# A new multiple testing procedure with applications to quantitative biology and wavelet thresholding

Alessio Farcomeni*[1], Giovanna J. Lasinio[1], Chiara Alisi[2] and Luigi Chiarini[2]

[1] University of Rome "La Sapienza"
[2] ENEA-Casaccia.

## 1 Introduction

It is well known that, when doing many tests at once, a correction for multiplicity is needed to avoid explosion of the number of false rejections. Moreover, many procedures (like the Bonferroni correction) that do this do not work well when the number of tests is large, leading to essentially no rejections.

Modern applications in bioinformatics (DNA Microarrays, brain imaging, etc.), in which the number of tests can be in the order of the thousands or even hundreds of thousands, motivated research in this area for more powerful and sensible corrections; following the seminal paper of Benjamini and Hochberg (1995), who introduced the False Discovery Rate (FDR).

Recently, a new Type I error rate, the False Discovery eXceedance (FDX), has been proposed (Genovese and Wasserman 2004; Dudoit *et al.*, 2004). The FDX is known to be more sensible than FDR in certain cases, and in particular when the "signal" in the data may be sparse/weak.

Unfortunately, some of the existing procedures to control the FDX may share the problems of Bonferroni correction, even if they control a far less strict error rate. In this work we propose a new procedure, tailored for the case of large number of tests, and show applications in quantitative biology and wavelet thresholding. We prove it controls the FDX for any number of tests. Moreover, simulations show that the new procedure usually achieves an higher power than existing methods, while still controlling the Type I error rate. A formal wider statement on power is still under investigation.

### 1.1 Background

Consider a multiple testing situation in which $m$ tests are being performed. Suppose $m_0$ of the $m$ hypotheses are true, and $m_1$ are false. Table 1 shows a categorization of the outcome of the tests. $R$ is the number of rejections. Note that $N_{1|0}$ and $N_{0|1}$ are the exact (unknown) number of errors committed, respectively false positives and false negatives. $N_{1|1}$ and $N_{0|0}$ are respectively the number of hypotheses correctly rejected and correctly retained.

| | $H_0$ **not rejected** | $H_0$ **rejected** | **Total** |
|---|---|---|---|
| $H_0$ **True** | $N_{0\|0}$ | $N_{1\|0}$ | $m_0$ |
| $H_0$ **False** | $N_{0\|1}$ | $N_{1\|1}$ | $m_1$ |
| **Total** | $m - R$ | $R$ | $m$ |

Table 1: Categorization of the outcome

In the frequentist framework, one controls the probability of a single Type I error (i.e., false rejection). When $m >> 1$, this can result in an explosion of the number of false rejections. For

this reason, a different error measure (Type I error rate) is considered. Traditional methods for multiple testing, like the well known Bonferroni correction, attempted control on the Family-wise error rate (FWER), i.e., $P(N_{1|0} > 0)$. FWER proves to be too strict as an error measure, leading essentially to no rejections as $m$ grows.

Define now the False Discovery Proportion (FDP) as the proportion of incorrect rejections over the number of rejections: $FDP = \frac{N_{1|0}}{R+1_{\{R=0\}}}$, where $1_{\{\cdot\}}$ is the indicator function. Modern applications attempt a control on the expected value of this random variable (the False Discovery Rate, or FDR) or on the tail probability (False Discovery eXceedance, or FDX), i.e., the probability that it exceeds a specified cut-off $c$ (typically, 0.1).

The setting is as follows: call $p_i$ the test statistic computed for the i-th hypothesis. The test statistic will usually be a $p$-value. Suppose we reject the i-th hypothesis if $p_i$ is smaller than a certain $t \in (0, 1)$, so that $I_i = 1_{\{p_i<t\}}$ will be equal to 1. Since the $p$-values are ordered, the only problem is to set a threshold $t$ such that the chosen error measure is below a pre-specified level $\rho$ (typically, 0.05).

# 2 A new control procedure for the FDX

We will now propose a new procedure to control the FDX.

The algorithm is as follows:

1. Choose $q \in (0, 1)$.
2. Reject the $S = |S_q|$ $p$-values smaller than $q$ Here $S_q$ is the set of indices of all $p$-values smaller than $q$ and $|\cdot|$ gives the cardinality of a set.
3. Let $i^*$ be
$$\min\{i : \sum_{k=i}^{S} \binom{m}{k} q^k (1-q)^{m-k} \leq \rho\}. \tag{2.1}$$
Note that $i^*$ is easily found for fixed $m$ and $\rho$, consisting in the evaluation of the binomial distribution with parameters $m$ and $q$.
4. If $\frac{(i^*-1)}{|S_q|} \leq c$, let $k_n(c, \rho) = \max\{j \in \{0, \ldots, m - |S_q|\} : \frac{j+i^*-1}{j+|S_q|} \leq c\}$.

If $k_n(c, \rho)$ exists and is positive, we can do an "augmentation" of $S_q$: any choice of $k_n(c, \rho)$ additional hypotheses will (still) control the FDX at level $\rho$. If $\frac{(i^*-1)}{|S_q|} > c$ or $i^*$ does not exist, then at the first step we rejected too many hypotheses. One can pick any of this two choices:

1. Choose a smaller $q$ (for instance, divide by 2 the previous one), and repeat the procedure.
2. Do a negative augmentation in this way: Let
$$\begin{aligned} k'_n(c, \rho) \quad = \quad & \min\{k \in \{0, \ldots, |S_q|\} : \\ & 1_{\{|S_q|-k>0\}} \left( \sum_{i=0}^{|S_q|-k} \sum_{j=0}^{\min(k,i)} \right. \\ & 1_{\{(i-j)/(|S_q|-k)>c\}} \binom{m}{i} q^i (1-q)^{m-i} \\ & \left. \frac{\binom{i}{j}\binom{|S_q|-i}{k-j}}{\binom{|S_q|}{k}} \right) < \rho\}. \end{aligned} \tag{2.2}$$

Then reject only the $|S_q| - k'_n(c, \rho)$ most significant $p$-values. This will control the FDX at level $\rho$.

If the first choice is taken, we call the procedure "GAUGE Type I", otherwise "GAUGE Type II". GAUGE stands for Generalized AUGmEntation. It can in fact be seen that this procedure is a generalization of Van der Laan *et al.* (2003).

The innovation over existing FDX controlling procedures is that usually a larger number of rejections is done, hence achieving an higher power; and the difference becomes more and more evident as $m$ grows. Conditions on the signal (i.e., the distribution of the test statistics under the alternative hypothesis), $a$, and the parameters lead to prove that the deciding point $T$ is strictly bounded below by zero as the number of tests grows. For instance, it can be seen that this happens with probability one if the $p$-values, under the alternative hypothesis, put a mass at zero. If $T$ is strictly bounded below by zero, then power is not infinitesimal. (Much) weaker conditions are under investigation.

Another open problem is the choice of $q$. For the time being, we will set $q = \rho$ (i.e., do uncorrected testing at the first step); which intuitively is not optimal since it often should lead to need of negative augmentation.

There also are conditions on dependence of the test statistics, under which the procedure still controls the FDX, which are omitted for reasons of space.

# 3 Applications

## 3.1 Application to wavelet thresholding

We point the reader to Abramovich and Benjamini (1996), who introduced the use of multiple testing procedures to do wavelet thresholding, for a description and introduction to the problem. Let $d_{jk}$ be the detail coefficients to be thresholded. It obviously happens that $p_i = f(d_{jk})$, where $f(\cdot)$ is not-increasing, for a certain $f(\cdot)$. Let $t$ be the multiple testing cut-off. The threshold is $\lambda = \max\{l : l = f(t)\}$.

Table 3.1 compares mean square errors for an image reconstruction problem. It can be seen that the new procedure, in this case, achieves a smaller or at most equal mean square error than classical SURE and Universal thresholding procedures.

| | Universal | SURE | GAUGE Type I | GAUGE Type II |
|---|---|---|---|---|
| $\sigma = 0.01$ | 0.183 | 0.0011 | 0.0001 | 0.0001 |
| $\sigma = 0.1$ | 0.186 | 0.010 | 0.010 | 0.023 |
| $\sigma = 0.25$ | 0.196 | 0.053 | 0.053 | 0.085 |
| $\sigma = 0.5$ | 0.218 | 0.170 | 0.147 | 0.139 |
| $\sigma = 0.75$ | 0.240 | 0.293 | 0.217 | 0.209 |

Table 2: Average mean square error for image reconstruction for different thresholding methods and noise levels, $m = 65536$

## 3.2 Application to metabolic profiling

We describe in this section analysis on unpublished data on metabolic profiling.

Burkholderia ambifaria is one of the Burkholderia cepacia complex (Bcc) species most frequently associated with roots of crop plants. In our laboratory we started studying the metabolic

profiles of Bcc species colonizing the rhizosphere of maize by means of the Biolog system using GN2 and SFN2 plates and different parameters related to optical density (OD). In the present study we carried out the metabolic profiling of a subset of 44 isolates of Burkholderia ambifaria, originally isolated from a single maize field in Northern Italy. The present work is an attempt to perform an analysis of the relative importance of the different substrates present in the Biolog system in shaping the metabolic profiles of the various bacterial isolates.

We have two replicates of most of the metabolic profiles organized over 95 substrates (38 out of the 44 isolates). We analyze these data by fitting ANOVA models to the 38 replicated isolates in order to asses substrates relevance.

The analysis involved the use of multiplicity correction as proposed, and as we expected all of the models were declared significant even after correction. This is a good result, showing that the method can adapt well to cases of strong signal by rejecting many hypotheses when it is opportune, even in a case in which the number of tests is small.

Then we perform cluster analysis with PAM for each replicate (the first one involving all 44 isolates) according to two perspectives:

1. we check how substrates cluster w.r.t. isolates
2. we check how isolates cluster w.r.t. substrates.

# References


Abramovich, F. and Benjamini, Y. (1996). Adaptive thresholding of wavelet coefficients. *Computational Statistics and Data Analysis*, **22**, 351-361.

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

Dudoit, S. and van der Laan, M.J. and Birkner, M.D. (2004). Multiple Testing Procedures for Controlling Tail Probability Error Rates. *Tech. Rep. 166*, Division of Biostatistics, UC Berkeley.

Farcomeni, A. (2005). More Powerful Control of the False Discovery Rate under Dependence. *Statistical Methods & Applications*. To appear.
Available at the page: `http://afarcome.interfree.it/cv.html`

Genovese, C.R. and Wasserman, L. (2004). Exceedance control of the False Discovery Proportion. *Tech. Rep.*. Department of Statistics, Carnegie Mellon University.

Storey, J.D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society (Ser. B)*, **74**, 479-498.

Van der Laan, M.J. and Dudoit, S. and Pollard, K.S. (2003). Multiple testing. Part III. Procedures for control of the generalized family-wise error rate and proportion of false positives. *Tech. Rep. 141*. Division of Biostatistics, UC Berkeley.