Supplementary material for FDR Control with Pseudo-Gatekeeping Based on a Possibly Data Driven Order of the Hypotheses

A. Farcomeni*

Department of Public Health and Infectious Diseases
Sapienza - University of Rome, Piazzale Aldo Moro, 5, 00185 Rome, Italy
*email: alessio.farcomeni@uniroma1.it

and

L. Finos *

Department of Statistical Sciences
University of Padua, Via Cesare Battisti 241, 35121 Padova, Italy
*email: livio@stat.unipd.it

SUMMARY: We propose a multiple testing procedure controlling the false discovery rate. The procedure is based on a possibly data driven ordering of the hypotheses, which are tested at the uncorrected level \( q \) until a suitable number is not rejected. When the order is data driven, larger effect sizes are considered first, therefore selecting more interesting hypotheses with larger probability. The proposed procedure is valid under independence for the test statistics. We also propose a modification which makes our procedure valid under arbitrary dependence. It is shown in simulation that we compare particularly well when the sample size is small. We conclude with an application to identification of molecular signatures of intracranial ependymoma. The methods are implemented in an R package (someMTP), freely available on CRAN. This paper includes supplementary material online.

KEY WORDS: Data driven order; False Discovery Rate; Multiple testing.

This paper has been submitted for consideration for publication in Biometrics
This Web Appendix is organized as follows: in Section 1 we provide the proofs to Lemma 1, Theorem 1, Theorem 2 and Theorem 3 of the main paper. Section 2 reports the results of additional simulation studies. In Section 3 we briefly describe an example in which the test statistics are ordered \textit{a-priori}, to complement the example with data driven order in the main paper.

\section{Proofs of Lemma and Theorems}

\subsection{Proof of Lemma 1}

We begin by noting that $N_{1|0}(i)$ is a sum of $i$ independent indicators, and the process stopped at the $i$-th hypothesis since $J(i, q)$ successes, each of probability at least $1 - q$, are obtained. A success for this negative binomial trial is defined as $p(j) > q$. At the same time, for all $j < i$, there were less than $J(j, q)$ successes.

We have the following assumption on $J(i, q)$:

\begin{equation}
J(i_1, q) \leq J(i_2, q) \quad \text{for any } i_1 \leq i_2.
\end{equation}

We begin by noting that if $J(i, q) = J(q)$, that is, (1) holds with $J(i_1, q) = J(i_2, q)$ for all $i_1, i_2$, we have a negative binomial trial. This is exactly the setting of Finos and Farcomeni (2011), Theorem 1. If $J(i, q) = J(q)$, note that

\begin{equation}
\Pr(N_{1|0}(i) \geq k | J(q)) = \Pr(N_{1|0} \geq k \cap I = i | J(q))
\end{equation}

\begin{equation}
= \Pr(I = i | N_{1|0} \geq k \cap J(q)) \Pr(N_{1|0} \geq k | J(q))
\end{equation}

\begin{equation}
\leq \Pr(N_{1|0} \geq k | J(q)),
\end{equation}

where $I$ denotes the random variable indicating the position of the last hypothesis reached by the algorithm, we have explicited that the algorithm stops after $J(q)$ successes by conditioning; and the first equality follows from the very definition of $N_{1|0}(i)$. For the marginal number of type I errors $N_{1|0}$ it is shown in Finos and Farcomeni (2011), Theorem
1, that the thesis holds under the assumption that \( J(i, q) \) does not depend on \( i \). Given that the proof of this statement is exactly the same as that of point (i) in Finos and Farcomeni (2011), Theorem 1, we omit it. We therefore have
\[
\Pr(N_{i|0} \geq k|J(q)) \leq 1 - F_{Bneg(J(q), 1-q)}(k)
\]
and consequently the thesis:
\[
\Pr(N_{i|0} \geq k|J(q)) \leq 1 - F_{Bneg(J(q), 1-q)}(k - 1).
\]
To see that the thesis holds also under (1) we can use Theorem 1 in Finos and Farcomeni (2011) and the following reasoning: when the number of successes prescribed before stopping is smaller, the algorithm stops earlier and hence there is a lower number of type I errors.

We now explicit this reasoning more formally. Let \( \mathcal{H}_0 \) denote the set of indices associated with true null hypotheses. We have:
\[
\Pr(N_{i|0} \geq k) = \Pr\left( \sum_{j \in \mathcal{H}_0 \cap 1, \ldots, i} I(p_{(j)} > q) \geq k \right)
\]
\[
\leq \Pr\left( \sum_{j \in \mathcal{H}_0 \cap 1, \ldots, i} I(p_j > q) \geq k \right)
\]
\[
\leq \Pr\left( \sum_{j=1}^{i} I(p_j > q) \geq J(i, q) \cap \sum_{j=1}^{i-1} I(p_j > q) < J(i-1, q) \cap \ldots \cap I(p_1 > q) = 0 \right)
\]
\[
\leq \Pr\left( \sum_{j=1}^{i} I(p_j > q) \geq J(i, q) \cap \sum_{j=1}^{i-1} I(p_j > q) < J(i-1, q) \cap \ldots \cap I(p_1 > q) = 0 \right)
\]
where we explicited the algorithm at the first step, used the hypothesis on a priori ordering at the second, (1) at the third. At the third step we now obtain a classical negative binomial trial, with number of successes \( J(i, q) \). Consequently, as long as the algorithm is stopped
on or before the $i$-th hypothesis, we can use Finos and Farcomeni (2011), Theorem 1, to show the last step. The result holds for any $i > 0$, hence the thesis.

1.2 Proof of Theorem 1

Note that (1) trivially holds in our definition $J(i, q) = i(1 - q)/(2 - q)$. We can therefore use Lemma 1 to study the expected value of $N_{1|0}(i)$:

\[
E[N_{1|0}(i)] = \sum_{j=1}^{i} \Pr(N_{1|0}(i) \geq j) \tag{2}
\]

\[
\leq \sum_{j=1}^{i} 1 - F_{B neg}(J(i,q),1-q)(j - 1)
\]

\[
\leq \sum_{j=1}^{\infty} 1 - F_{B neg}(J(i,q),1-q)(j - 1)
\]

\[
= J(i,q)q/(1 - q).
\]

Our proposed procedure leads us to stop testing after $u$ $p$-values have been reached by the sequential testing. Note that $u$ is a random variable, and that there are at least $u - J(u,q)$ rejections if $u > 0$, and 0 rejections if $u = 0$. We can now bound the FDR:

\[
FDR = E \left[ \frac{N_{1|0}(u)}{R(u)} \right]
\]

\[
\leq \sup_i E \left[ \frac{N_{1|0}(i)}{R(i)} \right]
\]

\[
\leq \sup_i \frac{E[N_{1|0}(i)]}{i - J(i,q)}
\]

\[
\leq \sup_i \frac{J(i,q)q}{(1 - q)(i - J(i,q))},
\]

where we used (2) at the last step. If we now fix $J(i,q) \leq i(1 - q)/(2 - q)$, we have the following inequality for the last expression:

\[
\sup_i \frac{J(i,q)q}{(1 - q)(i - J(i,q))} \leq \sup_i \frac{iq(1 - q)}{(2 - q)(i - i(1 - q)/(2 - q))}
\]

\[
= \frac{i(1 - q)}{i(1 - q)} = q.
\]

Consequently, if we fix $u$ using the proposed algorithm, we have $FDR \leq q$ under independence.
1.3 Proof of Theorem 2

Define $p_{jki} = \Pr(p_j \leq q_D \cap u = i \cap R = k|j \in \mathcal{H}_0)$ (i.e. the probability that a true null hypothesis is rejected and the procedure stops at step $i - 1$ after $k$ rejections). Recall that $\mathcal{H}_0$ denotes the set of indices associated with true null hypotheses. Note that $p_{(j)ki} = p_{jki}$ given the assumptions on a priori ordering. The FDR representation we now use has some similarities with that in the main result of Benjamini and Yekutieli (2001).

Let us remark that

$$\sum_{i=1}^{m} \sum_{k=1}^{i} p_{jki} = \Pr(p_j \leq q_D \cap (\cup_{i=1}^{m} \cup_{k=1}^{j} u = i \cap R = k)|j \in \mathcal{H}_0) = \Pr(p_j \leq q_D|j \in \mathcal{H}_0) \leq q/\sum_{j=1}^{m}(2 - q)/(j + 1).$$

The FDR can be written as (note that if $k < i/(2 - q)$ there are no rejections)

$$E\left[\frac{N_{1|0}}{R}\right] = \sum_{j \in \mathcal{H}_0} \sum_{i=1}^{m} \sum_{k=1}^{i} \frac{p_{jki}}{k} = \sum_{j \in \mathcal{H}_0} \sum_{i>j} \sum_{k \geq i/(2 - q)} \frac{p_{jki}}{k}$$

Consequently,

$$E\left[\frac{N_{1|0}}{R}\right] \leq \sum_{j \in \mathcal{H}_0} \sum_{i>j} \sum_{k \geq i/(2 - q)} \frac{2 - q}{k} p_{jki} \leq \sum_{j \in \mathcal{H}_0} \sum_{i>j} \sum_{k \geq i/(2 - q)} p_{jki} \leq \sum_{j \in \mathcal{H}_0} \frac{2 - q}{j + 1} \sum_{i>j} \sum_{k \geq i/(2 - q)} \frac{2 - q}{j + 1} \leq q$$

1.4 Proof of Theorem 3

Under the model and the null hypothesis, $M_j$ is a sufficient complete statistic for $\sigma_j$. By assumption the test statistic $T_j$ is ancillary to the dispersion parameter $\sigma_j$ under its corresponding $H_0$, therefore $M_j$ is independent of the test statistic $T_j$ through Basu theorem. This holds for all $p$-values calculated by transformation of the null CDF of test statistics $T_j$, even with arbitrarily dependent errors (see for instance Läuter et al. (1998), Finos (2011)).

As a consequence, the ordering on which any hypothesis $j$ is tested does not depend on the value of $T_j$ whenever the corresponding null hypothesis is true. The corresponding $p$-value
is therefore upper bounded by a uniform distribution under the null for assigned testing rank. We can now use Theorem 1 under independence or Theorem 2 without assumptions on dependence.

2. Additional simulation studies

We give in this section an account of additional simulation studies. The setting is as follows: we perform one-sample t-tests, with data generated from standard normals under the null hypothesis. We let \( n = 5, 10, 20, 50; \) \( m = 100, 1000, 10000; \) and \( q = 0.01, 0.05, 0.10. \) We fix the mean under the alternative hypotheses so that the single tests have a prescribed power of 70% when \( q = 0.05 \) (hence, the mean under the alternative decreases as the sample size \( n \) increases). The position of the false hypotheses are selected randomly at each iteration, and unless stated otherwise the number of false nulls is set so that 10% of the \( m \) hypotheses are false.

For each setting we generate the data, compute \( p \)-values, and apply the Benjamini and Hochberg (1995) (BH) and Benjamini and Yekutieli (2001) (BY) procedures; together with our procedure with data-driven order of the hypotheses (Ord) and the version for general dependence (OrdGD). We perform 10000 Monte Carlo iterations and estimate power through the fraction of correctly rejected hypotheses over the number of false nulls.

2.1 Variability

In this section we replicate the independence setting in the main paper, but we now report on the variability of the measures. Given that the Ord procedure seems to be strict at the initial stages and liberal afterwards, it could happen that our procedure could stop very early in certain cases and proceed very far in other cases. This would imply a large variability of the power measure

\[
Pow = \frac{N_{1\vert 1}}{M_1}.
\]
In order to check this point, we compare our procedure \((Ord)\) with the Benjamini and Hochberg (1995) \((BH)\) procedure. We compute the coefficient of variation of \(Pow\), as defined above, and of the quantity

\[ FDP = \frac{N_{1|0}}{R + I(R = 0)}. \]

We report the ratio of the two coefficients of variation, and the raw coefficients of variation for \(Pow\) in Table 3 for different values of \(n\) and \(m\).

The most important message of Table 3 is given by the column regarding \(CV(Pow_{Ord})\), which reports fairly small coefficients of variation in most cases, except maybe when \(n = 50\) and \(m = 100, 1000\). Hence, Ord procedure with small or moderate sample size does not yield a power with large variability. When \(n = 5, 10\) the \(CV(Pow)\) of the Ord procedure is actually smaller than or approximately equal to that of the BH procedure. When \(n \geq 20\), the BH procedure leads to a very small coefficient of variation for our measure of power.

For what concerns the variability of the error measure, we have a similar result: when \(n < 20\), Ord procedure gives a smaller variability than the BH procedure, and for \(n \geq 20\) the reverse holds. As a referee noted, the variance of the FDP may be inflated for both procedures under dependence (see e.g. Owen (2005); Farcomeni (2006, 2007)).

This simulation study complements the simulations in the main paper and in this Web Appendix, and confirms that the Ord procedure is particularly suited for cases in which the sample size is relatively small.

### 2.2 FDR control under block dependence

In this Section we generate data as described in Section 2, with dependent test statistics. We use a block dependence assumption in order to mimic a genomic situation (it is frequently assumed that genes are dependent in blocks of size 20-30). We then generate data in blocks
of size 25, in which observations within each block arise from a multivariate normal with
a covariance matrix with diagonal elements equal to 1 and off-diagonal elements equal to
\( \rho = 0.3 \). Figure 1 and Figure 2 report the estimated power and the estimated level of the
multiple test, respectively.

[Figure 1 about here.]

[Figure 2 about here.]

It can be seen that the nominal error rate is not exceeded by the procedures. It actually is
very hard to find simulated examples which lead the error rates to exceed the nominal error
rate, see for instance Benjamini and Yekutieli (2001) and Reiner-Benaim (2007). It can be
seen that the procedures valid under general dependence become more conservative. Our
Ord and OrdGD procedures outperform the respective competitors for \( n < 20 \), while for
\( n = 50 \) they are outperformed by BH and BY, respectively. When \( n = 20 \) the comparison
is mixed. It shall be noted that even if there is some mild dependence in the data, all
procedures still show a good power, which is comparable to that under independence as
illustrated in the main paper.

3. 1,25-dihydroxycholecalciferol in kidney transplant patients

1,25-dihydroxycholecalciferol (1,25-DHCC) is a physiologically active form of Vitamin D.
Vitamin D can be acquired either with the diet or by the metabolism of 7-dehydrocholesterol,
after exposure to ultraviolet light. Vitamin D is then transported to the liver where it
is converted into 25-hydroxycholecalciferol (25-HCC). 25-HCC is finally transported to
the kidney, which converts it to the active 1,25-DHCC. The active form is an important
hormone which regulates calcium absorption and mobilization, among other things. While
the benefits and importance of an adequate Vitamin D intake (by means of the diet or by
exposition to sunlight) are well recognized, it is apparent that the ability of the body to actually use the Vitamin D is linked to liver and kidney health.

In this study we compare 135 patients who have undergone a kidney transplant with 290 controls. For each of them we have a measurement of 25-HCC (in one of five classes) and 1,25-DHCC from blood samples. Our aim is to identify doses of 25-HCC in which transplant patients have a significantly different amount of 1,25-DHCC. If these are significantly lower, the study would help us measure what amounts of 25-HCC can a transplanted kidney bear as compared to a native kidney. A paper discussing the findings from a clinical perspective is in preparation.

We can proceed by pre-specifying the order of testing by considering the $p$-value associated with the lower 25-HCC class first, and then considering increasing classes of 25-HCC. The class with the largest 25-HCC is tested last (if at all). It is important to underline here that, as a matter of fact, any pre-specified ordering would lead to control of the FDR according to Theorem 1 in the main paper. The a priori order we suggest is justified according to the following reasoning: first of all, we have increasing doses of 25-HCC, which provide a natural ordering of the tests; secondly, we would like to focus on low 25-HCC doses, that is, results of clinical relevance are that associated with low 25-HCC doses. We are mostly interested in subjects with a low vitamin D intake since those subjects are at higher risk. Subjects with a limited kidney functionality but high vitamin D intake may still end up with enough 1,25-DHCC for a safe mineral equilibrium; hence any difference found associated with high initial amounts of 25-HCC may not be clinically relevant.

In Table 3 we report class-specific medians of 1,25-DHCC for each group and $p$-values arising from Mann-Whitney testing. Note that given that the order is pre-specified, we do not need any assumption on the error distribution.
With our procedure and $q = 0.05$ we end up rejecting the null hypothesis of no differences between transplants and controls in all but the third 25-HCC class. This corresponds to rejecting all $p$-values for which $p_j < 0.05$, as if we were in a single inference situation. On the other hand, Benjamini and Hochberg (1995) procedure leads us to reject the null hypothesis only for the last two classes, i.e., for $25\text{-HCC} \geq 26$. These correspond to very low $p$-values, but are much less clinically relevant than the first two classes. The additional, important, rejections obtained with the Ord procedure lead us to the following conclusions: when there is a low vitamin D intake, the transplanted kidney may not receive enough 25-HCC to promote the use of enzymes leading to as much active hormone as normal kidneys. Consequently, while suitable levels of vitamin D intake is an important recommendation for every one, it seems even more important for kidney transplants.

References


*Received*. *Revised*.

*Accepted*
Figure 1. Proportion of correctly rejected hypotheses for different values of $q$, $n$, $m$. The proportion of false nulls is set at 10%, the power of each single test at 70%. Block dependent test statistics.
<table>
<thead>
<tr>
<th>q</th>
<th>m</th>
<th>Sample size</th>
<th>FDR</th>
<th>Ord</th>
<th>OrdGD</th>
<th>BH</th>
<th>BY</th>
</tr>
</thead>
<tbody>
<tr>
<td>.01</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.01</td>
<td>1000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.01</td>
<td>10000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.05</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.05</td>
<td>1000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.05</td>
<td>10000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.10</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.10</td>
<td>1000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.10</td>
<td>10000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 2.** FDR for different values of q, n, m. The proportion of false nulls is set at 10%, the power of each single test at 70%. Block dependent test statistics.
Table 1
Coefficients of variation and their ratios for $\frac{\hat{N}^{\text{Ord}}}{R} (FDP)$ and $\frac{\hat{M}^{\text{Ord}}}{M_1} (Pow)$, obtained with the Ord and BH procedures under independence. The results are averaged over $B = 10000$ replications.

<table>
<thead>
<tr>
<th>n</th>
<th>m</th>
<th>CV($FDP_{\text{Ord}}$)</th>
<th>CV($FDP_{\text{BH}}$)</th>
<th>CV($Pow_{\text{Ord}}$)</th>
<th>CV($Pow_{\text{BH}}$)</th>
<th>CV($FDP_{\text{BH}}$)</th>
<th>CV($FDP_{\text{Ord}}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>100</td>
<td>0.50</td>
<td>0.26</td>
<td>1.19</td>
<td>0.31</td>
<td>3.27</td>
<td>1.64</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>0.90</td>
<td>0.85</td>
<td>0.59</td>
<td>0.50</td>
<td>2.31</td>
<td>2.07</td>
</tr>
<tr>
<td>20</td>
<td>100</td>
<td>1.59</td>
<td>2.07</td>
<td>0.41</td>
<td>0.84</td>
<td>1.91</td>
<td>3.05</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>2.51</td>
<td>3.87</td>
<td>0.34</td>
<td>1.30</td>
<td>1.79</td>
<td>4.50</td>
</tr>
<tr>
<td>5</td>
<td>1000</td>
<td>0.21</td>
<td>0.18</td>
<td>1.17</td>
<td>0.20</td>
<td>2.71</td>
<td>0.56</td>
</tr>
<tr>
<td>10</td>
<td>1000</td>
<td>0.74</td>
<td>0.74</td>
<td>0.25</td>
<td>0.19</td>
<td>0.79</td>
<td>0.58</td>
</tr>
<tr>
<td>20</td>
<td>1000</td>
<td>1.42</td>
<td>2.92</td>
<td>0.14</td>
<td>0.41</td>
<td>0.65</td>
<td>0.92</td>
</tr>
<tr>
<td>50</td>
<td>1000</td>
<td>4.92</td>
<td>10.72</td>
<td>0.11</td>
<td>1.20</td>
<td>0.61</td>
<td>2.98</td>
</tr>
<tr>
<td>5</td>
<td>10000</td>
<td>0.10</td>
<td>0.12</td>
<td>1.19</td>
<td>0.15</td>
<td>2.21</td>
<td>0.22</td>
</tr>
<tr>
<td>10</td>
<td>10000</td>
<td>0.81</td>
<td>1.29</td>
<td>0.08</td>
<td>0.11</td>
<td>0.25</td>
<td>0.20</td>
</tr>
<tr>
<td>20</td>
<td>10000</td>
<td>1.32</td>
<td>3.83</td>
<td>0.05</td>
<td>0.17</td>
<td>0.20</td>
<td>0.27</td>
</tr>
<tr>
<td>50</td>
<td>10000</td>
<td>6.02</td>
<td>18.90</td>
<td>0.04</td>
<td>0.67</td>
<td>0.19</td>
<td>1.16</td>
</tr>
<tr>
<td>25-HCC class</td>
<td>Transplants</td>
<td>Controls</td>
<td>p-value</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
<td>----------</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 13</td>
<td>35</td>
<td>37.10</td>
<td>0.0451</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[13 − 20)</td>
<td>37.2</td>
<td>43.50</td>
<td>0.03260</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[20, −26)</td>
<td>45</td>
<td>50.60</td>
<td>0.48900</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[26 − 34.5)</td>
<td>44.85</td>
<td>52</td>
<td>0.003420</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>⩾ 34.5</td>
<td>48</td>
<td>59.70</td>
<td>0.01720</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Class-specific median of 1,25-DHCC and p-value for comparing transplants and controls within that class.