Generalized Augmentation for Control of the k-Familywise Error Rate

Alessio Farcomeni*

Department of Public Health and Infectious Diseases, Sapienza - University of Rome, Italy

Abstract: When performing many hypothesis tests at once a correction for multiplicity is needed to both keep under control the number of false discoveries and be able to detect the true departures from the null hypotheses. A recently introduced method which has been proved to be useful in genomics, neuroimaging and other fields consists in probabilistically controlling that the number of falsely rejected hypotheses does not exceed a pre-specified (low) k. We introduce a new multiple testing procedure which is based on the idea of generalized augmentation: at first a number of hypotheses is rejected without any correction, then this number is adjusted by adding or removing rejections. The procedure is shown to keep under control the probability of k or more false rejections. We show a small simulation study which suggests that the new procedure is very powerful, especially when the number of tests at stake is large. We conclude with an illustration on a benchmark data set on classification of colon cancer.

1. INTRODUCTION

Consider the problem of simultaneously testing a finite number of null hypotheses. A simple approach would be to test each hypothesis at level α, without taking into account the multiplicity. However, when the number of true hypotheses is large, we shall be nearly certain that there will be false positives in the list of rejections. Under reasonable assumptions it actually happens that the number of false positives is increasing with the number of tests.

A possible solution would be to design a simoultaneous inference procedure leading to a list of rejections with no false positives with high probability. This is known as control of the FamilyWise Error Rate (FWER), to be more rigorously defined below, and can be simply achieved with the famous Bonferroni correction (that is, testing each hypothesis at level α / m, where m is the number of tests). There of course are available many other corrections controlling the FWER.

In analogy with the concept of power for a single test, one must also take into account the ability of multiple testing procedures to detect departures from the null hypotheses at population level. When the number of tests is large, control of the FWER becomes so stringent that a very low number of hypotheses (if any) are rejected, no matter the true departure from the null hypothesis. It is easily seen for instance that the

Bonferroni correction leads to an infinitesimal level (with m).

One of the aims of recent developments in multiple testing has been to overcome this problem. Mainly two routes have been followed. First, other error measures besides the FWER have been defined and motivated for applications. Secondly, for a fixed Type I error rate, multiple testing corrections which are possibly more and more powerful (at least in special situations) have been developed.

A simple generalization of the FWER, denoted as k-FWER, dates back at least to [1], and has been recently the focus of few papers, like [2-5] and others.

The aim of this paper is to and develop a multiple testing procedure which controls the k-FWER and is powerful, especially in the difficult situation in which the number of tests is large. We will do so by building on the results of [6], who has recently introduced the idea of Generalized AUGmEntation (GAUGE) in multiple testing: at first, a certain number of hypotheses is rejected with a single-inference procedure; then this number is adjusted (augmented or negatively augmentated) in order to achieve control of the desired Type I error rate.

For a deep discussion of the main ideas in the area of multiple testing refer to [7], and to [8] for a review of recent developments. In the following section we will briefly recall the necessary background. The rest of the paper is as follows: Section 2 will introduce the new k-FWER controlling procedure based on the assumption of independence among the test statistics, and briefly discuss the case of dependent test statistics. In Section

*Address corresponding to this author at the Department of Public Health and Infectious Diseases, Sapienza - University of Rome, Italy;
Tel: 00390649639263; E-mail: alessio.farcomeni@uniroma1.it

3 we provide a simulation study. In Section 4 we show an application to a benchmark data set on DNA microarrays. In Section 5 we give some concluding remarks.

1.1. Background

Table **1** summarizes the outcome of a multiple testing procedure in which m tests are being performed. We denote with $H_0(j)$ the null hypothesis associated with the j-th test, with $H_1(j)$ the corresponding alternative. M_0 of the m nulls are true, M_1 are false; R of the m null hypotheses are rejected. With $N_{1|0}$ and $N_{1|1}$ we denote respectively the number of false and true positives. Similarly, $N_{0|1}$ and $N_{0|0}$ denote the number of false and true negatives.

Table 1: Outcome in Testing m Hypotheses

	Null Hypotheses		
	Not Rejected	**Rejected**	**Total**
True	$N_{0|0}$	$N_{1|0}$	M_0
False	$N_{0|1}$	$N_{1|1}$	M_1
Total	$m - R$	R	m

Tests are performed through the use of *test statistics* $\mathbf{T} = (T_{n_j}(j) : j = 1,\dots,m)$. Test statistics are defined on the basis of the problem and data at hand, and so that higher values indicate a larger discrepancy between the data and the null hypothesis. Commonly used test statistics include t-statistics for testing on the mean of a single population or comparing the means of two populations, F-statistics for comparing the means of three or more populations; or the nonparametric rank-based counterparts like the Mann-Whitney and Kruskal-Wallis test statistics. See for instance [9] for a discussion and an extensive simulation on the use of parametric and non-parametric tests in multiple testing. Of course it is not uncommon to use other test statistics, like chi-squares, log-ranks, log-odds and so on.

We assume for each test a p-value has been computed (possibly by resampling). Assuming two-sided alternatives, the j-th p-value is defined as

$$p_j = Pr(|T_{n_j}(j)| > |t_{n_j}(j)| \,|\, H_0(j)), \qquad (1)$$

where $t_n(j)$ is the observed value of the test statistic $T_n(j)$.

The probability in equation (1) is then computed based on the asymptotic distribution under the null hypothesis of each statistic, which is known or in many cases approximately normal for large samples. In other cases, a permutation or bootstrap can be applied and the p-values estimated by resampling.

Throughout we adopt the notation $p_{(j)}$ to denote the j-th ordered statistic of the vector of p-values, with $p_{(0)} = 0$ and $p_{(m+1)} = 1$. After the p-values are computed, the multiple testing problem reduces to fixing a cut-off T which is theoretically proved to bring a pre-specified Type I error rate below a certain α for any configuration of true and false hypotheses, when rejecting all the hypotheses corresponding to $p_j \leq T$.

While traditional methods involve control of the FWER, defined as the probability of making one or more false rejections, there are now many other possible Type I error rates which can be chosen. Generalizations have been proposed among others in [3, 2, 10, 11]. Refer also to [8] for guidelines on the choice among the plethora of available Type I error rates. [12, 13] also introduce a Type II error rate, the False Non-discovery Rate or False Negatives Rate (FNR), defined as the expected ratio of false negatives to the number of not rejected hypotheses, if any.

In this paper we focus on the k-FamilyWise Error Rate (k-FWER), defined as the probability of having k or more false positives:

$$\Pr(N_{1|0} \geq k). \qquad (2)$$

The 1-FWER reduces to the classical FWER. The idea behind the use of $k > 1$ is that by allowing for a small (say $k-1$) number of false positives, the number of rejected hypotheses can be largely increased with respect to requiring that there are no false positives (the 1-FWER).

One of the first proposal for control of the k-FWER is in [2], by comparing the j-th p-value with the rank dependent cut off $k\alpha / (m + \min(k - j, 0))$. This approach is improved in [14], who propose refinements which anyway are valid only under independence or at least under assumptions on the dependence of the test statistics.

A two step procedure is given in [3], which is based on the idea of augmentation: control the 1-FWER at level α and then add the $k-1$ most significant not rejected hypotheses. In this paper we will compare with [3] using the Bonferroni correction at the first step.

Other methods are proposed in [4, 15], but are based on computationally intensive resampling approaches.

The main drawback of all k-FWER controlling procedures is that they are conservative, especially for large number of tests; meaning that the error rate is controlled at a level well below the nominal, leaving room for more rejections and higher power. This will be illustrated below with simulations, and is well known about augmentation. The main reason behind conservativeness of the [3] methods is that 1-FWER controlling procedures used at the first step usually get more and more conservative as the number of tests grows. The fewer hypotheses are rejected at the first stage, the fewer at the second. This is the consideration that lead [6] to introduce generalized augmentation, which is based on uncorrected testing at the first step. The number of hypotheses rejected at the first step, when uncorrected testing is used, usually *grows* with the number of tests; leading to possibly much higher power. The GAUGE procedure will be seen to be still controlling the k-FWER at level below the nominal α, but it is much less conservative than the available competitors.

As we will point out below, the procedures of [3] and the one-step version of the procedure proposed in [2] can be seen as special cases of our approach.

2. THE GAUGE PROCEDURE FOR k-FWER CONTROL

The idea of generalized augmentation [6] is as follows: at fist, uncorrected testing at level q is performed. If at this first step the level q was too stringent, augmentation can be applied by adding a suitable number of rejections chosen among the most significant not rejected hypotheses. If the first step was too liberal and control of the Type I error measure is not guaranteed, a *negative* augmentation must be applied by removing some of the least significant rejected hypotheses.

The number of hypotheses to add/remove, and whether to add/remove or not, is determined through a probabilistic reasoning.

This probabilistic reasoning is dependent on the Type I error rate which one wants to control using generalized augmentation. We now develop the rules needed to control the k-FWER.

Denoting with 1_C the indicator function of condition C, we can summarize the generalized augmentation procedure we propose as:

1. Do uncorrected testing at level q. Reject S hypotheses. If $S = 0$, stop.

2. Consider the set of positive augmentations

$$P = \left\{ \begin{array}{l} P(\alpha) \in \{0,\dots,k-1\} : \max\limits_{m_0=k,\dots,m} \sum\limits_{i=\max(k-P(\alpha),0)}^{m_0} \\ \binom{m_0}{i} q^i (1-q)^{m_0-i} \leq \alpha \end{array} \right\}$$

If P is not empty, reject the $S + \max_{P(\alpha) \in P} P(\alpha)$ most significant hypotheses, and stop.

3. If P is empty, consider the set of negative augmentations

$$N = \left\{ N(\alpha) \in \{1,\dots,S\} : \max\limits_{m_0=k,\dots,m} 1_{\{S-N(\alpha)>0\}} \right.$$

$$\left. \left(\sum_{i=0}^{m_0} \sum_{j=0}^{\min(N(\alpha),i)} 1_{\{i-j \geq k\}} \binom{m_0}{i} q^i (1-q)^{m_0-i} \frac{\binom{i}{j}\binom{m-i}{N(\alpha)-j}}{\binom{m}{N(\alpha)}} \right) < \alpha \right\}.$$

Reject the $S - \min_{N(\alpha) \in N} N(\alpha)$ most significant hypotheses, if any, and stop.

A formal proof of k-FWER control is given in next theorem:

Theorem 1 If the test statistics are independent, the generalized augmentation procedure provides control of $k - FWER$ at level α for any $q \in (0,1)$.

Proof. From [16] we have that under independence the number of false positives obtained at the first step is $Binomial(M_0, q)$.

If we augment with $P(\alpha)$ hypotheses, We can evaluate the k-FWER as

$$\Pr(N_{1|0} \geq k) \leq \sum_{i=0}^{M_0} 1_{\{i+P(\alpha) \geq k\}} \binom{M_0}{i} q^i (1-q)^{M_0-i}$$

$$= \sum_{i=k-P(\alpha)}^{M_0} \binom{M_0}{i} q^i (1-q)^{M_0-i}$$

$$\leq \max_{m_0=k,\dots,m} \sum_{i=k-P(\alpha)}^{m_0} \binom{m_0}{i} q^i (1-q)^{m_0-i}.$$

The max operator at the last step is taken from k to m, since the k-FWER is trivially zero for $M_0 < k$. The last expression is bounded by α at the positive augmentation step, when it is possible.

When a positive (or null) augmentation is not possible, let V_1 denote the number of false rejections at

the first step and W be the number of hypotheses not rejected after negative augmentation that were in fact true nulls. That is, W denotes the number of false rejections taken out from the set of rejections after uncorrected testing. It is straightforward to check that W is stochastically dominated by W_1, the number of false rejections taken out from the set of rejections after uncorrected testing if the hypotheses chosen for un-rejection were chosen at random. In fact, we un-reject the least significant hypotheses, which are the less likely to be correspondent to true nulls among the rejected. We can now bound $\Pr(W_1 \leq j \mid V_1 = i)$ as follows:

$$\Pr(W_1 \leq j \mid V_1 = i) = \sum_{s=i}^{m} \Pr(W_1 \leq j \mid V_1 = i, S = s)$$

$$\Pr(S = s \mid V_1 = i)$$

$$\leq \Pr(W_1 \leq j \mid V_1 = i, S = m)\sum_{s=i}^{m} \Pr(S = s \mid V_1 = i)$$

$$\leq \Pr(W_1 \leq j \mid V_1 = i, S = m).$$

It is straightforward to see that $W_1 \mid V_1 = i, S = m$ follows an Hypergeometric distribution with parameters $(m, i, N(\alpha))$.

We can then finally bound the k-FWER after negative augmentation as follows:

$$\Pr(N_{1|0} \geq k) = \sum_{i}\sum_{j} \Pr(N_{1|0} \geq k \mid V_1 = i, W = j)$$

$$\Pr(W = j \mid V_1 = i)\Pr(V_1 = i)$$

$$\leq \max_{m_0 = k, \ldots, m} 1_{\{S - N(\alpha) > 0\}} \left(\sum_{i=0}^{m_0} \sum_{j=0}^{\min(N(\alpha), i)} \right.$$

$$\left. 1_{\{i - j \geq k\}} \binom{m_0}{i} q^i (1-q)^{m_0 - i} \frac{\binom{i}{j}\binom{m-i}{N(\alpha) - j}}{\binom{m}{N(\alpha)}} \right)$$

and the last expression is bounded by α at the negative augmentation step.

We now briefly discuss the rationale behind the procedure. The k-FWER of uncorrected testing, performed at the first step, can be easily evaluated since the number of false positives under independence is easily seen [16] to be a binomial with parameters M_0 and q. We can then evaluate the probability of having exactly i false positives among the S, and bound this probability simply by taking the maximum for $M_0 \in k, \ldots, m$ (since the actual number of true nulls in unknown). Positive or null (i.e., no

adjustment to the set of S rejections) augmentation is performed by conservatively assuming that each added rejection is a false rejection. This is done also in [3], who recognize that a more refined probabilistic approach would improve the procedure but is actually very hard to formalize. If the k-FWER after the first step is seen to be above α, a negative augmentation is needed. When doing negative augmentation we take into account the probability that an hypothesis that is taken out is actually true, which leads to use of the hypergeometric distribution. We then evaluate the k-FWER as a sum of the probability of each configuration of number of false positives at the first step (i) and number of false positives removed at the negative augmentation step (denoted with j in the formula). More rigorous insights are given in the proof of Theorem 1.

Note that the procedure is easily implemented and needs only evaluating the binomial and possibly the hypergeometric distribution, at most m times. The approach is then also not computationally intensive.

By setting $q = \alpha / m$ (hence controlling the 1-FWER at the first step) the proposed approach reduces to the [3] *augmentation* procedure. Here we replaced FWER control at the first step with uncorrected testing. The advantage is that the number of hypotheses rejected at the first stage is much higher, leading to more flexibility and finally an higher power, especially for large number of tests. Note further that setting $q = k\alpha / m$ a null augmentation ($P(\alpha) = 0$) is obtained, which gives back the one-step procedure of [2].

In our approach q can be any value on $(0,1)$, provided it is chosen in advance. A natural choice is obviously $q = \alpha$. In our experience values in the interval $(\alpha / m, \alpha)$ usually lead to higher power, while the results are often not very smooth with respect to the choice of q. For illustration, in this paper we will set $q = 0.0005$ leading the procedure to coincide with augmentation when $m = 100$ (since we fix $\alpha = 0.05$), and to be less stringent at the first step for $m > 100$. When $m < 100$ or the magnitude of the smaller p values is expected to be moderate, we suggest using larger values of q in order to have $S > 0$ at first stage with high probability.

If certain assumptions on the dependence are available, the original procedure can be used with no need for modification. These results are achieved in case of positive or negative association [17, 18], or α-mixing [19]. Conditions on the α-mixing sequence of coefficients essentially make sure that the dependence decreases fast enough as lag between test statistics grows. More precisely, refer to conditions (1-6) of Theorem 1 of [20], whose results directly extend. Since

our procedure is permutation invariant, we only need that there exist at least one permutation satisfying the condition. Other arguments relate to Poisson approximation of dependent indicators. The theoretical conditions of [20] are useful for the applications since for instance genes spotted in DNA microarrays are usually thought to be dependent in blocks, which directly implies the conditions. Hence the proposed approach can be applied to genomic data, especially when the number of genes is large. For a deeper discussion on the issue of applicability of asymptotic results under dependence and/or Poisson approximation results in the area of genomics refer to [20] and to [6].

3. SIMULATION STUDY

A brief simulation study is used to illustrate the newly proposed procedure. We generate independent unit variance normal variates, with sample size $n = 1$, for different values of m. We test null hypotheses of the mean being equal to zero against a one-sided alternative, assuming known variance. Table **2** compares 3-FWER controlling procedures when mean values for each true alternative hypothesis are sampled from a uniform on $[0,5]$, and $M_0 = \lceil m * 0.9 \rceil$. AUG stands for augmentation [3], LR for the stepdown [2] procedure, GR for the [14] procedure; while GAUGE denotes the proposed procedure. Table **3** shows the same when mean values for each true alternative hypothesis are sampled from a uniform on $[0,2.5]$, making it much more difficult to distinguish between true and false nulls.

Table 2: **FNR and 3-FWER (in Parentheses) for Different Procedures at Level α = 0.05, $M_0 = \lceil 0.9m \rceil$, Data Generated from Independent Standard Normals, Alternatives Sampled from a *U*[0,2.5]. The Results are Based on 1000 Iterations**

m	AUG	LR	GAUGE	GR
200	0.0626 (0.001)	0.0651 (0.001)	0.0633 (0.001)	0.0598 (0.006)
500	0.0720 (0.000)	0.0702 (0.000)	0.0657 (0.002)	0.0653 (0.003)
700	0.0742 (0.002)	0.0719 (0.003)	0.0666 (0.009)	0.0678 (0.007)
1000	0.0765 (0.000)	0.0736 (0.002)	0.0678 (0.009)	0.0696 (0.002)

The comparison is done in terms of average observed FNR, which is smaller for more powerful procedures. We can then compare the procedures in terms of power, and see what happens as the number

of tests grows. The average observed 3-FWER is reported in parentheses, with the only aim to check that it is below the nominal level $\alpha = 0.05$. It also confirms that all methods are still conservative and can be improved.

Table 3: **FNR and 3-FWER (in Parentheses) for Different Procedures at Level α = 0.05, $M_0 = \lceil 0.9m \rceil$, Data Generated from Independent Standard Normals, Alternatives Sampled from a *U*[0,2.5]. The Results are Based on 1000 Iterations**

m	AUG	LR	GAUGE	GR
200	0.0906 (0.004)	0.0948 (0.001)	0.0895 (0.026)	0.0921 (0.006)
500	0.0951 (0.000)	0.0967 (0.001)	0.0941 (0.009)	0.0946 (0.002)
700	0.0960 (0.001)	0.0970 (0.000)	0.0944 (0.009)	0.0957 (0.006)
1000	0.0970 (0.000)	0.0975 (0.001)	0.0955 (0.009)	0.0963 (0.004)

The new procedure compares well in terms of power in these settings. In particular, we can see from Table **2** that GAUGE seems to be slightly less powerful than GR when the number of tests in moderate, and it slightly outperforms GR when the number of tests is larger. When the signal is weaker, GAUGE seems to be the most powerful procedure for all m. The differences are more and more evident as the number of tests grows. This feature of generalized augmentation is seen also when controlling other error measures [6]. Similar results are seen in other settings (i.e., changing M_0 and the strenght of the signal) and also under different scenarios of dependence (simulations not shown). In pratice, with a larger number of tests and weaker signal GAUGE seems to be preferable. This evidence, provided by the simulations, does not change when one changes M_0.

4. APPLICATION TO CLASSIFICATION OF COLON CANCER

In order to demonstrate the usefulness of our approach on real data, in this section we analyze a benchmark high-density microarray data set from [21]. The dataset consists of records on 2000 genes from 62 individuals, 22 safe and 40 ill of colon cancer. There are two main goals: first, identification of a subset of genes differentially expressed, and hence connected to colon cancer. Secondly, to correctly classify new individuals as safe or ill on the basis of measurements on the selected (small) subset of genes. Hence, a good procedure selects few genes which are useful for

classification. We do not focus on classification methods in this paper, and use the simple 3-Nearest Neighbor classifier [22]. We use two-sample t-test statistics to test for differential gene expression between the two groups for each gene.

First, we apply the methods to the entire data set, obtaining a list of 74 genes with GAUGE, 40 with GR, 28 with LR and 13 with AUG; when all procedures are set to control the 3-FWER at level $\alpha = 0.05$.

Secondly, we randomly split the data into a training set of 30 samples and a test set of 32 samples. For each procedure, we select a subset of statistically significant differentially expressed genes on the training set, and use the selected genes to classify the tissues on the test set. We set $k = 3$ and $\alpha = 0.05$. We repeat the operation 1000 times and report the average number of selected genes and the corresponding estimated classification error in Table **4**.

The last column of Table **4** shows, for each procedure, the proportion of times it resulted in the lowest classification error over the 1000 replications. When $v > 1$ procedures achieved the same minimal classification error, they received a weight equal to $1/v$ for that iteration.

Table 4: **Average Number of Selected Genes and Estimated Classification Error for the [21] Data. Pr(Best) is the Proportion of Iterations in which the Procedure Achieved the Lowest Error Rate in Comparison to the others**

Multiple Testing	Selected	Classification	Pr(Best)
Procedure	Genes	Error	
GAUGE	26.90	0.139	0.391
AUG	8.34	0.161	0.196
LR	10.61	0.226	0.172
GR	16.94	0.157	0.242

We can see that the proposed procedure leads to an higher number of selected genes and a smaller classification error for this data set. The difference between classification errors is not substantial, but steady over replications, and likely due to the higher number of selected genes.

We stress that procedures leading to an higher number rejected hypotheses are of interest when the focus is on gene discovery in this application. The set of genes corresponding to rejected null hypotheses are considered for confirmatory methods, like Polymerase Chain Reaction, and a larger list results in a wider choice for validation and in an higher likelihood of inclusion of biologically significant genes. In this application our approach lead to an higher number of selected genes than the other methods, but this number is still reasonable for the screening with confirmatory methods.

5. DISCUSSION

We proposed a new multiple testing procedure, based on the idea of augmentation: at first, a certain number of hypotheses is rejected through a one-step method; then this number is adjusted to achieve control of the k-FWER. We not that our procedure cannot be deemed to be stepwise, so that a formal comparison, e.g., along the lines of [23], may be hard to perform.

The procedure was seen to be competitive with other k-FWER controlling procedures, especially when the number of tests is large. As illustrated by simulations, generalized augmentation may allow for a higher power when the number of tests is large, but it still is conservative, leaving room for improvement by a tighter control of the k-FWER.

It has been argued that the proposed approach is applicable in common genomics applications and its usefulness has been illustrated on a benchmark data set on colon cancer. Further, the approach is easy to implement and it is not computationally intensive.

REFERENCES

[1]   Hommel G, Hoffman T. Controlled uncertainty. In: Bauer P, Hommel G, Sonnemann E, Eds. Multiple Hypoteses Testing. Heidelberg: Springer 1988; pp. 154-161.
http://dx.doi.org/10.1007/978-3-642-52307-6_13

[2]   Lehmann EL, Romano JP. Generalizations of the Familywise Error Rate. Ann Stat 2005; 33: 1138-54.
http://dx.doi.org/10.1214/009053605000000084

[3]   van der Laan MJ, Dudoit S, Pollard KS. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. Stat Appl Gen Mol Biol 2004; 3(1).

[4]   Pollard KS, van der Laan MJ. Resampling-based multiple testing: Asymptotic control of Type I error and applications to gene expression data. J Stat Plan Inf 2005; 125: 85-100.
http://dx.doi.org/10.1016/j.jspi.2003.07.019

[5]   Finos L, Farcomeni A. *k*-FWER control without multiplicity correction, with application to detection of genetic determinants of multiple sclerosis in Italian twins. Biomet 2011; 67: 174-81.
http://dx.doi.org/10.1111/j.1541-0420.2010.01443.x

[6]   Farcomeni A. Generalized Augmentation to control the False Discovery Exceedance in Multiple Testing. Scand J Stat 2009; 36: 501-17.

[7]   Westfall PH, Young SS. Resampling-based Multiple Testing: Examples and Methods for *p*-value Adjustment. Wiley 1993.

[8]   Farcomeni A. A review of Modern Multiple Hypothesis testing with particular attention to the False Discovery Proportion. Stat Meth Med Res 2008; 17: 347-88.
http://dx.doi.org/10.1177/0962280206079046

[9]   Farcomeni A. Parametric assumptions in single and multiple testing: when should we rely on them? Biomed Stat Clin Epid 2008; 2: 57-69.

[10]   Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Roy Statist Soc Ser B 1995; 57: 289-300.

[11]   Sarkar SK. Stepup procedures controlling generalized FWER and generalized FDR. Department of Statistics, Temple University 2005.

[12]   Genovese CR, Wasserman L. Operating Characteristics and Extensions of the FDR Procedure. J Roy Statist Soc Ser B 2002; 64: 499-18.
http://dx.doi.org/10.1111/1467-9868.00347

[13]   Sarkar SK. FDR-controlling stepwise procedures and their false negatives rates. J Stat Plan Inf 2004; 125: 119-37.
http://dx.doi.org/10.1016/j.jspi.2003.06.019

[14]   Guo W, Romano J. A Generalized Sidak-Holm procedure and control of generalized error rates under independence. Stat Appl Genet Molec Biol 2007; 6(1).
http://dx.doi.org/10.2202/1544-6115.1247

[15]   Romano JP, Wolf M. Control of generalized error rates in multiple testing. Ann Stat 2007; 35: 1378-408.
http://dx.doi.org/10.1214/009053606000001622

[16]   Finner H, Roters M. Multiple hypotheses testing and expected number of Type I errors. Ann Stat 2002; 30: 220-38.

[17]   Esary JD, Proschan F, Walkup DW. Association of Random Variables, with Applications. Ann Math Stat 1967; 38: 1466-74.
http://dx.doi.org/10.1214/aoms/1177698701

[18]   Kumar JD, Proschan F. Negative Association of Random Variables with Applications. Ann Statis 1983; 11: 286-95.
http://dx.doi.org/10.1214/aos/1176346079

[19]   Doukan P. Mixing. Lectures Notes in Statistics, 85; Springer-Verlag 1994.

[20]   Farcomeni A. Some Results on the Control of the False Discovery Rate under Dependence. Scand J Stat 2007; 34: 275-97.
http://dx.doi.org/10.1111/j.1467-9469.2006.00530.x

[21]   Alon U, Barkai N, Notterman DA, *et al*. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissue probed by oligonucleotide arrays. Proc Natl Acad Sci USA 1999; 96: 6745-50.
http://dx.doi.org/10.1073/pnas.96.12.6745

[22]   Cover T, Hart P. Nearest neighbor pattern classification. IEEE trans inf th 1967; IT-13: 21-27.

[23]   Gordon AY. Explicit formulas for generalized family-wise error rates and unimprovable step-down multiple testing procedures. J Stat Plan Inf 2007; 137: 3497-12.
http://dx.doi.org/10.1016/j.jspi.2007.03.027