Working Paper Series, N. 7, May 2009

k-FWER control without multiplicity correction, with application to detection of genetic determinants of multiple sclerosis in Italian twins

L. Finos
Department of Statistical Sciences
University of Padua
Italy
-
Department of Medical Statistics and Bioinformatics
Leiden University Medical Center
The Netherlands

A. Farcomeni
Dipartimento di Medicina Sperimentale
Sapienza - University of Rome

Abstract: 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 non parametric setting. The approach is not only very simple and computationally light, but also the data-driven order enhances power when the sample size is small (and also when k and/or m 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.

Keywords: Data driven order; Gene discovery; k-Familywise Error Rate; Multiple Sclerosis; Multiple testing.

Submitted 19-04-2009

k-FWER control without multiplicity correction, with application to detection of genetic determinants of multiple sclerosis in Italian twins

Contents

Department of Statistical Sciences
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

Corresponding author:
Livio Finos
livio.finos@unipd.it

k-FWER control without multiplicity correction, with application to detection of genetic determinants of multiple sclerosis in Italian twins

L. Finos
Department of Statistical Sciences
University of Padua
Italy
-
Department of Medical Statistics and Bioinformatics
Leiden University Medical Center
The Netherlands

A. Farcomeni
Dipartimento di Medicina Sperimentale
Sapienza - University of Rome

Abstract: 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 non parametric setting. The approach is not only very simple and computationally light, but also the data-driven order enhances power when the sample size is small (and also when k and/or m 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.

Keywords: 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 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 assigned to the ill twin and a green dye to the safe 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 poweful 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 have 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 lead 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 *et al.* (2003), to the books by Westfall and Young (1993) and by 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 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 *et al.* (2004) show 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 which dramatically improve power under independence of the test statistics and Romano and Wolf (2007) show methods that can be used also under dependence, which anyway 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.

Goal of this paper 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 lack of information. We anticipate the power of our procedure will be enhanced in cases of approximate homoschedasticity 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 cut-off 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, regression. One can adjust the p-values for confounders and non-parametric approaches can be accomodated 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 *et al.* (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. We also give an extension of the procedure which does not rely on any assumption concerning the dependency structure. R (R Development Core Team, 2007) code for the proposed procedures is available from the authors upon request.

The rest of the paper is as follows: in Section 2 we show 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

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 few other cases. See for instance Marcus *et al.* (1976); Hsu and Berger (1999); Maurer *et al.* (1995); Strassburger *et al.* (2007).

The k-FWER can be controlled by performing tests sequentially at the uncorrected level α. Sequential testing is stopped after 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 summary, denoting with $p_{(1)}, \ldots, p_{(m)}$ the m p-values ordered with respect to the natural ordering of the m hypotheses, we propose the following *sequential procedure*:

Algorithm: Sequential procedure for naturally ordered hypotheses

- Let

$$J(k,\alpha) = \max\{J : \sum_{j=k}^{J+k-1} \binom{J+k-1}{j} \alpha^j (1-\alpha)^{J+k-1-j} \leq \alpha\} \qquad (1)$$

- Set $j := 0$, $i := 1$

```
WHILE($i < m$ & $j < J(k,\alpha)$) $i := i + 1$
IF($p_{(i)} \geq \alpha$) $j := j + 1$ ENDIF
ENDWHILE
```

- Reject all the 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.

Unlike many other approaches, the proposed sequential procedure does not correct 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. *Call $\mathcal{S}_J$ the set of hypotheses tested before the J^{th} unrejected hypothesis. Let $\mathcal{S}_J^0 \subseteq \mathcal{S}_J$ be the set of true null hypotheses tested before the J^{th} unrejected hypothesis. Let $\phi_i = 1$ $(i \in 1,\dots,m)$ if the i-th 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 are stochastically dominated by 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 J^{th} jump is bounded by the survival function of a negative binomial random variable:

$$P(\sum_{i \in \mathcal{S}_J^0} \phi_i \geq k) \leq 1 - F_{\mathcal{B}neg(J,1-\alpha)}(k-1) \quad \forall (k,\alpha), \forall \mathcal{S}_J^0 \subseteq \mathcal{H}_0$$

ii The sequential procedure 2.1 with $J(k,\alpha)$ defined as in (1) controls the k-FWER at level α.

Proof. For fixed J, the k-FWER of procedure 2.1 is given by $P(\sum_{i \in \mathcal{S}_J^0} \phi_i \geq k)$, since all hypotheses corresponding to the other p-values are not rejected by definition.

Note that $P(\phi_i = 1) \leq \alpha$ for all $i \in \mathcal{H}0$. Hence

$$P(\sum_{i \in \mathcal{S}_J^0} \phi_i \geq k) \leq 1 - F_{\mathcal{B}neg(J,1-\alpha)}(k-1). \qquad (2)$$

To fix the ideas, suppose all null hypotheses are true. Each hypothesis is not rejected with probability at least $1-\alpha$. When the procedure is stopped, the number of type I errors before the J^{th} not rejected hypothesis is then a negative binomial with parameters J and probability of success in each trial (i.e. test) equal to $1-\alpha$. When some of the null hypotheses are false, the inequality holds because of the hypothesis of stochastic domination. The inequality (2) is strict unless $\mathcal{S}_J^0 = \mathcal{S}_J$ and the tests are exactly α-size. This proves the first part.

To prove the second part of the theorem, we can use the properties of Negative Binomial random variables to show that:

$$P(\sum_{i \in \mathcal{S}^0_{J(k,\alpha)}} \phi_i \geq k) \leq 1 - F_{\mathcal{B}neg(J(k,\alpha),1-\alpha)}(k-1) = 1 - F_{\mathcal{B}(J(k,\alpha)+k-1,\alpha)}(k-1),$$

where $\mathcal{B}(n,\pi)$ denotes binomial random variable. It is straightforward to see that $1 - F_{\mathcal{B}(J(k,\alpha)+k-1,\alpha)}(k-1) = \alpha$ by substituting the expression for (1). □

The idea behind the proof is that the number of failures (false rejections) which occur in a sequence of Bernoulli trials before J successes (i.e. J^{th} unrejected hypothesis) is reached can be seen as a negative binomial random variable. The rest of the proof follows from straightforward algebra.

Computation of $J(k,\alpha)$ through (1) relies only on evaluation of the upper tail of binomial distributions. It is straightforward to check that $J(1,\alpha) = 1$ and then when controlling the classical FWER we get back the Maurer *et al.* (1995) procedure. In the k-FWER control setting this 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
$\alpha = .10$	1	4	9	15	21	27	34	40	47	54	128
$\alpha = .05$	1	6	14	25	36	48	61	74	87	101	249
$\alpha = .01$	1	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 $\alpha = .05$ and $k = 10$. One can reject all hypotheses below $\alpha = .05$ (and in fact $J(10,.05) = 101$ jumps are allowed) since the probability of having ten or more false positives would be lower than or equal to .011. Note also that in presence of many true null hypotheses the sequential testing will be stopped very early *regardless* of m.

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 surprising feature is that $J(k, \alpha)$ is decreasing in α. One should anyway consider that even if the number of allowed jumps in the sequence increases when controlling more strictly the k-FWER, the probability that a single p-value is below α rapidly decreases. Hence, the number of jumps in a fixed-length sequence will be much higher for lower α.

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 2.1.

In the following, we assume that each p-value arises from a test on linear hypotheses on the parameters of the model:

$$Y_j = Z_j \beta_j + \epsilon_j, \tag{3}$$

where Y_j is a numerical response, Z_j is a fixed matrix of covariates (which may include dummy variables and/or a constant column), β_j is a vector of parameters and ϵ_j is distributed like a zero-centered Gaussian with variance σ_j^2. 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 (3), estimated constraining the parameters under the null hypothesis.

The idea is easily understood if one thinks about the one-sample t-test for a zero mean, in which

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

with $\bar{y}_j = \sum_i y_{ij}/n$, $\hat{\sigma_j}^2$ the (biased) estimated variance and $\hat{\delta_j}$ 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 to be comparing two independent samples. In that case, Z is defined as a two column matrix with a column of ones and a column which containts 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 the second moment $E(Y_j^2)$ can be expressed as:

$$E(Y_j^2) = E((Z\beta_j + \epsilon_j)^2) = \|Z\beta_j^2\| + Var(\epsilon_j).$$

($\| \cdot \|$ 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).

A proof that the ordering according to $M2_j$ does not inflate the k-FWER is based on the theory of sphericall distributed matrices (Fang and Zhang, 1990), and is a direct extension of Lauter *et al.* (1998), Theorem 1.

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, thus producing a list of rejected hypotheses potentially more interesting for the practitioner (see Kirk (2007), and references therein).

The non-parametric setting can be accomodated in three different ways. First, one can simply use the rank-transformation of Conover and Iman (1982). Secondly, one can compute p-values from non-parametric rank based methods, and order the hypotheses according to (possibly adjusted) medians in case of one sample tests and to (possibly adjusted) interquartile ranges in case of $C \geq 2$ 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 which 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 for dependent test statistics

The m test statistics are in general not independent. Nevertheless, in many situations multiple testing procedures devised for independent test statistics can be used under dependence according to the results of the recent breakthrough paper of Clarke and Hall (2009). Formally, they show that if the distributions of the test statistics under the null hypotheses are not heavy-tailed and dependence does not increase with the number of tests, procedures devised for the independence case are asymptotically valid also under dependence. Their conditions apply to many real data applications. For instance, block (sometimes called “clumpy”) dependence is usually expected in microarray experiments, like the one we discuss in Section 5, as argued for instance in Storey and Tibshirani (2003). Further their conditions apply to situations in which weak dependence is expected, as the applications discussed in Farcomeni (2007). In summary, according to the results of Clarke and Hall (2009), our procedure (together with the procedures in Guo and Romano (2007) and other procedures devised for independent test statistics) can be used in many real situations when the number of tests m is large, even if the dependence among the test statistics is strong.

For cases in which the conditions of Clarke and Hall (2009) may not be met, 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 $\alpha' = \frac{k\alpha}{J(k,\alpha)+k-1}$, obtaining a slightly more conservative procedure which anyway is valid under general dependence. A proof of this statement given in next theorem:

Theorem 2. *Assume the distributions of the p-values under the null hypotheses are stochastically dominated by the uniform, and the distributions of the p-values under the alternative are stochastically dominated by the null distribution(s). Let* $\alpha' = \frac{k\alpha}{J(k,\alpha)+k-1}$*. Under general dependence among the test statistics we have that the sequential procedure 2.1 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. The proof relies on the Markov's inequality:

$$P(\sum_{i=1}^{J(k,\alpha)+k-1} \phi_i \geq k) \leq \frac{E[\sum_{i=1}^{J(k,\alpha)+k-1} \phi_i]}{k} = \frac{(J(k,\alpha)+k-1)\alpha'}{k} = \alpha.$$

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

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 $n = 5, 10, 20, 50$; $m = 500, 1000, 10000$; $\alpha = .01, .05$. 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%.

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). We repeat the operation $B = 1000$ times and record the average fraction of correctly rejected hypotheses as a measure of power, and the k-FWER (that is, the fraction of datasets with a number of false rejections larger than or equal to k) in order to check that it is below the nominal error rate level α. The results for different values of k are reported in tables 2, 3 and 4 for $\alpha = .05$ and respectively $m = 500, 1000, 10000$. In tables 5, 6 and 7 we report the results for respectively $m = 500, 1000, 10000$ when $\alpha = .01$. Note that the case $k = 1$ is reported only for reference, since control of the 1-FWER is not the main focus of this paper.

The main conclusion from the simulations is that the 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. In other settings, the k-FWER control with data-driven order of the hypotheses behaves approximately like LR,

Table 2: Average proportion of correctly rejected hypotheses (and k-FWER in parentheses) for different values of k and n, with $m = 500$ and $\alpha = .05$. The proportion of false nulls is set at 10% and the power of each single test at 70%.

	$n = 5$			$n = 10$			$n = 20$			$n = 50$		
k	LR	GR	ORD	LR	GR	ORD	LR	GR	ORD	LR	GR	ORD
1	.011	.011	.216	.066	.067	.229	.170	.172	.130	.273	.275	.049
	(.044)	(.046)	(.006)	(.044)	(.045)	(.033)	(.047)	(.047)	(.042)	(.042)	(.043)	(.052)
3	.030	.130	.710	.135	.348	.601	.274	.503	.435	.385	.595	.248
	(.001)	(.040)	(.019)	(.000)	(.039)	(.038)	(.000)	(.044)	(.037)	(.000)	(.044)	(.039)
5	.047	.253	.790	.185	.506	.732	.335	.638	.594	.445	.705	.401
	(.000)	(.034)	(.027)	(.000)	(.041)	(.040)	(.000)	(.045)	(.044)	(.000)	(.046)	(.038)
8	.071	.403	.823	.241	.641	.815	.397	.740	.723	.502	.787	.556
	(.000)	(.035)	(.037)	(.000)	(.041)	(.042)	(.000)	(.045)	(.047)	(.000)	(.047)	(.051)
10	.086	.481	.832	.270	.699	.845	.428	.781	.775	.529	.819	.631
	(.000)	(.037)	(.037)	(.000)	(.043)	(.042)	(.000)	(.040)	(.048)	(.000)	(.045)	(.050)

Table 3: Average proportion of correctly rejected hypotheses (and k-FWER in parentheses) for different values of k and n, with $m = 1000$ and $\alpha = .05$. The proportion of false nulls is set at 10% and the power of each single test at 70%.

	$n = 5$			$n = 10$			$n = 20$			$n = 50$		
k	LR	GR	ORD	LR	GR	ORD	LR	GR	ORD	LR	GR	ORD
1	.005	.006	.138	.040	.041	.169	.120	.121	.093	.213	.215	.032
	(.039)	(.040)	(.002)	(.044)	(.045)	(.022)	(.046)	(.047)	(.044)	(.042)	(.043)	(.046)
3	.016	.072	.617	.087	.245	.503	.204	.403	.337	.311	.506	.170
	(.000)	(.038)	(.007)	(.000)	(.038)	(.029)	(.001)	(.042)	(.038)	(.000)	(.045)	(.036)
5	.025	.150	.737	.122	.380	.640	.255	.533	.477	.365	.618	.284
	(.000)	(.032)	(.014)	(.000)	(.039)	(.036)	(.000)	(.042)	(.045)	(.000)	(.042)	(.045)
8	.038	.254	.791	.163	.509	.737	.308	.641	.600	.418	.707	.408
	(.000)	(.035)	(.024)	(.000)	(.036)	(.033)	(.000)	(.039)	(.047)	(.000)	(.048)	(.046)
10	.046	.315	.807	.185	.569	.774	.336	.687	.655	.444	.744	.472
	(.000)	(.030)	(.025)	(.000)	(.038)	(.037)	(.000)	(.043)	(.049)	(.000)	(.049)	(.040)

and sometimes slightly worse than 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 $\alpha = .05$ and $n = 20, 50$ it never outperforms GR, while with $\alpha = .01$ it compares better and behaves more or less like GR with large k and smaller m.

A somewhat surprising behaviour of our ORD procedure is that power is sometimes seen to decrease for larger n. Recall anyway that single-inference power has been fixed to 70%. This has been done in order to make this behaviour evident. This happens because of the decreasing capability of the data-driven ordering of putting false nulls at the beginning of the list as n increases. Roughly speaking, effect size is blurred by a larger sum of squares of the errors when n is larger. If the mean under the alternative were left fixed as n increased, as expected, the proportion of correctly

Table 4: Average proportion of correctly rejected hypotheses (and k-FWER in parentheses) for different values of k and n, with $m = 10000$ and $\alpha = .05$. The proportion of false nulls is set at 10% and the power of each single test at 70%.

	$n = 5$			$n = 10$			$n = 20$			$n = 50$		
k	LR	GR	ORD	LR	GR	ORD	LR	GR	ORD	LR	GR	ORD
1	.001	.001	.025	.006	.007	.050	.032	.033	.028	.084	.085	.007
	(.044)	(.046)	(.000)	(.047)	(.048)	(.016)	(.049)	(.050)	(.037)	(.040)	(.042)	(.045)
3	.002	.009	.196	.016	.057	.223	.062	.153	.123	.134	.255	.044
	(.000)	(.037)	(.000)	(.000)	(.039)	(.010)	(.000)	(.039)	(.032)	(.000)	(.045)	(.041)
5	.003	.020	.360	.024	.105	.326	.083	.230	.191	.165	.340	.079
	(.000)	(.033)	(.000)	(.000)	(.037)	(.013)	(.000)	(.042)	(.031)	(.000)	(.041)	(.041)
8	.004	.039	.510	.034	.163	.418	.107	.308	.261	.197	.419	.120
	(.000)	(.029)	(.000)	(.000)	(.029)	(.017)	(.000)	(.037)	(.034)	(.000)	(.039)	(.044)
10	.006	.051	.569	.040	.195	.461	.120	.347	.297	.214	.456	.144
	(.000)	(.028)	(.000)	(.000)	(.033)	(.017)	(.000)	(.038)	(.034)	(.000)	(.040)	(.041)

Table 5: Average proportion of correctly rejected hypotheses (and k-FWER in parentheses) for different values of k and n, with $m = 500$ and $\alpha = .01$. The proportion of false nulls is set at 10% and the power of each single test at 70%.

	$n = 5$			$n = 10$			$n = 20$			$n = 50$		
k	LR	GR	ORD	LR	GR	ORD	LR	GR	ORD	LR	GR	ORD
1	.031	.007	.112	.199	.073	.356	.430	.240	.314	.596	.420	.132
	(.043)	(.011)	(.000)	(.049)	(.012)	(.003)	(.047)	(.009)	(.004)	(.046)	(.011)	(.010)
3	.083	.196	.750	.351	.540	.882	.584	.731	.833	.715	.817	.642
	(.001)	(.011)	(.004)	(.001)	(.009)	(.007)	(.000)	(.009)	(.008)	(.001)	(.011)	(.010)
5	.126	.409	.751	.437	.737	.903	.655	.853	.916	.766	.898	.836
	(.000)	(.008)	(.006)	(.000)	(.010)	(.009)	(.000)	(.010)	(.010)	(.000)	(.008)	(.009)
8	.182	.627	.751	.522	.858	.908	.718	.920	.942	.808	.941	.938
	(.000)	(.009)	(.010)	(.000)	(.007)	(.012)	(.000)	(.008)	(.010)	(.000)	(.011)	(.009)
10	.215	.720	.751	.564	.898	.908	.746	.941	.945	.828	.955	.957
	(.000)	(.010)	(.010)	(.000)	(.010)	(.010)	(.000)	(.010)	(.010)	(.000)	(.010)	(.009)

rejected hypotheses by the ORD procedure would have been non-decreasing.

The same results are obtained in other simulations settings, also under dependence, which we do not show for reasons of space. In the simulations shown we have set a prescribed power for each single test at 70%. In cases in which the single-inference power is set lower our procedure compares much better also in settings with a larger number of observations.

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 world-

Table 6: Average proportion of correctly rejected hypotheses (and k-FWER in parentheses) for different values of k and n, with $m = 1000$ and $\alpha = .01$. The proportion of false nulls is set at 10% and the power of each single test at 70%.

	$n = 5$			$n = 10$			$n = 20$			$n = 50$		
k	LR	GR	ORD	LR	GR	ORD	LR	GR	ORD	LR	GR	ORD
1	.017	.003	.063	.131	.044	.245	.343	.179	.243	.518	.351	.096
	(.046)	(.009)	(.000)	(.045)	(.009)	(.001)	(.045)	(.012)	(.007)	(.042)	(.009)	(.010)
3	.045	.113	.745	.247	.412	.852	.487	.639	.757	.639	.752	.519
	(.000)	(.005)	(.002)	(.000)	(.007)	(.007)	(.001)	(.011)	(.007)	(.000)	(.008)	(.009)
5	.070	.259	.749	.319	.611	.892	.560	.778	.869	.695	.848	.716
	(.000)	(.005)	(.004)	(.000)	(.009)	(.007)	(.000)	(.010)	(.006)	(.000)	(.011)	(.010)
8	.105	.441	.750	.396	.757	.904	.626	.864	.922	.743	.905	.856
	(.000)	(.007)	(.003)	(.000)	(.010)	(.008)	(.000)	(.008)	(.007)	(.000)	(.011)	(.009)
10	.125	.533	.750	.435	.812	.906	.656	.894	.935	.765	.924	.903
	(.000)	(.007)	(.007)	(.000)	(.009)	(.010)	(.000)	(.008)	(.010)	(.000)	(.012)	(.010)

Table 7: Average proportion of correctly rejected hypotheses (and k-FWER in parentheses) for different values of k and n, with $m = 10000$ and $\alpha = .01$. The proportion of false nulls is set at 10% and the power of each single test at 70%.

	$n = 5$			$n = 10$			$n = 20$			$n = 50$		
k	LR	GR	ORD	LR	GR	ORD	LR	GR	ORD	LR	GR	ORD
1	.002	.000	.009	.026	.007	.054	.129	.056	.083	.288	.170	.030
	(.046)	(.009)	(.000)	(.038)	(.007)	(.000)	(.045)	(.010)	(.002)	(.043)	(.007)	(.009)
3	.005	.015	.221	.059	.121	.583	.213	.325	.455	.391	.504	.213
	(.000)	(.006)	(.000)	(.000)	(.007)	(.000)	(.000)	(.009)	(.003)	(.001)	(.010)	(.007)
5	.009	.039	.500	.084	.228	.756	.263	.465	.603	.444	.624	.336
	(.000)	(.007)	(.000)	(.000)	(.007)	(.002)	(.000)	(.008)	(.004)	(.000)	(.009)	(.009)
8	.013	.081	.735	.114	.345	.830	.315	.581	.711	.495	.712	.457
	(.000)	(.007)	(.000)	(.000)	(.008)	(.003)	(.000)	(.007)	(.006)	(.000)	(.007)	(.008)
10	.016	.109	.745	.131	.404	.852	.342	.631	.754	.519	.748	.516
	(.000)	(.006)	(.000)	(.000)	(.008)	(.002)	(.000)	(.010)	(.005)	(.000)	(.007)	(.008)

wide. In most cases, a diagnosis is made between the ages of 20 and 3. 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 Center for Experimental Neurological Therapy of Sant'Andrea hospital in Rome (Italy) a case-control study was designed by enrolling thirteen cases who had an homozygotic twin safe from the disease. 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 the 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 100000, the number of couples enrolled can not be expected to be large.

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-colour DNA microarray experiment was designed by using thirteen 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, subject, and slide 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 j-th gene for the i-th couple; and assume $Y_{ij} \sim N(\mu_j, \sigma_j^2)$. For each gene, we test the null hypothesis $H_0 : \mu_j = 0$. For each test, p-values arise from one-sample t-tests, and $M2_j = \sum_i y_{ij}^2$.

Figure 1 shows 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 cut-off α staring from the rightmost and proceeding leftwards, until $J(k, \alpha)$ p-values are found above α.

Results are reported in Table 8 for Lehman and Romano procedure (LR), Guo and Romano step-down approach (GR) and our ordered procedure with the data-driven order of hypotheses (ORD), for different α and k.

Table 8: Number of rejected hypotheses for Lehman and Romano (LR), Guo and Romano step-down (GR) and ordered (ORD) procedures for different α and k. In parentheses, $J(k, \alpha)$ for the ORD procedure.

	$\alpha = .05$			$\alpha = .01$			$\alpha = .001$		
k	LR	GR	ORD($J(k, \alpha)$)	LR	GR	ORD($J(k, \alpha)$)	LR	GR	ORD($J(k, \alpha)$)
1	14	14	52(1)	8	8	7(1)	2	2	6(1)
5	37	159	232(36)	14	117	188(125)	7	74	118(737)
10	58	326	379(101)	21	274	325(406)	8	214	236(2956)

As expected, the number of selected genes decreases with α for all the procedures. Only in one case ($k = 1, \alpha = .01$) the ORD procedure selects a lower number of genes than its competitors, while the number of selected genes is much higher (but

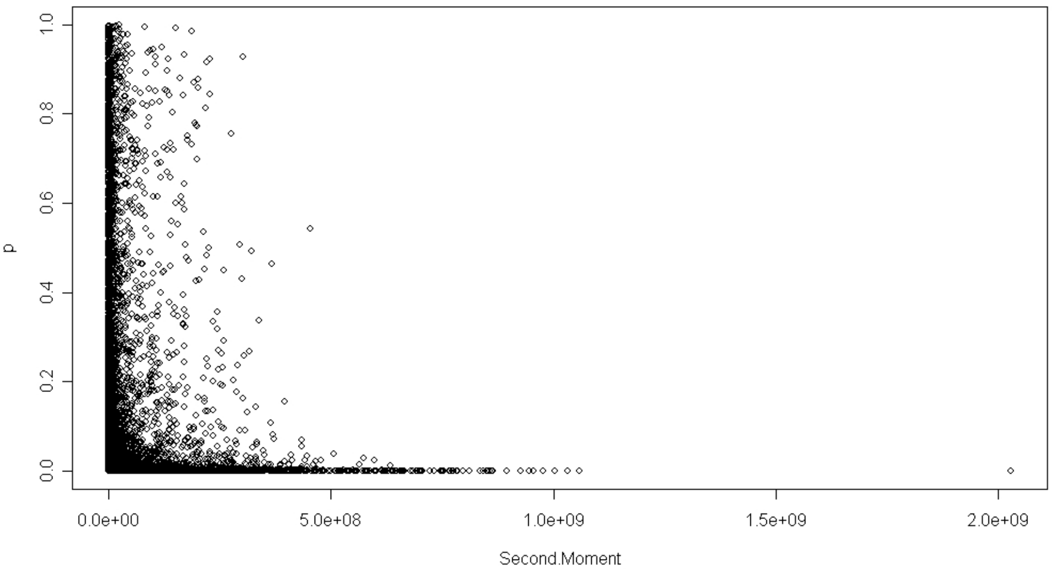

Figure 1: p-values for the multiple sclerosis twin data plot against the second moment for each test.

still reasonable for further screening) in many settings, suggesting a possibly higher power for the ORD procedure for the data at hand. An 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.

References

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

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

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

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

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

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

L. Finos and L. Salmaso (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.

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

Y. Hochberg and A.C. Tamhane (1987). *Multiple Comparisons Procedures*. Wiley.

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

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

S. Kropf and J. Läuter (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.

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

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

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

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

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

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

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

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

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

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

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

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

M.J. van der Laan, S. Dudoit, and K.S. Pollard (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, 1**.

P. H. Westfall and S. S. Young (1993). *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*. Wiley.

Acknowledgements

The authors are grateful to Department of Neurology and Center for Experimental Neurological Therapy of SantAndrea hospital, Sapienza - University of Rome, for permission to use the twin data.