Recent Developments in Multiple Comparison Procedures
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 47 (2004) 143–154

Weighted FWE-controlling methods in high-dimensional situations

Peter H. Westfall[1], **Siegfried Kropf**[2], **and Livio Finos**[3]

Texas Tech University, Otto von Guericke University and Padova University

Abstract: With high dimensionality, standard Bonferroni-style procedures can suffer from loss of power, since the significance level α must be divided by k to declare significance. Kropf and Läuter (KL) show that certain data-dependent quadratic forms can be used to "pre-specify" hypotheses, which can then be tested in a fixed, data-dependent order, without multiplicity adjustment. In this article we extend the KL procedure to a class of weighted procedures, using the same quadratic forms. The class includes the KL method, the Bonferroni-Holm method, and other, new procedures. We establish strong FWE control for all procedures, and compare power and level of various weighting methods using analytical and simulation results. The method is applied using a high-dimensional mixture model that is suggested by the analysis of real gene expression data.

1. Introduction

With the genomics revolution, methods for detecting signals in large data sets are increasingly in demand. In studies relating single nucleotide polymorphisms (SNPs) to disease status, there can easily be thousands of genotypes to be tested; the number of tests is in the millions and even billions when interacting genotype effects are considered. With so many tests, the standard Type I error rate criterion upon which tests are evaluated becomes less meaningful, as hundreds of "significances" might easily be, in effect, Type I errors. On the other hand, attempts to rigorously control the familywise Type I error rate (FWE) typically are excessively conservative. For example, if the Bonferroni method is used to control the FWE with k tests, then a test will have to be significant at the α/k level to be declared "real." The two most often voiced complaints about this method are (i) the dependence of the procedure on k seems arbitrary, and related to that (ii) the method is exceedingly conservative for large k.

Holm's (1979) step-down approach is a slight improvement on the simple Bonferroni method, but gains little in terms of power when k is large. Holm's method rejects the hypothesis corresponding to the most significant (smallest p-value) test if the p-value is less than α/k; if this hypothesis is rejected, then the second-smallest p-value is compared to $\alpha/(k-1)$, and so on.

The false discovery rate (FDR) controlling method of Benjamini and Hochberg (1995; hereafter the BH method) has been proposed as alternative to FWE-controlling methods that, to an extent, meliorates problems (i) and (ii). FDR-controlling

[1] Area of Information Systems and Quantitative Sciences, Texas Tech University, Lubbock, TX 79409, USA. e-mail: `westfall@ba.ttu.edu`

[2] Otto-von-Guericke-Universität Magdeburg, Institut für Biometrie und Medizinische Informatik, Leipziger Str. 44, D-39120 Magdeburg, Germany. e-mail: `Siegfried.Kropf@medizin.uni-magdeburg.de`

[3] Department of Statistical Science, Padova University, Via Cesare Battisti 241/243, 35121 Padova, Italia. e-mail: `lfinos@stat.unipd.it`

Keywords and phrases: a-priori ordered hypotheses, gene expression, multiple comparisons.

AMS 2000 subject classifications: primary 62F03, 62F07; secondary 62C25, 62E17.

procedures do not generally control the FWE, thus they allow that some fraction of detected significances are in error. While FDR-controlling procedures have more power and attractive optimality properties, significances obtained using FWE-controlling procedures remain desirable since the inferences are stronger. In this article we consider a class of weighted FWE-controlling procedures.

Weighted methods are useful when some hypotheses H_i are deemed more important than others. For example, in clinical trials, the various patient outcomes might be ranked a priori, and the testing procedure designed to give more power to the more important hypotheses. The simplest weighting multiple testing procedure, discussed in e.g., Rosenthal and Rubin (1983), is to reject H_i if $p_i \leq w_i\alpha$, where the weights w_i lie in the simplex $w_i \geq 0$; $\Sigma w_i = 1$, and where p_i is the p-value of the test. The w_i may be chosen based purely on a priori importance of the hypotheses, or to optimize power based on prior information (Spjøtvoll, 1972; Benjamini and Hochberg, 1997; Westfall et al., 1998).

Using concurrent data sets to generate and test hypotheses is generally considered "data snooping," and such methods typically inflate Type I error rates. However, when properly chosen, weights can be taken from the concurrent data set so as to improve power without compromising significance levels. For example, in the two-sample Fisher exact test of binomial proportions, the marginal totals contain no information about the significance level; thus, "unusually large" marginal totals may be used to pre-select particular tests (Louis and Bailey, 1990; Westfall and Soper, 2001). In the parametric setting, Läuter, Glimm, and Kropf (1996) noted that linear combination weight vectors may depend on the data through certain quadratic forms, and the resulting (ordinary) t-tests retain their levels. Thus, by choosing the weight vectors suitably, the procedures may be used to weight the procedure in favor of particular hypotheses selected by the data.

In either of the cases mentioned above, determination of "unusually large" targeting functions requires either historical data or an assumption of marginal homogeneity. For example, in the binomial case, if there is historical data suggesting that the rate should be π_{0i}, then H_i might be weighted using a measure of discrepancy of the marginal total from the *a priori* expected total (Westfall and Soper, 2001). On the other hand, if historical information is unavailable but the binomial proportions can be assumed to be reasonably homogeneous across tests, then the targeting function may be taken as the marginal total itself. This latter approach is essentially taken by Kropf and Läuter (2002; hereafter KL), but in the case of normally distributed data, where the targeting functions are certain quadratic forms.

In this paper we consider a class of weighted FWE-controlling testing procedures that utilize the KL quadratic forms. As in Westfall and Krishen (2001, hereafter WK), we let the weight be indexed by a parameter η, $0 \leq \eta \leq \infty$, where $\eta = 0$ corresponds to the Holm procedure, $\eta = \infty$ corresponds to the fixed sequence KL procedure, and intermediate values of η refer to new procedures. Strong FWE control is proven for all members of the class, power comparisons among various members of the class are considered, and recommendations are offered.

Popular FWE-controlling alternatives to the methods considered here include non-parametric and semi-parametric resampling methods, including bootstrap and permutation methods. The Westfall-Young "method" (Westfall and Young, 1993, actually a collection of methods) involves resampling data under the complete null hypothesis and computing step-wise p-value adjustments that incorporate distributional and correlation characteristics. These methods are reviewed and extended in recent publications with particular emphasis on gene expression data; see Dudoit et al. (2003), Ge et al. (2003), and Troendle et al. (2003). While it is useful to consider how different methods work in different situations, the resampling methods are of somewhat tangential interest as regards the current paper, as we are concerned primarily with a class of exact parametric tests. Nevertheless, some analyses using resampling methods are given in this paper for brief comparison. For discussions of the problem form a nonparametric standpoint, see Kropf et al. (2004).

2. A class of weighted testing procedures

Holm (1979) presented the following weighted testing procedure: Order the weighted p-values $q_i = p_i/w_i$ as $q_{(1)} \leq \cdots \leq q_{(k)}$, where $q_{(j)} = q_{i_j}$; i.e., i_j denotes the index of the jth ordered weighted p-value. Define the sets $\mathbf{S}_j = \{i_j, \ldots, i_k\}$, $j = 1, \ldots, k$. Letting the hypothesis corresponding to $q_{(j)}$ be denoted $H^w_{(j)}$, the method rejects $H^w_{(j)}$ if $q_{(i)} \leq \alpha/\Sigma_{h \in \mathbf{S}_i} w_h$, for all $i = 1, \ldots, j$. When the weights are equal, the method reduces to the ordinary step-down Holm method.

Consider a sequence of positive weight vectors $\mathbf{w}(\eta) = (w_1(\eta), \ldots, w_k(\eta))$, where $\eta \to \infty$, satisfying $w_{i+1}(\eta)/w_i(\eta) \to 0$, for $i = 1, \ldots, k-1$. Westfall and Krishen proved that the critical function of the weighted Holm method converges almost surely to that of the fixed sequence procedure wherein hypotheses are tested in the fixed sequence $H_1, H_2, \ldots$, stopping as soon as an insignificant result is obtained (see, e.g., Maurer et al. 1995).

Suppose now that there exist absolutely continuous data-dependent positive random variables $g_i = g_i(\mathbf{x})$ where $\mathbf{x}$ denotes the data (later we shall define the g_i as KL quadratic forms). For almost all $\mathbf{x}$, we have strict equality of the order statistics, $g_{(1)} > \cdots > g_{(k)}$. Letting $w_i(\eta) = g_i^\eta$, we have the condition $w_{(i+1)}(\eta)/w_{(i)}(\eta) \to 0$, for $i = 1, \ldots, k-1$, for almost all $\mathbf{x}$. Thus, applying Theorem 4 of WK, the weighted step-down Holm method converges almost surely to the fixed-sequence procedure, with the ordering of the hypotheses determined by the order of the data-dependent g_i.

Thus, we have a class of data-dependent multiple testing procedures, indexed by η, for which $\eta = 0$ implies the ordinary step-down Holm procedure, and for which $\eta = \infty$ gives the KL procedure. Strong FWE control for the KL procedure is given by KL; while strong FWE control for all members in the class is shown (for suitable g_i) in the following section. It is also worth noting that the FWE and power functions for the weighted procedures converge to those of the KL procedure; this is a consequence of the Lebesgue dominated convergence theorem as can be shown using the method of Theorem 3 of WK.

3. Assumptions and FWE control

3.1. The multivariate single-sample case

Assume that we have a sample of size n from a k-dimensional normal population

$$\mathbf{x}_j = \begin{pmatrix} x_{j1} \\ \vdots \\ x_{jk} \end{pmatrix} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad j = 1, \ldots, n,$$

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \ldots & \sigma_{1k} \\ \vdots & \ddots & \vdots \\ \sigma_{k1} & \ldots & \sigma_{kk} \end{pmatrix}$$

and that we want to test the local null hypotheses $H_i : \mu_i = 0$ $(i = 1, \ldots, k)$. Let $\overline{x}_i = \sum_{j=1}^n x_{ji}/n$, $s_i^2 = \sum_{j=1}^n (x_{ji} - \overline{x}_i)^2/(n-1)$; giving the usual one-sample t test statistic $t_i = \overline{x}_i/(s_i/\sqrt{n})$ and the two-sided p-values $p_i = 2(1 - F_{t(n-1)}(\mid t_i \mid))$. The hypotheses will be targeted using weights $w_i(\eta) = g_i^{\eta}$, where $g_i = \sum_{j=1}^{n} x_{ji}^2$ $(i = 1, \ldots, k)$.

Theorem 1. *Holm's weighted testing procedure described in Section* 2*, with data-dependent weights* $w_i(\eta)$*, strongly controls the FWE.*

Proof. For convenience, let $w_i = w_i(\eta)$.

Let M_0 be the fixed but unknown "null set" of variables x_i, for which $\mu_i = 0$, with size k_0. Theorem 2 of WK states that the weighted step-down procedure is simply a closed testing procedure using $\min p_i/w_i$ as a test statistic at each node. Thus, we can apply the main theorem of Marcus et al. (1976; stated as Theorem 1 of WK) to prove FWE control, provided that $\min_{i:x_i \in M_0} p_i/w_i < \alpha/\sum_{i=1}^{k} w_i$ occurs with probability α at most. If M_0 is empty ($k_0 = 0$) then no Type I error can occur, so we consider $k_0 > 0$.

We denote by $\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix}$ the matrix of k_0-dimensional subvectors $\mathbf{y}_j$ from the sample vectors $\mathbf{x}_j$ $(j = 1, \ldots, n)$, consisting of those variables only that belong to M_0. Then the row vectors of $\mathbf{Y}$ are *iid* $\mathbf{y}_j \sim N_{k_0}(\mathbf{0}, \mathbf{\Sigma}_0)$ with some positive semidefinite matrix $\mathbf{\Sigma}_0$, or in matrix notation $\mathbf{Y} \sim N_{n \times k_0}(\mathbf{0}_{n \times k_0}, \mathbf{I}_n \otimes \mathbf{\Sigma}_0)$. Hence, $\mathbf{Y}$ is left-spherically distributed (see Fang and Zhang, 1990, e.g.).

We now consider the distribution of $\mathbf{Y}$ and that of the derived test statistics for fixed $\mathbf{W}_0 = \mathbf{Y}'\mathbf{Y}$. According to the properties of left-spherically-distributed matrices, this is again a left-spherical matrix distribution. Also all columns (corresponding to the single variables in M_0) are left-spherically distributed. According to Fang and Zhang (1990), Theorem 5.1.1, a test statistic $t(\mathbf{X})$ is distribution-free on the class of left-spherically distributed matrices of size $n \times p$ (here $n \times 1$), if $t(\mathbf{XA}) = t(\mathbf{X})$ for each constant upper triangular $p \times p$-matrix $\mathbf{A}$ with positive diagonal elements. As the t tests t_i are invariant to scale changes, their conditional distributions for fixed $\mathbf{W}_0$ are the same as with *iid* standard normal variables, such that they exactly maintain the Type I error. For fixed $\mathbf{W}_0 = \mathbf{Y}'\mathbf{Y}$, the weights w_i are also fixed for the variables in M_0 because the w_i are functions of the diagonal elements of $\mathbf{W}_0$. Therefore, for each variable in M_0 we have

$$P\left(\frac{p_i}{w_i} \leq \frac{\alpha}{\sum_{l:x_l \in M_0} w_l}\right) = P\left(p_i \leq \frac{\alpha w_i}{\sum_{l:x_l \in M_0} w_l}\right) = \frac{\alpha w_i}{\sum_{l:x_l \in M_0} w_l}$$

and therefore with non-negative weights for the variables outside M_0

$$\begin{aligned} P\left(\min_{i:x_i \in M_0} \frac{p_i}{w_i} \leq \frac{\alpha}{\sum_{l=1}^{k} w_l}\right) &\leq P\left(\min_{i:x_i \in M_0} \frac{p_i}{w_i} \leq \frac{\alpha}{\sum_{l:x_l \in M_0} w_l}\right) \\ = P\left(\bigcup_{i:x_i \in M_0} \left(\frac{p_i}{w_i} \leq \frac{\alpha}{\sum_{l:x_l \in M_0} w_l}\right)\right) &\leq \sum_{i:x_i \in M_0} \frac{\alpha w_i}{\sum_{l:x_l \in M_0} w_l} = \alpha. \end{aligned}$$

As this is true for each matrix $\mathbf{W}_0$ to be conditioned on, it is true for the unconditional distribution as well. □

3.2. The multivariate two-sample case

In the case of two independent samples from two k-dimensional normal populations with equal covariance matrix

$$\mathbf{x}_{hj} = \begin{pmatrix} x_{hj1} \\ \vdots \\ x_{hjk} \end{pmatrix} \sim N_k(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}), h = 1, 2; \quad j = 1, \ldots, n_h, \ n = n_1 + n_2,$$

we consider the usual two-sample t statistics

$$t_i = \frac{\overline{x}_{1i} - \overline{x}_{2i}}{\sqrt{\sum_{h=1}^{2}\sum_{j=1}^{n_h}(x_{hji} - \overline{x}_{hi})^2}}\sqrt{\frac{(n-2)n_1 n_2}{n}}$$

and their p-values $p_i = 2(1 - F_{t(n-2)}(\mid t_i \mid))$. The weights are defined as $w_i(\eta) = g_i^\eta$, where $g_i = \sum_{h=1}^{2}\sum_{j=1}^{n_h}(x_{hji} - \overline{x}_i)^2$ and $\overline{x}_i = \sum_{h=1}^{2}\sum_{j=1}^{n_h} x_{hji}/n$. The formal procedure is then the same as in the one-sample case.

For the proof that the procedure keeps the familywise error, we appeal once again to closure arguments and consider the null set M_0, of size k_0, of variables x_i for which the local hypothesis $H_i : \mu_{1i} = \mu_{2i} = \tilde{\mu}_i$ is true. The corresponding $n \times k_0$-submatrix of the matrix of sample vectors is again denoted as $\mathbf{Y} \sim N_{n \times k_0}(\mathbf{1}_n \tilde{\boldsymbol{\mu}}', \mathbf{I}_n \otimes \boldsymbol{\Sigma}_0)$, where $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_{1,\ldots,}\tilde{\mu}_{k_0})'$. In this two-sample case, the matrix $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_{k_0})$ is no longer left-spherically distributed as the expectation is different from zero. However, we can consider the transformed data matrix $\tilde{\mathbf{Y}} = \mathbf{E}'\mathbf{Y}$ now, where the $n \times (n-1)$-matrix $\mathbf{E}$ is defined by $\mathbf{E} = (\mathbf{k}_1, \mathbf{K}_2)$ with the $n \times 1$-vector $\mathbf{k}_1 = \sqrt{\frac{n_1 n_2}{n}}\begin{pmatrix} +\mathbf{1}_{n_1}/n_1 \\ -\mathbf{1}_{n_2}/n_2 \end{pmatrix}$, $\mathbf{1}_m$ as vector of m 1s, and $\mathbf{K}_2$ as $n \times (n-2)$-matrix such that $(\mathbf{1}_n/\sqrt{n}, \mathbf{k}_1, \mathbf{K}_2)$ is an orthogonal matrix. Because all columns of $\mathbf{E}$ are orthogonal to each other and to the vector $\mathbf{1}_n$, $\tilde{\mathbf{Y}} \sim N_{(n-1)\times k_0}(\mathbf{E}'\mathbf{1}_n\tilde{\boldsymbol{\mu}}', \mathbf{E}'\mathbf{I}_n\mathbf{E} \otimes \boldsymbol{\Sigma}_0) = N_{(n-1)\times k_0}(\mathbf{0}, \mathbf{I}_{n-1} \otimes \boldsymbol{\Sigma}_0)$ under the null hypothesis of no mean differences between both populations, such that $\tilde{\mathbf{Y}}$ is left-spherically distributed. Then again the conditional distribution of $\tilde{\mathbf{Y}}$ and of all its columns for fixed

$$\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}} = \mathbf{Y}'\mathbf{E}\mathbf{E}'\mathbf{Y} = \mathbf{Y}'\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\right)\mathbf{Y}$$

is also left-spherical. Finally, utilizing

$$\begin{aligned}\sum_{h=1}^{2}\sum_{j=1}^{n_k}\left(y_{hji} - \overline{y}_{hi}\right)^2 &= \sum_{h=1}^{2}\sum_{j=1}^{n_k} y_{hji}^2 - n\overline{y}_i^2 - \frac{n_1 n_2}{n}\left(\overline{y}_{1i} - \overline{y}_{2i}\right)^2 \\ &= \mathbf{y}_i'\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n' - \mathbf{k}_1\mathbf{k}_1'\right)\mathbf{y}_i = \mathbf{y}_i\mathbf{K}_2\mathbf{K}_2'\mathbf{y}_i\end{aligned}$$

($\overline{y}_{hi}$ and $\overline{y}_i$ are the groupwise and total mean of the ith variable, whose values are collected in the ith column $\mathbf{y}_i$ of $\mathbf{Y}$), we can reformulate the local test statistics t_i in terms of $\tilde{\mathbf{Y}} = (\tilde{y}_{ji})$ as

$$t_i = \frac{\overline{y}_{1i} - \overline{y}_{2i}}{\sqrt{\sum_{h=1}^{2}\sum_{j=1}^{n_k}(y_{hji} - \overline{y}_{hi})^2}}\sqrt{\frac{(n-2)n_1 n_2}{n}} = \frac{\mathbf{k}_1'\mathbf{y}_i}{\sqrt{\frac{\mathbf{y}_i'\mathbf{K}_2\mathbf{K}_2'\mathbf{y}_i}{n-2}}} = \frac{\tilde{y}_{1i}}{\sqrt{\frac{\sum_{j=1}^{n-1}\tilde{y}_{ji}^2}{n-2}}}.$$

As these test statistics each have the t distribution with d.f. $n-2$ for independent standard normal variables $\tilde{y}_{ji}$ $(i = 1, \dots, n-1)$, they have the same distribution with left-spherically distributed variables (according to the above mentioned theorem). If we now notice that the weights g_i are fixed for fixed $\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}$ (they are just the diagonal elements), then the rest of the proof is quite analogous to the one-sample case.

4. A simulation study

We consider the following model for multivariate two-sample data. First, the variances of each of the $k = 5000$ measurements are assumed to be independently generated as $\sigma_i^2 \sim \tau_0\lambda/\chi^2_\lambda$, $i = 1, \dots, k$, where χ^2_λ denotes a chi-square distributed random variable with λ degrees of freedom. The parameter τ_0 is a nuisance parameter reflecting overall scale; for convenience it is taken to be unity in the simulations. The parameter λ is specified in the simulations; small λ corresponds to large variance heterogeneity across variables, while $\lambda = \infty$ corresponds to variance homogeneity across variables; we take $\lambda = 1$ and $\lambda = 200$ in the simulations. We assume variance homogeneity across the two samples: $\sigma_{i1}^2 = \sigma_{i2}^2 = \sigma_i^2$, all $i = 1, \dots, k$.

Next, conditional on σ_i^2, the effect sizes $V_i = \theta_i/\sigma_i$ are assumed to be drawn independently from a mixture of $N(0, \sigma^2_{\text{eff}})$ and single point (0) distributions, where $\theta_i = \mu_{1i} - \mu_{2i}$. The parameter σ^2_{eff} is specified in the simulations; larger σ^2_{eff} denotes generally larger alternatives. We take $\sigma^2_{\text{eff}} = 10$ in all simulations. The mixing parameter is denoted π, with $\pi = P(\theta_i = 0)$, and is specified as $\pi = 0.8$ in all simulations. Finally, conditional upon the means and variances, the measurement vector is assumed to come from a k-dimensional multivariate normal distribution with compound symmetry correlation for some fixed ρ.

FWE control proven in the previous section holds for fixed variances and noncentrality parameters; since FWE control holds conditionally, it holds unconditionally as well.

Power is taken to be the average fraction of true rejected hypotheses per simulation. Specifically, defining $S_1 = \{i \,|\, H_i \text{ is rejected}\}$ and $S_2 = \{i \,|\, \theta_i \neq 0\}$, we define $\text{Power} = E(|S_1 \cap S_2|/|S_2| \; \big| \; |S_2| > 0)$.

Figure 1 shows the effects of sample size, variance heterogeneity and correlation on the power of weighted procedures that use $w_i(\eta) = g_i^\eta$ as weights, when $k = 5000$, using 1000 simulated data sets to estimate power. FWF = 0.05 for all cases. The "maxT" (MT) version of the Westfall-Young procedure (Dudoit et al., 2003) using 199 samples from the permutation distribution is included for comparison. Clearly, the limiting case $\eta = \infty$ of KL is attractive for the case of small sample sizes and variance homogeneity, while the standard Bonferroni-Holm and MT procedures becomes more attractive for larger sample sizes and/or large variance heterogeneity.

5. Analysis of the gene expression data sets

5.1. Yeast genome expression with and without amino acids

A gene expression data set (courtesy of Jennifer Fostel, Pharmacia Corporation) consists of six Affymetrix-type arrays using the yeast genome. Three arrays are grown in a medium lacking amino acids, another three are grown in a medium containing amino acids, thus $n_1 = n_2 = 3$. There are $k = 9732$ genes evaluated using six arrays. For convenience, the genes are labeled simply as "1, 2, 3,. . . ", where the ordering is determined by the ordering of the combined sample quadratics $g_i = \sum_{h=1}^{2}\sum_{j=1}^{3}(x_{hji} - \overline{x}_i)^2$ and $\overline{x}_i = \sum_{h=1}^{2}\sum_{j=1}^{3} x_{hji}/6$. (The raw data and more detailed gene labels are available from the first author.)

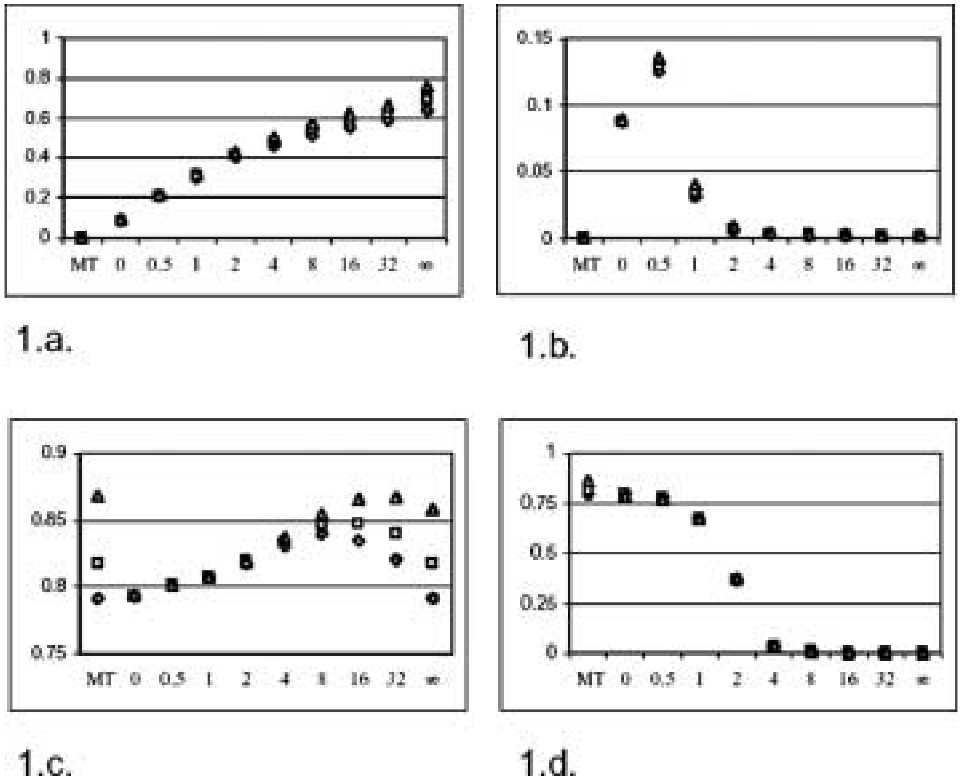

Figure 1: Power as a function of η, for $n_i = 3$, $\lambda = 200$ (1a), $n_i = 3$, $\lambda = 1$ (1b), $n_i = 10$, $\lambda = 200$ (1c), and $n_i = 10$, $\lambda = 1$ (1d). Diamonds indicate $\rho = 0.0$, squares $\rho = 0.6$, and triangles $\rho = 0.9$

Figures 2a and 2b displays the spectrum of results for the weighted procedures, from $\eta = 0$ (Holm's method) to $\eta = \infty$ (KL method). In this example, the Holm and KL extremes are comparable in that both reject about the same number of hypotheses, although it appears in this example that intermediate values η may provide greater sensitivity. It is unfortunate that the genes selected are quite different at the extremes: for Holm and KL, there is only one common selection (gene 11).

Figure 2b reports all testing results using adjusted p-values; for the weighted procedures these are defined sequentially as $\tilde{p}_{(1)} = \Sigma_{h \in \mathbf{S}_1} w_h q_{(1)}$, $\tilde{p}_{(2)} = \max\{\Sigma_{h \in \mathbf{S}_2} w_h q_{(2)}, \tilde{p}_{(1)}\}$, $\ldots$; for the KL procedure these are defined as $\tilde{p}_{(1)} = p_1$, $\tilde{p}_{(2)} = \max\{p_2, \tilde{p}_{(1)}\}$, $\ldots$, assuming that the labels $1, 2, \ldots$ correspond to the ordering $g_1 > g_2 > \cdots$. All genes with adjusted p-values less than α are thus rejected at FWE= α, and the size of the adjusted p-value indicates multiplicity adjusted strength of evidence. In this example, no dramatic differences in adjusted p-values are noted.

As suggested by Figure 1, the MT procedure yields a result that is most comparable to the $\eta = 0$ results, since the standard MT procedure does not use weights. In this case with such small samples, direct enumeration of the $\binom{6}{3} = 20$ permutation samples is possible. Because the sample sizes are very small, the MT procedure is not as useful because of the "graininess" of the permutation distribution. The

Figure 2a: Gene labels rejected at FWE = .10 level for various η, yeast genome/amino acid study

$\eta = 0.0$	$\eta = 0.5$	$\eta = 1$	$\eta = 2$	$\eta = 4$	$\eta = 8$	$\eta = 16$	$\eta = 32$	$\eta = \infty$
1484	1484	11	11	11	11	1	1	1
377	11	377	7	7	1	3	3	2
3526	377	1484	377	1	3	11	2	3
11	398	398	226	3	7	2	5	4
1553	226	226	8	8	2	7	6	5
2643	1553	97	97	9	8	6	4	6
5127	3526	7	9	2	9	5	7	7
398	640	424	3	19	6	8	11	8
2175	424	640	398	17	5	4	8	9
3661	97	8	1	6	4	9	9	10
226	2643	53	43	43	17	10	10	11
	579	131	19	5	10	12	12	12
	980	43	53	97	19	17		
	2175	579	17	10	12			
	1713	1553	2	53				
	698	9	131					
	507	507						

Figure 2b: Adjusted p-values corresponding to genes listed in Figure 1a

$\eta = 0.0$	$\eta = 0.5$	$\eta = 1$	$\eta = 2$	$\eta = 4$	$\eta = 8$	$\eta = 16$	$\eta = 32$	$\eta = \infty$
0.0012	0.0008	0.0004	0.0002	0.0002	0.0010	0.0030	0.0030	0.0030
0.0128	0.0014	0.0022	0.0088	0.0068	0.0042	0.0052	0.0044	0.0042
0.0131	0.0023	0.0028	0.0102	0.0097	0.0068	0.0052	0.0059	0.0042
0.0278	0.0069	0.0068	0.0214	0.0111	0.0068	0.0087	0.0696	0.0612
0.0319	0.0122	0.0085	0.0228	0.0111	0.0132	0.0163	0.0696	0.0612
0.0334	0.0224	0.0188	0.0228	0.0152	0.0132	0.0526	0.0696	0.0612
0.0336	0.0268	0.0258	0.0294	0.0234	0.0200	0.0526	0.0696	0.0612
0.0364	0.0345	0.0367	0.0329	0.0467	0.0465	0.0526	0.0696	0.0612
0.0734	0.0352	0.0507	0.0329	0.0540	0.0465	0.0627	0.0696	0.0612
0.0806	0.0381	0.0660	0.0342	0.0716	0.0815	0.0627	0.0696	0.0612
0.0921	0.0449	0.0692	0.0453	0.0762	0.0815	0.0627	0.0696	0.0612
	0.0547	0.0706	0.0473	0.0787	0.0815	0.0627	0.0696	0.0612
	0.0642	0.0716	0.0493	0.0787	0.0815	0.0627		
	0.0777	0.0725	0.0693	0.0892	0.0815			
	0.0799	0.0812	0.0752	0.0892				
	0.0800	0.0822	0.0834					
	0.0820	0.0970						

Figure 2: Analysis of yeast data

smallest observed MT adjusted p-value for these data is 0.10, occurring for only 9 of the top 11 genes flagged as significant by B-H (corresponding to $\eta = 0$), and no MT significances were found at the FWE=0.05 level as compared to 8 flagged.by B-H (Figure 2b).

5.2. Comparing expression levels of different leukemia types

Data analyzed by Golub et al. (1999) (http://www-genome.wi.mit.edu/cancer/) relate gene expression from 7129 genes to disease status (the article by Golub et al. actually screens out several of the 7129, leaving 6817 for their analysis). There are $n_1 = 11$ patients with acute myeloid leukemia (AML) and $n_2 = 27$ with acute lymphoblastic leukemia (ALL).

Figures 3a and 3b displays the results for these data. Unfortunately, in this data set, the largest g_i happened to correspond to an insignificant ($p = 0.5835$, unadjusted) gene, so the case $\eta = \infty$ produced no significant result. However, there is large variance heterogeneity in the data, which can be somewhat alleviated using the cube root transformation as suggested by Tusher et al. (2001), applying this method yields two significances for $\eta = \infty$ at the $\alpha = 0.10$ level. Cube root or not, in this example there are dramatic differences favoring small η, likely because of the larger sample sizes, as suggested by Figure 1. The analyses of Figure 3 use the untransformed data.

We find an additional 10 significant genes at the 0.10 level and an additional 3 significant genes at the 0.05 level using MT and 1,000,000 samples from the permutation distribution. One suspects that additional genes are flagged here because the method utilizes correlation information (see Westfall and Young, 1993, for more details); however, since the method is non-parametric, using the permutation distribution instead of the normal distribution, it is possible that additional genes are flagged because of non-normal characteristics of the data. Indeed, it sometimes happens that the MT method flags fewer genes than Bonferroni-Holm, despite accounting for correlation structures (as shown in the previous example), because of the graininess of the permutation distribution, and because it accounts for non-normalities exactly, while the parametric B-H method is only approximate.

6. Discussion

We have shown that FWE-controlling multiple testing methods are possible with such high-dimensional data as occurs with gene expression data. Further, we have shown that by using the random weights g_i, not only can we “pre-specify” and weight tests so as to maintain FWE control, but we can increase the power of the tests, especially in the case of small sample sizes and variance homogeneity across variables.

Our simulations show that there are cases where intermediate η provide more power than the extremes at 0 and ∞. Clearly, larger values of η are preferred for small studies with reasonably homogenous variance. Experience with similar data and appropriate transformations may suggest whether to use a small (perhaps 0) or large (perhaps ∞) value of η. Further research may be needed to specify a more specific η a priori, or to estimate such an η adaptively from the data, while maintaining FWE control.

Figure 3a: Gene labels rejected at FWE = .10 level for various η, leukemia study

$\eta = 0.0$	$\eta = 0.5$	$\eta = 1$	$\eta = 2$	$\eta = 4$	$\eta = 8$	$\eta = 16$	$\eta = 32$	$\eta = \infty$
3320	3320	4847	4847	6201	6201			
4847	4847	3320	6201	1674				
2020	2020	2020	3320	1882				
1745	1745	6201	4196	2186				
5039	4196	4196	1882	4196				
1834	5039	1745	1674	4847				
461	6201	1882	2186	2402				
4196	2288	2288	6200	6200				
3847	1834	5039	2288	6803				
2288	1882	3258	2402	1394				
1249	3258	1674	2020	6806				
6201	1249	6200	3258	6797				
2242	6200	1249	6803					
3258	3847	2186	6806					
1882	2121	2402	1394					
2111	1674	2121	2121					
(30 more)	(40 more)	(33 more)	(14 more)					

Figure 3b: Adjusted p-values corresponding to qenes listed in Figure 3a

$\eta = 0.0$	$\eta = 0.5$	$\eta = 1$	$\eta = 2$	$\eta = 4$	$\eta = 8$	$\eta = 16$	$\eta = 32$	$\eta = \infty$
0.0000	0.0000	0.0000	0.0000	0.0001	0.0262			
0.0000	0.0000	0.0000	0.0000	0.0008				
0.0000	0.0000	0.0000	0.0000	0.0013				
0.0001	0.0000	0.0000	0.0001	0.0029				
0.0001	0.0001	0.0000	0.0001	0.0033				
0.0001	0.0001	0.0001	0.0001	0.0083				
0.0003	0.0001	0.0001	0.0003	0.0100				
0.0004	0.0001	0.0001	0.0003	0.0109				
0.0005	0.0001	0.0002	0.0004	0.0333				
0.0006	0.0002	0.0002	0.0004	0.0393				
0.0012	0.0002	0.0002	0.0006	0.0594				
0.0012	0.0003	0.0002	0.0007	0.0750				
0.0014	0.0005	0.0003	0.0008					
0.0015	0.0005	0.0003	0.0014					
0.0023	0.0007	0.0004	0.0016					
0.0026	0.0007	0.0004	0.0016					
(30 more)	(40 more)	(33 more)	(14 more)					

Figure 3: Analysis of leukemia data

References

[1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A new and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 1289–1300. MR1325392

[2] Benjamini, Y. and Hochberg, Y. (1997). Multiple hypothesis testing with weights. *Scandinavian Journal of Statistics* **24**, 407–418. MR1481424

[3] Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* **18**, 71–103. MR1997066

[4] Fang, K.-T. and Zhang, Y.-T. (1990). *Generalized Multivariate Analysis*, Science Press, Beijing. MR1079542

[5] Ge, Y., Dudoit, S. and Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *Test* **12**, 1–77. MR1993286

[6] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., 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] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6**, 65–70. MR538597

[8] Kropf, S. and Läuter, J. (2002). Multiple tests for different sets of variables using a data-driven ordering of hypotheses, with an application to gene expression data. *Biometrical Journal* **44**, 789–800. MR1934963

[9] Kropf, S., Läuter, J., Eszlinger, M., Krohn, K. and Paschke, R. (2004). Nonparametric multiple test procedures with data-driven order of hypotheses and with weighted hypotheses, to appear in *Journal of Statistical Planning and Inference* **125**, 31–47.

[10] Läuter, J. (1996). Exact t and F tests for analysing studies with multiple endpoints. *Biometrics* **52**, 964–970. MR1411742

[11] Läuter, J., Glimm, E. and Kropf, S. (1996). New multivariate tests for data with an inherent structure. *Biometrical Journal* **38**, 5–23. Erratum:*Biometrical Journal* **40**, 1015. MR1405950

[12] Louis, T.A. and Bailey, J. K. (1990), Controlling marginal error rates using prior information and marginal totals to select tumor sites, *Journal of Statistical Planning and Inference* **24**, 297–316. MR1046967

[13] Maurer, W., Hothorn, L.A. and Lehmacher, W. (1995). Multiple comparisons in drug clinical trials and preclinical assays: A-priori ordered hypotheses. In Vollman, J., editor, *Biometrie in der chemische-pharmazeutichen Industrie*, Volume 6. Stuttgart: Fischer Verlag.

[14] Rosenthal, R. and Rubin, D.B. (1983). Ensemble-adjusted p-Values. *Psychological Bulletin* **94**, 540–541.

[15] Spjøtvoll, E. (1972). On the optimality of some multiple comparison procedures. *Ann. Math. Statist.* **43**, 398–411. MR301871

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

[17] Tusher, V. G., Tibshirani and R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* **98**(9), 5116–5121.

[18] Westfall, P. H., Krishen, A. and Young, S. S. (1998). Using prior information to allocate significance levels for multiple endpoints. *Statistics in Medicine* **17**, 2107–2119.

[19] Westfall, P. H. and Soper, K. A. (2001). Using priors to improve multiple animal carcinogenicity tests, *Journal of the American Statistical Association* **96**, 827–834. MR1963409

[20] Westfall, P. H. and Krishen, A. (2001). Optimally weighted, fixed sequence, and gatekeeping multiple testing procedures, *Journal of Statistical Planning and Inference* **99**, 25–40. MR1858708

[21] Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment.* Wiley, New York.