Biometrics 67, 174–181
March 2011
DOI: 10.1111/j.1541-0420.2010.01443.x

k-FWER Control without *p*-value Adjustment, with Application to Detection of Genetic Determinants of Multiple Sclerosis in Italian Twins

L. Finos[1,*] and A. Farcomeni[2,]**

[1]Department of Statistical Sciences, University of Padua, Padua, Italy and Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, The Netherlands
[2]Department of Hygiene and Public Health Sapienza—University of Rome, Rome, Italy
[*]*email:* livio@stat.unipd.it
[**]*email:* alessio.farcomeni@uniroma1.it

Summary. We show a novel approach for *k*-FWER control which does not involve any correction, but only testing the hypotheses along a (possibly data-driven) order until a suitable number of *p*-values are found above the uncorrected α level. *p*-values can arise from any linear model in a parametric or nonparametric setting. The approach is not only very simple and computationally undemanding, but also the data-driven order enhances power when the sample size is small (and also when *k* and/or the number of tests is large). We illustrate the method on an original study about gene discovery in multiple sclerosis, in which were involved a small number of couples of twins, discordant by disease. The methods are implemented in an `R` package (`someKfwer`), freely available on CRAN.

Key words: Data-driven order; Gene discovery; *k*-Familywise error rate; Multiple sclerosis; Multiple testing.

1. Introduction

The statistical analysis of DNA microarrays often leads to the evaluation of the significance of thousands of hypotheses simultaneously. These applications are also often characterized by the lack of information due to small sample size, weak effect sizes, very small fraction of true positives, and dependence among the test statistics.

Our motivation for this work comes from an original study on multiple sclerosis. Thirteen couples of homozygotic Italian twins, discordant by disease, were enrolled at Center for Experimental Neurological Therapy of Sant'Andrea hospital in Rome (Italy). A small quantity of mRNA was drawn from each twin of the 13 couples; a red dye was assigned to the diseased twin and a green dye to the healthy twin for gene expression mapping through a DNA microarray experiment. mRNAs for each couple were put on a slide, finally recording the expression levels of $m = 8570$ genes on 13 slides.

Large-scale transcriptional expression profiling allows screening for differentially expressed genes in a discovery-driven fashion. As a complement, real-time reverse transcription and/or polymerase chain reaction (RT-PCR) can be used for more targeted profiling after gene selection through multiple testing. Gene expression profiling is a powerful tool for identifying novel molecular biomarkers. Powerful statistical tools for multiple testing are needed at the screening level, in order not to exclude important biomarkers from the list of genes candidate for further investigation through the complementary techniques.

In our motivating example, the use of twins leads to having an overwhelming majority of genes equally expressed in the pair. The signal is then sparse and weak. The sample size is small, especially if compared to the number of genes involved. This led us to investigate the possibility of a powerful approach of multiple testing especially devised for cases in which the number of samples is small. More details and an analysis of this data set can be found in Section 5.

The problem of gene discovery is easily cast in the area of multiple hypothesis testing, as discussed above. For reviews refer to Dudoit, Shaffer, and Boldrick (2003); to the books by Westfall and Young (1993) and Hochberg and Tamhane (1987); and to Farcomeni (2008) for recent developments. In a multiple testing problem, the *p*-values should be corrected in order to take into account the multiplicity and control a suitable generalization of the single-inference Type I error rate. This usually reduces to comparing the raw *p*-values with a rank-dependent threshold, which is also a function of the number of tests and is often much smaller than the overall significance level α.

There are many different generalizations of the Type I error rate that can be put forward. One possibility is given by the *k*-FWER (*k*-Familywise Error Rate), defined as the probability of having *k* or more false positives. This is a generalization of the well-known FWER (the 1-FWER according to our definition). Allowing for more than one false positive is seen to be liberal enough so to allow for satisfactory power when the number of tests is high. There now are available a number of methods controlling the *k*-FWER. A step-down approach is used in Lehmann and Romano (2005). van der Laan, Dudoit, and Pollard (2004) present augmentation procedures. One common drawback is that those methods are somewhat conservative, in that they often have an error rate well below the nominal α. In this regard, Guo and Romano (2007)

give procedures that dramatically improve power under independence of the test statistics and Romano and Wolf (2007) present methods that can be used also under dependence, which are based on a resampling approach. Sarkar (2008) makes use of the kth order joint null distributions obtaining k-FWER control under the assumption of positive dependence among the test statistics.

The goal of this article is to develop a simple but powerful approach for controlling the k-FWER which is not computationally intensive and that achieves high power especially with the lack of information. We anticipate the power of our procedure will be enhanced in cases of approximate homoscedasticity of the error terms. The strategy we suggest relies on pseudo-gatekeeping, in which hypotheses are tested in a (possibly data-driven) order without correction for multiplicity. The p-values are ordered with respect to a (data-driven) exogenous criterion, and compared sequentially with the single-step cutoff α. That is, at each step we simply perform uncorrected testing. Once an uncorrected p-value is found above the α level, we do not stop the procedure but keep rejecting until a number $J(k,\alpha)$ of p-values are found above the α level; where $J(k,\alpha)$ is to be defined below. After the algorithm is stopped, the processed hypotheses corresponding to p-values below α are rejected. p-values can arise from one or two sample t-testing, ANOVA, ANCOVA, or regression. One can adjust the p-values for confounders and nonparametric approaches can be accommodated via rank-based testing, permutation, or the rank transformation (Conover and Iman, 1982).

When the order of the hypotheses is not data driven, the procedure can be seen as an extension of the 1-FWER controlling procedure of Maurer, Hothorn, and Lehmacher (1995). When there is a data-driven order, the procedure is an extension of Kropf and Läuter (2002) and Kropf et al. (2004). The main difference with those methods is that we do not stop at the first uncorrected p-value above the α level, but allow for a suitable number of jumps, obtaining k-FWER control. Also Hommel and Kropf (2005) consider testing along a given order and allowing for a number of jumps, but their proposal can be directly used to control only the 1-FWER.

We also give an extension of the procedure, which does not rely on any assumption concerning the dependency structure. An `R` (`R` Development Core Team, 2007) package, under the name `someKfwer`, is available on CRAN (`http://cran.r-project.org`).

The rest of the article is as follows. In Section 2, we present our proposed procedure and prove it controls the k-FWER under independence. In Section 3, we discuss extensions under dependence. In Section 4, we illustrate and compare the method via a simulation study, and in Section 5 we analyze the multiple sclerosis data set.

2. k-FWER Control with Possibly Data-Driven Order of the Hypotheses under Independence

2.1 *k-FWER Control of Ordered Hypotheses*

First let us assume that the hypotheses are naturally ordered and shall be tested sequentially. This is not a theoretical situation: ordered hypotheses arise in dose–response studies, in toxicity studies, in observational studies when comparing a treatment to more than one type of control (Rosenbaum, 2008), and in other cases. See for instance Marcus, Peritz, and Gabriel (1976); Hsu and Berger (1999); Maurer et al. (1995); and Strassburger, Bretz, and Finner (2007).

The k-FWER can be controlled by performing tests sequentially at the uncorrected level α. Sequential testing is stopped after $J(k,\alpha)$ p-values are found above level α, where $J(k,\alpha)$ is to be defined below, and is fixed before the experiment. After the sequential testing is stopped, all hypotheses corresponding to p-values above α and hypotheses not yet reached by the sequential testing (regardless of their significance level) are not rejected.

In what follows, let $F_{\mathcal{B}neg(s,\pi)}(x)$ be the CDF of a negative binomial random variable with parameters s and π, and let $Q_{\mathcal{B}neg(s,\pi)}(y)$ be its y quantile.

In summary, denoting with $p_{(1)},\dots,p_{(m)}$ the m p-values ordered with respect to the natural ordering of the m hypotheses, we propose the *sequential procedure* stated in algorithm 1.

Algorithm 1: Sequential procedure for naturally ordered hypotheses

Set J as the α-quantile of a negative binomial random variable with parameters k and α:

$$J(k,\alpha) = Q_{\mathcal{B}neg(k,\alpha)}(\alpha). \tag{1}$$

Set $j := 0, i := 1$
while $i < m$ & $j \leqslant J(k,\alpha)$ **do**
  $i := i + 1$
  if $p_{(i)} \geqslant \alpha$ **then**
    $j := j + 1$
  end if
end while

Reject the i hypotheses considered until stopping which correspond to a p-value below α. Do not reject the hypotheses corresponding to a p-value above α and the hypotheses which have not been reached by the sequential testing even if they correspond to p-values below α. If the final number of rejected hypotheses is lower than $k-1$, augment the set of rejected hypotheses until $k-1$ are rejected, by adding the hypotheses corresponding to the most significant p-values.

Unlike many other approaches, the proposed sequential procedure does control multiplicity without correcting the level α of individual hypotheses, and k-FWER control is obtained by pseudo-gatekeeping: after $J(k,\alpha)$ p-values are found above α there is no further rejection.

We now formally state our main results:

THEOREM 1: *Define $\mathcal{S}_J$ the set of hypotheses tested before the Jth unrejected hypothesis. Let $\mathcal{S}_J^0 \subseteq \mathcal{S}_J$ be the set of true null hypotheses tested before the Jth unrejected hypothesis. Let $\phi_i = 1$ $(i \in 1,\dots,m)$ if the ith hypothesis is rejected at level α and $\phi_i = 0$ otherwise. Call $\mathcal{H}_0$ the collection of m_0 true null hypotheses and assume that the distributions of the p-values of its elements stochastically dominate the uniform. The remaining $m_1 = m - m_0$ hypotheses under the alternative are collected in $\mathcal{H}_1$ and the distribution of their p-values are stochastically dominated by the null distribution(s). Suppose the test statistics are independent. We have:*

(*i*) *For a fixed J, the probability of k or more type I errors before the Jth jump is bounded by:*

$$P\left(\sum_{i\in\mathcal{S}_J^0}\phi_i \geqslant k\right) \leqslant 1 - F_{\mathcal{B}neg(J,1-\alpha)}(k-1)$$

$$\forall(k,\alpha), \quad \forall\mathcal{S}_J^0 \subseteq \mathcal{H}_0$$

where $\mathcal{B}neg(r,\pi)$ *defines a negative binomial random variable with parameters r and* π.

(*ii*) *The sequential procedure in algorithm 1 with* $J(k,\alpha)$ *defined as in (1) controls the k-FWER at level* α.

Proof. We first prove point (*i*). Suppose tests are α-size: $P(\phi_i = 1 \,|\, H_0) \leqslant \alpha \;\forall i \in \mathcal{H}_0$.

Consider the quantity $\sum_{i\in\mathcal{S}_J^0}\phi_i$, which counts the number of true null hypotheses rejected before J p-values are found above α. This quantity is clearly smaller than or equal to the number of true null hypotheses rejected before J p-values *corresponding to true null hypotheses* are found above α, which we denote with $\sum_{i\in\mathcal{S}_{J_0}^0}\phi_i$. The tail probability of $\sum_{i\in\mathcal{S}_{J_0}^0}\phi_i$ is bound by the tail probability of a negative binomial with parameters J and $1-\alpha$. Consequently, we have the thesis: $P(\sum_{i\in\mathcal{S}_J^0}\phi_i \geqslant k) \leqslant P(\sum_{i\in\mathcal{S}_{J_0}^0}\phi_i \geqslant k) \leqslant 1 - F_{\mathcal{B}neg(J,1-\alpha)}(k-1)$.

To prove the point (*ii*) we note that (Johnson, Kemp, and Kotz, 2005): $1 - F_{\mathcal{B}neg(J,1-\alpha)}(k-1) = 1 - F_{\mathcal{B}(J-1+k,\alpha)}(k-1) = F_{\mathcal{B}(J-1+k,1-\alpha)}(J-1) = F_{\mathcal{B}neg(k,\alpha)}(J-1)(\mathcal{B}(n,\pi)$ defines a binomial random variable with n trials and probability π).

Since (*i*) holds for every J, we set $J(k,\alpha)$ as the highest J such that $1 - F_{\mathcal{B}neg(J,1-\alpha)}(k-1) \geqslant 1-\alpha$, which is $Q_{\mathcal{B}neg(k,\alpha)}(\alpha)$. This maximizes power.

Since (*i*) holds $\forall\mathcal{T} \subseteq \mathcal{S}^0$, this guarantees the k-FWER control.

Finally, if less than $k-1$ hypotheses are rejected, the k-FWER is trivially equal to zero. Hence, if the final number of selected hypotheses is lower than $k-1$, we can add rejections up to $k-1$ without violating k-FWER control. This is the reason behind the final sentence in Algorithm 1 (on augmentation see also van der Laan et al., 2004).

It is straightforward to check that $J(1,\alpha) = 0$ and then when controlling the classical FWER we get back the Maurer et al. (1995) procedure. In the k-FWER control setting our approach is particularly advantageous in terms of power with respect to other procedures, in particular when k is large, as we will illustrate below. Table 1 shows some values of $J(k,\alpha)$ as a function of k and α.

Table 1
Number of jumps $J(k,\alpha)$ *in sequential testing for different values of k and* α

	k										
	1	2	3	4	5	6	7	8	9	10	20
$\alpha = 0.10$	0	4	9	15	21	27	34	40	47	54	128
$\alpha = 0.05$	0	6	14	25	36	48	61	74	87	101	249
$\alpha = 0.01$	0	14	42	80	125	175	228	285	345	406	1093

There are two features that are somewhat surprising: first of all, $J(k,\alpha)$ does not depend on the number of tests. This could be expected since k does not depend on m. It can also happen that $J(k,\alpha)$ is large or even larger than m. An example is easily given: assume one is testing $m = 100$ hypotheses with $\alpha = 0.05$ and $k = 10$. One can reject all hypotheses below $\alpha = 0.05$ (and in fact $J(10, 0.05) = 101$ jumps are allowed) since the probability of having 10 or more false positives would be lower than or equal to 0.011. As the number of true null hypotheses grows, furthermore, sequential testing is stopped earlier with higher probability, regardless of m. In fact, p-values arising from true null hypotheses are above the threshold with higher probability due to stochastic dominance assumption.

According to this reasoning, it can be understood why the number of allowed false positives k shall be set smaller when the number of tests is smaller.

The second feature is that $J(k,\alpha) = Q_{\mathcal{B}neg(k,\alpha)}(\alpha)$, being a quantile of a negative binomial distribution, is decreasing in α and increasing in k. This leads to the consequence that the number of jumps in a fixed-length sequence will be much higher for smaller α. Nevertheless, it is natural to expect the number of rejections to be smaller for smaller α, since even if more jumps are allowed, fewer and fewer p-values will be below the threshold.

2.2 *Data-Driven Order*

The hypotheses order should be chosen a priori, on the basis of experimental hypotheses. However, in most cases there is no natural order of the hypotheses, as in our motivating example. While in general an a posteriori data-driven ordering may lead to inflation of the nominal error rate, we can propose in this section a strategy for a data-driven ordering which does not inflate the error rate and which is chosen in order to enhance power.

The final procedure we propose is to order the hypotheses according to the criteria specified in this section, and then apply procedure 1.

In the following, we assume the following model:

$$Y_j = \mathbf{Z}_j\boldsymbol{\beta}_j + \epsilon_j, \tag{2}$$

where Y_j is a numerical response, $\mathbf{Z}_j$ is a fixed matrix of covariates (which may include dummy variables and/or a constant column), $\boldsymbol{\beta}_j$ is a vector of parameters and ϵ_j is distributed like a zero-centered Gaussian with variance σ_j^2; $j = 1,\ldots,m$. These variances are independent. In practice, especially in high dimensions, it often happens that variablewise covariates are not different, so that $\mathbf{Z}_j = \mathbf{Z}$.$p$-values arise from a test on linear hypotheses of the type

$$H_0 : \mathbf{M}_j\boldsymbol{\beta}_j = \mathbf{0}_{b_j},$$

with $\mathbf{M}_j$ denoting a full rank matrix and 0_{b_j} a vector of zeros of length b_j. Once again, it often happens that variablewise contrasts are not different, so that $\mathbf{M}_j = \mathbf{M}$ and $b_j = b$.

This setting includes, but is not limited to, one- and two-sample paired and unpaired t-tests, F-tests, tests on the correlations; also adjusted for confounders, depending on the construction of Z_j and Y_j. Extension to other parametric and nonparametric testing situations are discussed below.

We propose to order the hypotheses according to decreasing values of the second moment of residuals of the model (2), estimated constraining the parameters under the null hypothesis. The residuals are clearly zero-centered under the null, and their distribution does not depend on parameters involved in testing. Under the alternative, the second moment of the residuals is expected to be higher due to location shift.

The idea is easily understood if one thinks about the one-sample t-tests for a zero mean, in which residuals coincide with observations, and hence

$$M2_j = \sum_i y_{ij}^2/n = \sum_i (y_{ij} - \overline{y}_j)^2/n + (\overline{y}_j)^2 = (1 + \widehat{\delta_j}^2)\widehat{\sigma_j}^2,$$

with $\overline{y}_j = \sum_i y_{ij}/n, \widehat{\sigma_j}^2$ the (biased) estimated variance and $\widehat{\delta_j}$ the estimated normalized effect. Then, the ordering with respect to $M2_j$ enhances power since it is a proxy for the ordering with respect to δ_j. The smaller and closer to each other the variances σ_j^2, the better. To give a further example, suppose we are comparing two independent samples. In that case, Z is defined as a two-column matrix with a column of ones and a column that contains the indicator of one of the two groups. The ordering shall be done with respect to the column-wise mean-centered matrix of measurements (i.e., the residuals with respect to the estimated intercept under the null hypothesis of equal mean samples). The same result can be reached for $C > 2$ samples. In Section 5, we develop in detail the case of two paired samples.

In the most general case, we have that $E(Y_j^2)$ can be expressed as:

$$E\left(Y_j^2\right) = E\left(\left(\mathbf{Z}\boldsymbol{\beta}_j + \epsilon_j\right)^2\right) = \|\mathbf{Z}\boldsymbol{\beta}_j\|^2 + \mathrm{Var}(\epsilon_j)$$

($\|\cdot\|$ denotes the Euclidian norm). When $k = 1$ and one-, two-sample t-tests or F-tests are performed, the procedure reduces to the sequential testing with data-driven order of Kropf and Läuter (2002).

The proof that the ordering according to $M2_j$ does not inflate the 1-FWER has been given in Kropf and Läuter (2002) or in a similar situation in Westfall, Kropf, and Finos (2004). It is based on the theory of spherical distributions (Fang and Zhang, 1990) and it is a direct consequence of Theorem 1 in Lauter, Glimm, and Kropf (1998). In these papers, an arbitrary covariance structure of the variables is assumed. For the present extension to the k-FWER we need according to the proof in Section 2.1 the more restrictive assumption of independent variables (or test statistics, respectively).

One important feature of this data-driven criterion for ordering is that it promotes rejection of hypotheses with larger effect sizes, even if they may also be associated with larger p-values (but lower than α anyway), thus producing a list of rejected hypotheses potentially more interesting for the practitioner (see Kirk, 2007, and references therein).

The nonparametric setting can be accommodated in three different ways. First, one can simply use the rank-transformation of Conover and Iman (1982). Second, one can compute p-values from nonparametric rank-based methods, and order the hypotheses according to medians (possibly adjusted for confounders) in case of one-sample tests and to (possibly adjusted) interquartile ranges in case of $C \geqslant 2$ sample tests. The resulting method is a generalization of the approach of Kropf et al. (2004) for the classical 1-FWER, and a proof that the ordering according to the latter criterion does not inflate the k-FWER directly follows from their results. The third approach regards the use of p-values arising from permutation testing and the usual $M2_j$ based ordering. Finos and Salmaso (2006) show that any ordering that does not depend on the vector of permuted indexes used for shuffling the data is valid. This includes ordering based on $M2_j$. Based on the results of Finos and Salmaso (2006), our results extend to a broader class of statistics (e.g., interquartile ranges) whenever p-values arise from permutation testing.

3. Extension to Dependent Test Statistics

The m test statistics are in general not independent. Nevertheless, in certain situations multiple testing procedures devised for independent test statistics can be used under weak dependence. For formal discussions refer to Farcomeni (2007) and Clarke and Hall (2009). Typically, it is required that dependence is overwhelmed when the number of tests grows larger and larger. For instance, in microarray experiments a form of block (sometimes called "clumpy") dependence is usually expected, as argued for instance in Storey and Tibshirani (2003). Other weak dependence conditions are discussed in Farcomeni (2007). Under these conditions, our procedure (together with the procedures in Guo and Romano, 2007, and other procedures devised for independent test statistics) can be directly used when the number of tests m is large. A simulation study, reported in the Web Appendix, confirms that the error rate is controlled when m is large, especially when k is small, under weak dependence. With stronger dependence, the nominal error rates may be exceeded.

In such cases, or when the number of tests is too small to invoke asymptotic results (with m), a simple device is given by testing the individual hypotheses at level $\alpha' = \frac{k\alpha}{J(k,\alpha)+k}$, obtaining a slightly more conservative procedure which anyway is valid under general dependence. A proof of this statement is given in next theorem:

THEOREM 2: *Assume the distributions of the p-values under the null hypotheses stochastically dominate the uniform, and the distributions of the p-values under the alternative are stochastically dominated by the null distribution(s). Let* $\alpha' = \frac{k\alpha}{J(k,\alpha)+k}$. *Under general dependence among the test statistics we have that the sequential procedure 1 with* $J(k,\alpha)$ *defined as in (1) and in which the individual test level is fixed as* α', *controls the k-FWER at level* α.

Proof. Note that $|\mathcal{S}_J^0| \leqslant J(k,\alpha) + k$. By using the Markov inequality, we have

$$P\left(\sum_{i\in\mathcal{S}_J^0} \phi_i \geqslant k\right) \leqslant \frac{E\left(\sum_{i\in\mathcal{S}_J^0} \phi_i\right)}{k} \leqslant \frac{(J(k,\alpha)+k)\alpha'}{k} = \alpha.$$

This inequality replaces conclusion (i) of Theorem 1. The result trivially follows as in the proof of (ii) in Theorem 1.

Note that for $k = 1$ the version for independent and dependent test statistics coincide.

The ratio used in the proof is very similar to that used in Lehmann and Romano (2005) (the former makes use of $\mathcal{S}_J^0$,

while the latter considers the whole number of hypotheses under test). In this sense, the procedure of Theorem 2 can be viewed as a variant of the generalized Bonferroni procedure in Lehmann and Romano (2005).

4. Simulation Study

A brief simulation study is used to illustrate our methodology. We perform one-sample *t*-tests, with data generated from standard normals under the null hypothesis. We let $n = 5, 10, 20, 50$; $m = 500, 1000, 10{,}000$; $\alpha = 0.05$. We fix the mean under the alternative hypotheses so that the single tests have a prescribed power of 70% (hence, the mean under the alternative decreases as the sample size n increases) and the proportion of false null hypotheses is fixed at 10%. The position of the false hypotheses is selected randomly at each iteration.

For each setting we generate the data, compute *p*-values, and apply the Lehmann and Romano (2005) (LR) and Guo and Romano (2007) (GR) step-down procedures; together with our procedure with data-driven order of the hypotheses (ORD) and the version for arbitrary dependence in Theorem 2 (ORD_{dep}). The LR procedure leads us to compare the jth ordered *p*-value with the step-down constant $\alpha_j = k\alpha/(m + \min(k - j, 0))$; the GR procedure instead assumes independence among variables and models the process with a binomial random variable $\mathcal{B}(n, \pi)$ with $n = m$ (number of variables) and $\pi = \alpha$. We use the step-down version of this procedure which is a slight improvement of the original one. We perform 10,000 Monte Carlo iterations. We estimate power through the fraction of correctly rejected hypotheses.

The estimated power for $\alpha = 0.05$ and different values of k is reported in Figure 1.

Note that the case $k = 1$ is reported only for reference, since control of the 1-FWER is not the main focus of this article. Analogous results are obtained in other simulations settings, also under dependence. A complete report of our simulation studies is reported in the Web Appendix. In the Web Appendix, we also illustrate that the ORD and other procedures may exceed the nominal error level under certain dependence structures.

The main conclusion from the simulations is that our procedure is particularly suited for the challenging cases in which the sample size is small. The differences are particularly evident as m and k get larger. As we noted, in real data applications it is sensible to allow for larger k as m gets larger. When n is large, for fixed single inference power, ORD is outperformed by GR. More precisely, with $n = 5$ and $n = 10$ our procedure always outperforms the competitors, often markedly. With $n = 20$ and $n = 50$ the ORD procedure behaves usually more or less like LR. With $n = 20, 50$, it never outperforms GR.

The fixed single inference power setting further underlines that, as n increases, the data-driven ORD is less and less able to put false nulls at the beginning of the list. Roughly speaking, effect size is blurred by a larger sum of squares of the errors, when n is large.

We have chosen to perform simulations under a fixed single inference power to underline these features. As reported in the Web Appendix, these behaviors are connected only with the fixed power setting: if we fix the effect size and let the sample size increase, we have that for small n ORD outperforms the other procedures, while for large n all procedures have approximately the same power. Further, with fixed effect size, power is always seen to increase with n, as it is expected.

5. Multiple Sclerosis Data

Multiple sclerosis (MS) is a demyelinating disorder of the central nervous system with inflammatory and neurodegenerative components affecting about 2.5 million worldwide. In most cases, a diagnosis is made between the ages of 20 and 30. A definitive therapy is not yet available, and medications available usually are prescribed to help victims cope with pain and slow down degradation of physical, mental, and speech abilities.

No clear causative factor has yet been identified. Further, there are a variety of clinical and pathological manifestations of MS which account for a large causative heterogeneity and make harder the disclosure of the relative contribution of genetic and environmental factors for this multifactorial disease.

Studies that aim at assessing gene relationships with the disease can then be of great help, at least by increasing understanding of disease mechanisms.

At the Center for Experimental Neurological Therapy of Sant'Andrea hospital in Rome (Italy) a case-control study was designed by enrolling 13 cases who had a healthy homozygotic twin. The choice of working with twins is related to the heterogeneity expected at the individual level for MS cases, since homozygotic twins are obviously expected to be similar at the genetic level. For a discussion about the advantages of using twins for this kind of studies refer to Salvetti et al. (2000). Of course, in a study involving homozygotic twins discordant by a disease whose prevalence in Italy is about 75 per 100,000, the number of couples enrolled cannot be expected to be large.

The main goal is gene discovery, that is, forming a list of significantly differentially expressed genes for further study through polymerase chain reaction and other methods.

A two-color DNA microarray experiment was designed by using 13 separate slides onto which the mRNA from each couple was spotted. The mRNAs in each slide were labeled using a green and a red dye. Microarrays were scanned using the GenePix scanner (Axon Instruments, Inc., Union City, CA) and expression levels for each gene and subject were recorded for data analysis, together with information about the background noise.

The expression levels were normalized and then log transformed. In order to apply our approach, for each gene the response Y_j is defined as the difference between the log-transformed normalized expression levels, and the null hypothesis for each gene specifies a zero mean for the difference on the log scale. This is a simple device for transforming this two-sample paired design in an equivalent one-sample design.

More formally, we let Y_{ij} be the difference for the log expression levels of the jth gene for the ith couple and assume $Y_{ij} \sim N(\mu_j, \sigma_j^2)$. For each gene, we test the null hypothesis $H_0 : \mu_j = 0$ against a two-sided alternative. For each test, *p*-values arise from one-sample *t*-tests, and $M2_j = \sum_i y_{ij}^2$.

Figure 2 shows the $-\log_{10}$ of the *p*-values (on the *y*-axis) plot against the second moments $M2_j$ (on the *x*-axis). According to our procedure, *p*-values are compared to the one-step

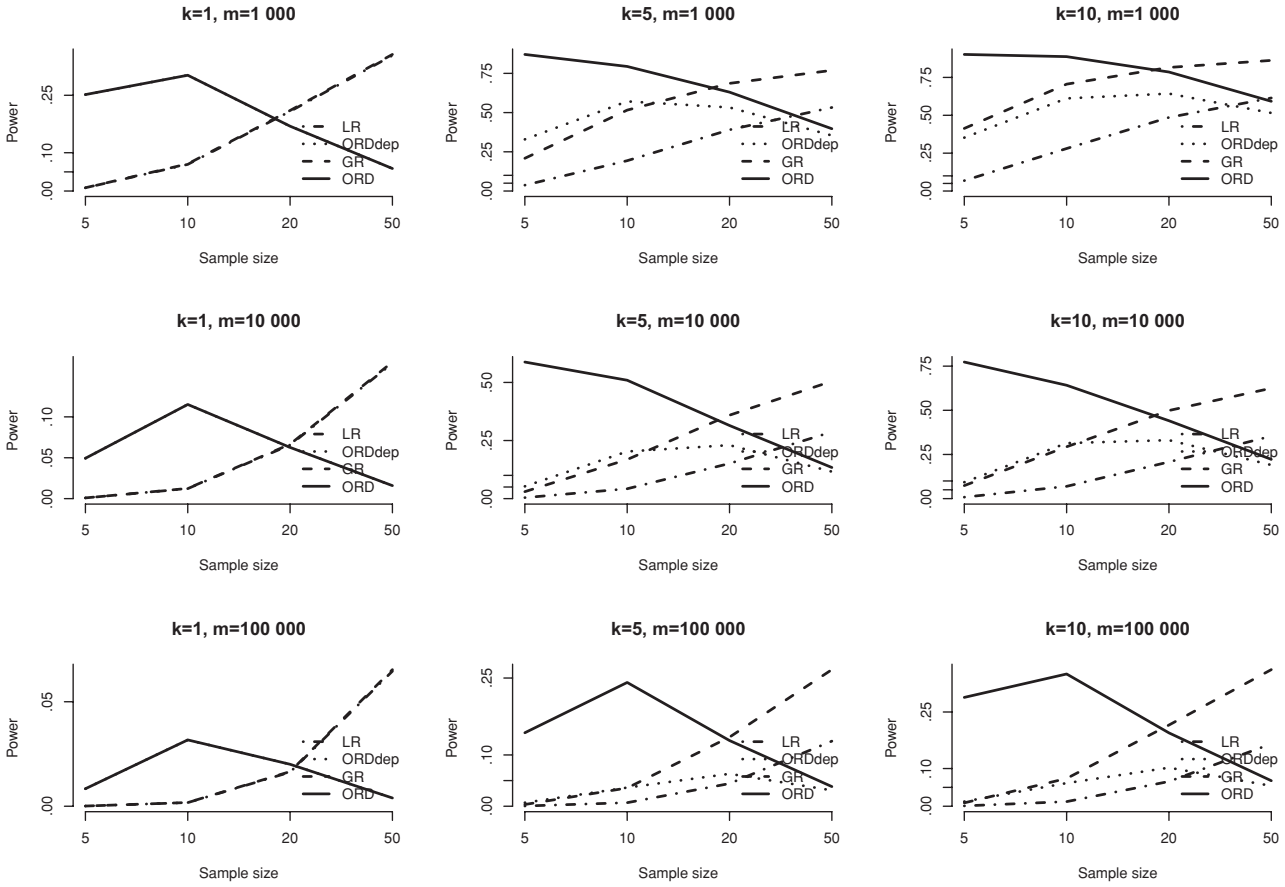

Figure 1. Proportion of correctly rejected hypotheses for different values of k, n, and m.α is set to 0.05, the proportion of false nulls is set at 10%, the power of each single test at 70% (i.e., effect size equal to 2.024, 1.325, 0.909, and 0.566 for $n = 5, 10, 20,$ and 50, respectively).

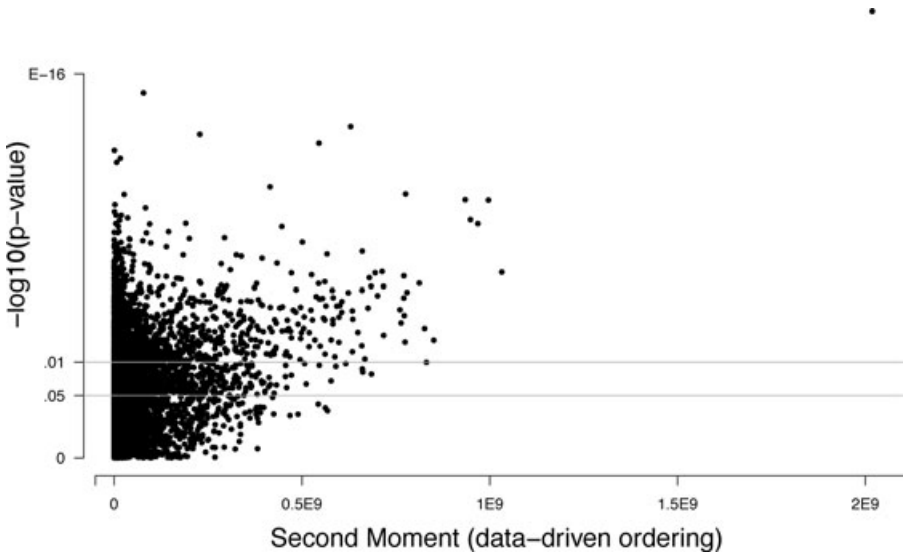

Figure 2. $-\log_{10}(p)$ for the multiple sclerosis twin data plot against the second moment for each test. The straight lines denote the $\alpha = 0.05$ and $\alpha = 0.01$ thresholds.

Table 2
Number of rejected hypotheses for Lehman and Romano (LR), Guo and Romano step-down (GR), and ordered (ORD) procedures for different α and k

	$\alpha = 0.05$				$\alpha = 0.01$				$\alpha = 0.001$			
k	LR	GR	ORD	ORD_{dep}	LR	GR	ORD	ORD_{dep}	LR	GR	ORD	ORD_{dep}
1	14	14	52	52	8	8	7	7	2	2	6	6
5	37	159	232	95	14	117	188	33	7	74	118	8
10	58	326	379	121	21	274	325	35	8	214	236	8

cutoff α starting from the rightmost and proceeding leftwards, until $J(k, \alpha)$ p-values are found above α.

Results are reported in Table 2 for Lehman and Romano procedure (LR), Guo and Romano step-down approach (GR); our ordered procedure with the data-driven ORD and its version extended for arbitrary dependence (ORD_{dep}), for different α and k.

As expected, the number of selected genes decreases with α for all the procedures. Only in one case ($k = 1, \alpha = 0.01$) the ORD procedure selects a lower number of genes than its competitors, while the number of selected genes is much higher (but still reasonable for further screening) in many settings, suggesting a possibly higher power for the ORD procedure for the data at hand. The extension under general dependence is as expected more conservative, but always rejecting a higher number of hypotheses than the LR procedure.

A higher number of selected genes reduces the odds of exclusion of important genes for further investigation, and as already noted our data-driven criterion further enhances selection of genes with larger effect sizes, which are put at the beginning of the list even if they may have larger p-values.

In this application, $k = 5$ has been a priori chosen by the geneticists, with the help of considerations related to expected number of true effects and feasibility of further screening. It is important to underline that k should be set a priori, in order to avoid data snooping. Using a larger k results in greater power; however, this increase in power is accompanied by an increased number of false positives. As we noted in Section 2.1, further, the number of allowed false positives k shall be set smaller when the number of tests is smaller. A different route for instance is taken in Chen and Storey (2006), who suggest to specify a vector of reasonable levels for different values of k, with higher levels prescribed for smaller values of k.

6. Supplementary Materials

The Web Appendix referenced in Sections 3 and 4 is available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

The authors are grateful to three referees for detailed comments which led to substantial improvements, to the AE for help in clarifying the proof of Theorem 1, and to Department of Neurology and Center for Experimental Neurological Therapy of Sant'Andrea hospital, Sapienza—University of Rome, for permission to use the twin data. This work was supported by Biorange grant SP 1.3.2 of the Netherlands Bioinformatics Center (NBIC).

References

Chen, L. and Storey, J. (2006). Relaxed significance criteria for linkage analysis. *Genetics* **173,** 2371–2381.

Clarke, S. and Hall, P. (2009). Robustness of multiple testing procedures against dependence. *Annals of Statistics* **37,** 332–358.

Conover, W. and Iman, R. (1982). Analysis of covariance using the rank transformation. *Biometrics* **38,** 715–724.

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

Fang, K. and Zhang, Y. (1990). *Generalized Multivariate Analysis*. Beijing: Science Press.

Farcomeni, A. (2007). Some results on the control of the false discovery rate under dependence. *Scandinavian Journal of Statistics* **34,** 275–297.

Farcomeni, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research* **17,** 347–388.

Finos, L. and Salmaso, L. (2006). Weighted methods controlling the multiplicity when the number of variables is much higher than the number of observations. *Journal of Nonparametric Statistics* **18,** 245–261.

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**.

Hochberg, Y. and Tamhane, A. (1987). *Multiple Comparisons Procedures*. New York: Wiley.

Hommel, G. and Kropf, S. (2005). Tests for differentiation in gene expression using a data-driven order or weights for hypotheses. *Biometrical Journal* **47,** 554–562.

Hsu, J. and Berger, R. (1999). Stepwise confidence intervals without multiplicity adjustment for dose-response and toxicity studies. *Journal of the American Statistical Association* **94,** 468–475.

Johnson, N. L., Kemp, A. W., and Kotz, S. (2005). *Discrete Univariate Distributions,* 3rd edition. New York: Wiley-Interscience.

Kirk, R. (2007). Effect magnitude: A different focus. *Journal of Statistical Planning and Inference* **137,** 1634–1646.

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.

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. *Journal of Statistical Planning and Inference* **125,** 31–47.

Lauter, J., Glimm, E., and Kropf, S. (1998). Multivariate tests based on left-spherically distributed linear scores. *Annals of Statistics* **26,** 1972–1988.

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

Marcus, R., Peritz, E., and Gabriel, K. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63,** 655–660.

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

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Romano, J. and Wolf, M. (2007). Control of generalized error rates in multiple testing. *Annals of Statistics* **35,** 1378–1408.

Rosenbaum, P. (2008). Testing hypotheses in order. *Biometrika* **95,** 248–252.

Salvetti, M., Ristori, G., Bomprezzi, R., Pozzilli, P., and Leslie, R. (2000). Twins: Mirrors of the immune system. *Immunology Today* **21,** 342–347.

Sarkar, S. (2008). Generalizing Simes' test and Hochberg's stepup procedures. *Annals of Statistics* **36,** 337–363.

Storey, J. and Tibshirani, R. (2003). Statistical significance for genome-wide experiments. *Proceedings of the National Academy of Sciences* **100,** 9440–9445.

Strassburger, K., Bretz, F., and Finner, H. (2007). Ordered multiple comparisons with the best and their applications to dose-response studies. *Biometrics* **63,** 1143–1151.

van der Laan, M., Dudoit, S., and Pollard, K. (2004). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology* **3**.

Westfall, P., Kropf, S., and Finos, L. (2004). Weighted fwe-controlling methods in high-dimensional situations. In *Recent Developments in Multiple Comparison Procedures*, *Institute of Mathematical Statistics Lecture Notes-Monograph Series Vol. 47,* Y. Benjamini, F. Bretz, and S. Sarkar (eds), 143–154. Beachwood, Ohio: Institute of Mathematical Statistics.

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

Received April 2009. *Revised March* 2010.
Accepted March 2010.