes for which the number of missing values was less than the number of specimens in the class with fewer observations minus 2.

The first comparison is for patients with grade 1 or 2 tumors ( $n = 54$ ) versus patients with grade 3 tumors ( $m = 45$ ) with  $s = 7498$  genes. Allowing for no errors ( $k = 1$ ), Bonferroni identifies three genes and the MPT-based procedure identifies six genes. However, allowing for 10 errors ( $k = 11$ ), the Bonferroni-type procedure identifies 28 genes and the MPT-based method identifies 94 genes. An interesting gene found by the MPT-based procedure and not by the Bonferroni-type procedure is BUB1, whose gene expression has been previously associated with survival in breast cancer patients (Glinsky, Berezovska, and Glinskii, 2005).

The second comparison is for patients with estrogen receptor negative status ( $n = 34$ ) versus patients with estrogen receptor positive status ( $m = 65$ ) with  $s = 7470$  genes. Allowing for no errors ( $k = 1$ ), Bonferroni identifies 163 genes and the MPT-based procedure identifies 172 genes. However, allowing for 10 errors ( $k = 11$ ), the Bonferroni-type procedure identifies 290 genes and the MPT-based method identifies 503 genes. Two interesting genes found in the 503 gene set but not in the 290 gene set are TSG101 and BAP1. TSG101 represses transcriptional activation by estrogen receptor (Sun et al., 1999), and BAP1 may be a breast cancer tumor suppressor gene (Jensen et al., 1998).

#### 4. Discussion

In addition to the lower power as compared to the MPT-based procedure, a severe limitation of the Bonferroni-type method is that it requires a valid univariate p-value for each hypothesis test in the sense of (1). In particular, for p-values derived from parametric tests this is a challenge for very small  $u$  (Ringwald, 1983). For example, a small amount of

nonnormality in the  $x$  and  $y$  data will lead to a violation of (1) for very small  $u$  for a  $t$ -test derived p-value unless the sample sizes are extremely large. In particular, if a normal critical value corresponding to a  $10^{-5}$  two-sided p-value was incorrectly used instead of a  $t$ -distribution with 100 degrees of freedom, then the actual rejection probability would be  $2.5 \times 10^{-5}$ , far from the nominal level. An approach to avoid this problem is to use p-values from univariate permutation tests, e.g., the Wilcoxon-rank sum test for the two-class comparison. Although these p-values will satisfy (1), they unfortunately lead to a Bonferroni-type procedure with very little power unless the sample sizes are moderately large. For example, for a two-class comparison with  $n = 10$  and  $m = 5$ , the smallest (two-sided) p-value obtainable from a rank test is  $6.66 \times 10^{-4}$ , leading to no possible rejection for reasonably sized  $k$  and  $s = 10,000$  variables.

The MPT-based procedure does not require (1) to hold for the procedure to be valid. In fact, any univariate statistic can be used, not necessarily a p-value. The choice of the statistic can obviously affect the power of the procedure, so we generally recommend using p-values from parametric tests. The MPT-based procedure is restricted to experimental designs that accommodate a permutation test. Fortunately, this covers many practical applications: paired or unpaired two-class comparisons,  $k$ -class comparisons, simple linear regression, simple logistic regression, and survival analysis with one independent variable. For situations in which a permutation test cannot be directly used, e.g., testing one independent variable in a multiple linear regression, one might attempt to obtain the required reference distribution by bootstrapping the data vectors rather than permuting them under a null model. Unfortunately, the properties of bootstraps with high-dimensional data are unpredictable unless the sample sizes are extremely large (Troendle, Korn, and Mcshane, 2004). We believe a more promising approach is to use an approximate permutation test based on permuting the residuals; this is an area of further research.

ACKNOWLEDGEMENTS

The authors wish to thank Dr Lara Lusa (Department of Experimental Oncology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy) for her help in interpreting the results of the example given in Section 3.

REFERENCES

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289–300.

Ge, Y., Dudoit, S., and Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis (with discussion). *TEST* **12**, 1–77.

Glinsky, G. V., Berezovska, O., and Glinskii, A. B. (2005). Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. *Journal of Clinical Investigation* **115**, 1503–1521.

Guo, W. and Romano, J. (2007). A generalized Sidak-Holm procedure and control of generalized error rates under independence. *Statistical Applications in Genetics and Molecular Biology* **6**, 1–33.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *Annals of Mathematical Statistics* **23**, 169–192.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70.

Hommel, G. and Hoffmann, T. (1988). Controlled uncertainty. In *Multiple Hypotheses Testing*, P. Bauer, G. Hommel, and E. Sonnendmann (eds), pp. 154–161. Heidelberg: Springer.

Jensen, D. E., Proctor, M., Marquis, S. T., Gardner, H. P., Ha, S. I., Chodosh, L. A., Ishov, A. M., Tommerup, N., Vissing, H., Sekido, Y. et al. (1998). BAP1: A novel ubiquitin hydrolase which binds to the BRCA1 RING finger and enhances BRCA1-mediated cell growth suppression. *Oncogene* **16**, 1097–1112.

Korn, E. L., Troendle, J. F., McShane, L. M., and Simon, R. (2004). Controlling the number of false discoveries: Application to high-dimensional genomic data. *Journal of Statistical Planning and Inference* **124**, 379–398.

Korn, E. L., Li, M.-C., McShane, L. M., and Simon, R. (2007). An investigation of two multivariate permutation methods for controlling the false discovery proportion. *Statistics in Medicine* **26** (in press).

Lehmann, E. L. and Romano, J. P. (2005). Generalizations of the familywise error rate. *Annals of Statistics* **33**, 1138–1154.

Li, S. S., Bigler, J., Lampe, J. W., Potter, J. D., and Feng, Z. (2005). FDR-controlling testing procedures and sample size determination for microarrays. *Statistics in Medicine* **24**, 2267–2280.

Ringwald, J. T. (1983). Robust multiple comparisons. *Journal of the American Statistical Association* **78**, 145–151.

Romano, J. P. and Wolf, M. (2007). Control of generalized error rates in multiple testing. *Annals of Statistics* (in press).

Sarkar, S. K. (2007). Stepup procedures controlling generalized FWER and generalized FDR. *Annals of Statistics*, in press.

Sotirou, C., Neo, S.-Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., Martiat, P., Fox, S. B., Harris, A. L., and Liu, E. T. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *PNAS* **100**, 10393–10398.

Sun, Z., Pan, J., Hope, W. X., Cohen, S. N., and Balk, S. P. (1999). Tumor susceptibility gene 101 protein represses androgen receptor transactivation and interacts with p300. *Cancer* **86**, 689–696.

Troendle, J. F., Korn, E. L., and McShane, L. M. (2004). An example of the slow convergence of the bootstrap in high dimensions. *American Statistician* **58**, 25–29.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing*. New York: Wiley.

Received December 2006. Revised May 2007.  
Accepted May 2007.

APPENDIX

Proof of the Theorem 1

- (i) The cumulative distribution function of an order statistic can be written in terms of a linear combination of the associated multivariate cumulative distribution function. Therefore, in particular,  $c_{\alpha,k,\{X_1,\dots,X_n,Y_1,\dots,Y_m\}}^{(n,m)}$  is the maximum (over  $c$ ) for which  $1 - \alpha$  is less than a linear combination of  $F_{n,m}(u_1, \dots, u_s | \{X_1, \dots, X_n, Y_1, \dots, Y_m\})$  for various values of  $(u_1, \dots, u_s)$  that are each equal to  $t_{n+m-2}^{-1}(1 - c/2)$  or  $\infty$ , where  $t_{n+m-2}^{-1}(\bullet)$  is the inverse cumulative distribution function of a  $t$  distribution with  $n + m - 2$  degrees of freedom. For example, for  $k = 1$ ,

$$c_{\alpha,1,\{X_1,\dots,X_n,Y_1,\dots,Y_m\}}^{(n,m)} = \text{MAX}\{c | F_{n,m}(t_{n+m-2}^{-1}(1 - c/2), \dots, t_{n+m-2}^{-1}(1 - c/2) | \{X_1, \dots, X_n, Y_1, \dots, Y_m\}) > 1 - \alpha\}.$$

The result follows by taking the limit as  $n, m \rightarrow \infty$ .

- (ii) Let  $(Z_1, \dots, Z_s)$  have a multivariate normal distribution with mean 0 and covariance matrix  $R$ , and let  $p_{(k)}$  be the  $k$ th smallest value of  $2[1 - \Phi(Z_i)]$ ,  $i = 1, \dots, s$ . Then, restating (2),  $P(p_{(k)} < \frac{k\alpha}{s}) \leq \alpha$  (for any  $n$  and  $m$ ). Taking the limit of this quantity, one has  $\lim_{n,m \rightarrow \infty} P(p_{(k)} < \frac{k\alpha}{s}) = d \leq \alpha$ . Because the distribution of the  $p_{(k)}$  is continuous, the result follows by contradiction by noting  $\lim_{n,m \rightarrow \infty} P(p_{(k)} < c_{\alpha,k}^\infty) = \alpha$ .