Summary and discussion of: "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing"

Statistics Journal Club, 36-825

Beau Dabbs and Philipp Burckhardt

9-19-2014

1 Paper Overview

1.1 Motivating Example

In hypothesis testing, the multiplicity problem occurs when performing a large number of hypotheses tests simultaneously. With moderately sized data sets, it might be possible to gloss over this issue, yet in an era increasingly characterized by massive data sets, this is no longer possible.

In genetics, DNA microarray experiments are used to gain a better understanding of the causes and effects of diseases by investigating changes in gene expression for thousands of genes. In such a microarray experiment, one might easily perform $10,000$ tests. Using a standard p-value of 0.05, one would expect 500 genes to be deemed significant by chance even if there was no effect at all.

To correct hypothesis testing procedures under these circumstances, several methods have been proposed, some based on the concept of the *False Discovery Rate (FDR)*, which forms the main object of interest for the rest of this summary paper. We start our discussion by introducing necessary notation and quickly reviewing the classical approaches to deal with the multiplicity problem. Then, the definition of FDR is introduced and the original method outlined in[1] is explained. Finally, simulation results are presented and points brought up in the in-class discussion are summarized.

1.2 Definition of Terms

To make precise statements about the problem at hand, we will have to introduce some notation. Let us denote with $H_1, \ldots, H_m$ the m hypotheses we are evaluating. Of these m_0 hypotheses are assumed to be true nulls, i.e. hypotheses for which the null hypothesis is true.

For the i-th hypothesis test, we declare a significant finding if the corresponding test statistic falls into the rejection region or its p-value p_i is smaller than previously set significance level α. In these cases, we reject hypothesis H_i. Let R_i be the associated indicator random variable which takes the value one if the i-th hypothesis is rejected and zero otherwise. Then we can define the total number of rejections as $R = \sum_{i=1}^{m} R_i$. Besides this

	True null	Non-true null	Total
Declared significant	V	S	R
Declared non-significant	U	T	$m-R$
Total	m_0	$m-m_0$	m

Table 1: Notation for a setting with m hypothesis tests. Of the defined variables, only m and R are actually observed, with the rest bing unknown. V is equal the number of *false positives*, i.e. the count of how many of the m_0 true null hypotheses were erroneously rejected, whereas R is the total number of significant results out of the m hypotheses.

observed quantity, we define several other unobserved ones, which are all displayed in Table **??**. Of these, V is the total number of false positives, hypotheses for which we decide to reject the null even though it is in fact true. We again define V as the sum of indicator functions, namely $V = \sum_{i=1}^{m} V_i$, where $V_i = 1$ if we decide to reject H_i while $R_i = 1$. Otherwise, $V_i = 0$.

Although traditionally much focus in statistics is devoted to control the number of false positives (*Type 1 errors*), it is likewise important to be able to detect true effects. We denote with T the number of committed *Type 2 errors* (cases for which the null hypotheses is false but our decision rule does not yield a significant result). The power of a statistical test is defined as the probability of correctly rejecting a false null hypothesis:

$$\text{power} = \mathbb{P}\big(\text{reject null hypothesis}\big|\text{null hypothesis is false}\big)$$

In our reported simulations (where all quantities from Table **??** are known), the power is estimated as power $= \frac{S}{S+T}$. Having introduced these quantities, we are ready of a formal analysis of the multiple testing problem and a discussion of the methods proposed to alleviate the problem.

Without taking the multiplicity of the testing procedures into account, we would fix $P(V_i = 1) = \mathbb{E}[V_i]$ at level α, which ensures that the probability of committing a type 1 error is equal to α for each hypothesis individually. However, this ignores the multiplicity problem, as we only control the per-comparison error rate (PCER) and not the family-wise error rate (FWER), which is defined as the probability of obtaining at least one false positive: $P(V \geq 1)$. To see why, observe that the expected number of false positives is

$$\mathbb{E}[V] = \sum_{i=1}^{m_o} \mathbb{E}[V_i] = m_0\alpha \leq m\alpha,$$

where the sum is only over all m_0 true nulls as V_i is zero automatically for non-true hypotheses. Now,

$$\begin{aligned}\mathbb{E}[V] &= 0P(V=0) + 1P(V=1) + 2P(V=2) + \ldots + m_oP(V=m_0)\\ &\geq P(V=1) + P(V=2) + \ldots + P(V=m_0)\\ &= P(V \geq 1),\end{aligned}$$

showing that the expected number of false positives provides an upper bound of the FWER. Alternatively, as V is non-negative, use Markov's inequality to conclude that $P(V \geq 1) \leq \mathbb{E}[V]$.

1.3 Review of Previous Methods

1.3.1 Bonferroni Correction

The results of the previous section suggest an easy fix which allows control of the FWER in a multiple testing situation. If we pick $\alpha^\star = \frac{\alpha}{m}$ as the significance level for each individual test, we get $\mathbb{E}[V] = m_0\alpha^\star \leq m\alpha^\star = \alpha$ and hence we have bounded the FWER at α because $P(V \geq 1) \leq \alpha$. We say that the Bonferroni correction controls the FWER *in the strong sense* as the FWER is bounded at α no matter how many of the underlying hypotheses are true or false.

One major disadvantage of using the Bonferroni procedure is that it comes with a heavy price tag, namely a substantial loss in statistical power compared to the per-comparison testing procedures. The method described next does strictly better in this regard.

1.3.2 Hochberg Procedure

There is an alternative method by Hochberg (1988)[2] for controlling the FWER. Given our set of m hypotheses to be tested, H_i, and their corresponding p-values, p_i, we first sort them to obtain the ordered p-values $p_{(1)} \geq p_{(2)} \geq \cdots \geq p_{(m)}$. We then determine the largest k such that:

$$p_(k) \leq \frac{\alpha}{m - k + 1}$$

If we then reject the hypotheses $H_(j)$ for each $j \leq k$, it turns out that we will still have $P(V \geq 1) \leq \alpha$. This statement holds no matter how many of the null hypotheses are actually true. This procedure rejects strictly more of the hypotheses being tested than the bonferroni procedure in every case except when only a single hypothesis is rejected, and thus has a gain in power. This can be seen by noting that $\alpha/m < \alpha/m - k + 1$ for all $k > 1$. However, in practice, the gain in power is relatively small.

1.4 False Discovery Rate

The two methods above show that it is possible to control the FWER, but that it's hard to do this and retain much power. In some situations, guaranteeing a low chance of making any errors about the true nulls isn't as important. For instance, when you are testing for multiple positive outcomes from a single drug. If a single one of your detected outcomes turns out to be false, this isn't a huge penalty as long as some of the null hypotheses are false. Thus, it makes sense to instead control the percentage of discoveries that were actually null. Formally we will control:

$$Q_e = E[Q] = E[V/R]$$

We also adopt the convention that if $R = 0$, and thus there are no significant discoveries, that $Q = 0$. If it is possible to gain power while still controlling the FDR, this would be

advantageous in many common scenarios. The main theorem of this paper shows that it is in fact possible to control the False Discovery Rate (FDR) at some level q by using a procedure similar to the Hochberg procedure above. Again, we consider the sorted p-values and find the largest k such that:

$$p_(k) \leq \frac{k}{m} q$$

We then reject all ordered hypotheses with $j \leq k$. It turns out that this method does control the FWER when all of the null hypotheses are true, but has no guarantee when $m_0 < m$. The theorem does, however, require that the p-values under the null hypothesis are independent and uniformly distributed between zero and one. The theorem places no requirements, however, on the p-values under the alternative hypotheses, and is proven by the following lemma.

Lemma. For any $0 \leq m_0 \leq m$ independent p-values corresponding to true null hypotheses, and for any values the $m_1 = m - m_0$ p-values corresponding to false null hypotheses can take, the multiple-testing procedure above satisfies:

$$E(Q|P_{m_0+1} = p_1, .., P_m = p_{m_1}) \leq \frac{m_0}{m} q.$$

Then, no matter what the distribution is over the false null hypotheses, we can integrate over their joint distribution to obtain the bound:

$$E(Q) \leq \frac{m_0}{m} q \leq q.$$

Thus, our bound is tightest when all null hypotheses are true. To see that we are gaining in power, even over the Hochberg procedure, notice that we now have $\frac{1}{m-k+1} \leq \frac{k}{m}$ for all k. It turns out that the inequality is strict for all k away from $k = m$ and $k = 1$. To see this difference graphically consider figure 1. The gap between the rejection criterion can be quite large, and so it's not surprising that in practice a large gain in power can be seen.

2 Examples

2.1 Shiny App

We have developed a Shiny application to illustrate the properties of the FDR and other classical procedures like Bonferroni on simulated data. The application can be accessed under the following URL: `https://philipp.shinyapps.io/FalseDiscoveryRate/`. Using a simple GUI, it is possible to vary the parameters of the simulation: The number of hypotheses can be set, the percentage of true nulls as well as the maximal true effect under the alternative hypothesis, $\mu_1^{\max}$. Data are drawn from $N(0, 1)$ under the null and $N(\mu_i, 1)$ under the alternative hypothesis of $\mu_i \neq 0$, where μ_i is a value between zero and the chosen maximal true effect. To be more precise, the non-zero expectations were binned into three groups at values $\frac{1}{4}\mu_1^{\max}$, $\frac{1}{2}\mu_1^{\max}$, $\frac{3}{4}\mu_1^{\max}$, and $\mu_1^{\max}$, where the number of hypotheses in each bin is either equal, linearly decreasing or linearly increasing. This setting is almost identical to the simulation experiment described in.[1]

Results show that both the Hochberg method and the Bonferroni procedure take a big hit in terms of power when compared with the naive method of discounting the multiplicity

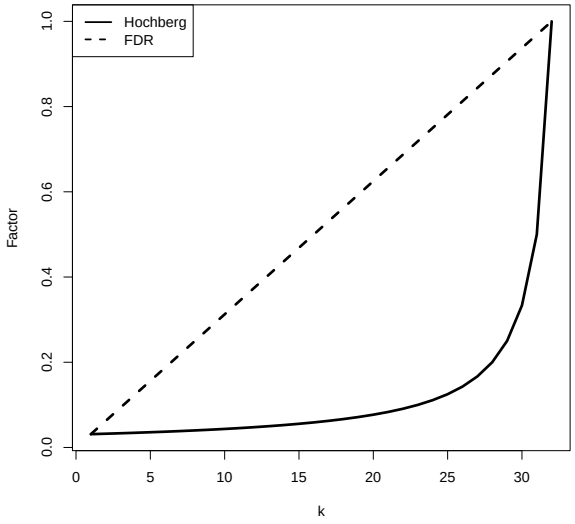

Figure 1: Multiplicative factor for Hochberg and FDR procedures

problem and the FDR procedure. However, the naive method does very poorly when evaluated in terms of the FWER: With as few as 50 simultaneous hypotheses, it is quite likely to obtain a FWER around 0.75 in this case.

2.2 Ryan's Example

This simulation study (code available on the course website) by Ryan Tibshirani demonstrates the effect of using different kind of pre-screening procedures before hypothesis testing using FDR. Throwing away observations is often not a good idea, but one reason might be the need to decrease the number of comparisons and thus increase the power of the testing procedure.

In the example provided by Ryan Tibshirani, we measure m genes for n healthy and n sick patients, with the question to identify the genes which are associated with the patients being either sick or healthy. Assume that for each gene i, X_{ij} measures the gene expression level for the j-th healthy person. We assume that the gene expression levels of gene i for the n healthy persons are independently drawn from a normal distribution

$$X_{i1}, \ldots, X_{in} \sim N\left(\mu_i, \sigma^2\right),$$

and likewise that for the n sick persons, we have that

$$Y_{i1}, \ldots, Y_{in} \sim N\left(\theta_i, \sigma^2\right).$$

The hypotheses in question are $H_i : \mu_i = \theta_i$, which are tested against the alternative that the expression level of gene i differs among the sick and healthy. Two screening procedures are considered: The first one being based on $BV_i = \left(\bar{X}_i - \bar{Y}_i\right)^2$, where $\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$

and $\bar{Y}_i = \sum_{j=1}^{n} Y_{ij}$. Using this procedure, we only keep the half of the observations with the highest BV_i score and discard all others. The second screening procedure is based on the total variance: $TV_i = \frac{1}{2n}\left(\sum_{j=1}^{n}(X_{ij} - M_i)^2 + \sum_{j=1}^{n}(Y_{ij} - M_i)^2\right)$, where M_i is the mean of the joint sample $X_1, \ldots, X_n, Y_1, \ldots, Y_n$. Figure 2 shows the estimated and true FDRs under both screening regimes.

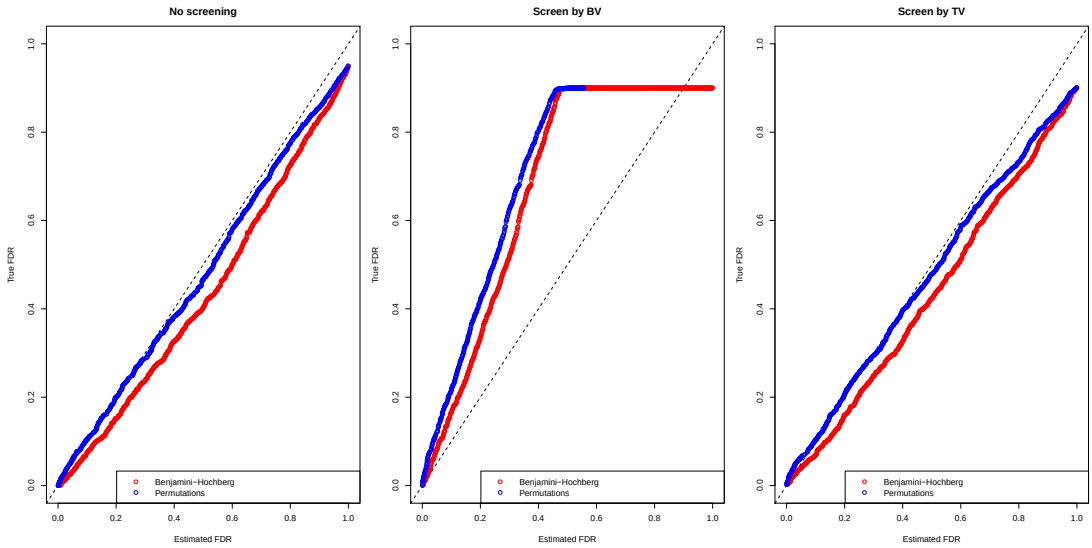

Figure 2: Estimated versus true FDR under different screening procedures.

As we can see, for the situation without screening and screening based on the total variance, the red and blue line lie which correspond to results obtained by the Benjamini-Hochberg procedure and a permutation test roughly lie on the 45-degree line, suggesting agreement between the true and estimated FDR. However, when basing the screening on BV_i, the result looks strikingly different: One heavily underestimates the true FDR in this case.

What is going on here?

Note that our inference results after screening will only continue to be valid if the screening procedure does not alter the distribution of the test statistics and consequently the p-values. As shown in the supplementary note of Ryan Tibshirani, this is the case for TV_i but not for BV_i. The main takeaway from this is that if one decides to pre-screen observations, one should decide upon a screening criterion which is independent of the test statistic.

2.3 Mean Comparison Example (Beau)

One multiple testing problem that is relatively common is the comparison of means. Suppose we have n variables $Y_1, ..., Y_n$ each with mean μ_i and common variance known to be $\sigma^2 = 1$. Since each variable is used in multiple tests, this means that there will be dependence between the tests. We considered one simulation in class where we had 8 variables generated

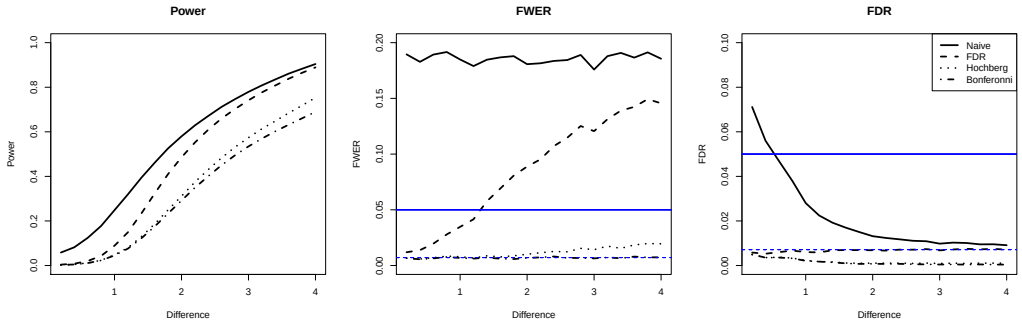

Figure 3: Power, FWER, and FDR curves for comparison of 8 means. All tests were at level $\alpha = q = 0.05$. The solid blue line represents the 0.05 level in both plots, and the dashed blue line represesnts $\frac{0.05}{7}$.

from a normal distribution with:

$$\mu_1 = \mu_2 = L \; ; \; \mu_3 = \mu_4 = 2L \; ; \; \mu_5 = \mu_6 = 3L \; ; \; \mu_7 = \mu_8 = 4L$$

Thus we have 28 tests of difference of mean, 4 of which are true under the null hypothesis of no difference. Figure 2.3 shows the resulting power, FWER, and FDR rates as we vary the magnitude L. In this simulation our procedures do provide the correct guarantees. One thing to note is that the bonferroni procedure actually controls the familywise error rate at $\alpha \cdot \frac{m_0}{m}$ which in our case happens to be $\frac{\alpha}{7}$. Since the bonferonni correction makes no assumptions about independence, it's not surprising that it is still controlled at the correct rate when we know the proportion of true hypotheses. Similarly the FDR procedure should actually control the FDR at level $\frac{q}{7}$ and this control appears to be fairly tight. The reason that we still have control despite dependence among some of the p-values is that the null hypotheses actually are independent, even though some of the tests that have false null hypotheses are dependent.

After the class discussion we did a further simulation with 3 levels for the means, $L, 2L$, and $3L$. We had three variables for each level resulting in 9 variables total. This new test induces dependence among the null hypotheses as well. Figure 2.3 shows the results for this new simulation. Again we also plotted the 0.05 and $0.05 \cdot \frac{m_0}{m}$ levels for the FWER and FDR. The bonferroni procedure still controls the FWER at the expected level, while our FDR procedure also controls the FDR at the expected level, despite the dependence.

3 Discussion

3.1 Balancing sample size and FDR

During the discussion, Rob Tibshirani gave offered some suggestions for usage of the FDR in practice. Instead of picking a single q in the Benjamini-Hochberg procedure, he advocated

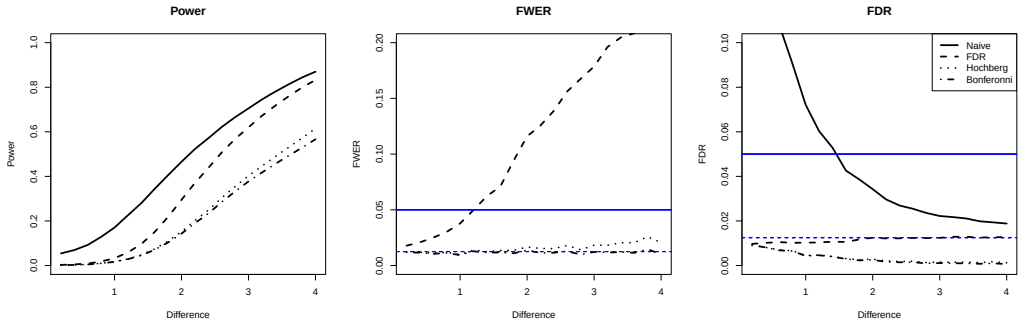

Figure 4: Power, FWER, and FDR curves for comparison of 9 means. All tests were at level $\alpha = q = 0.05$. The solid blue line represents the 0.05 level in both plots, and the dashed blue line represesnts $\frac{0.05}{4}$.

varying it and obtaining a range of results, differing in the number of rejected and kept hypotheses. Based on the research question or domain knowledge of the researchers, one can then decide upon which of the results is deemed most reasonable.

3.2 Alternate Versions of FDR

A few alternate version of FDR exist, one of the most common ones is the positive false discovery rate (pFDR) which can be found in the Benjamini, Hochberg paper[1] and is elaborated on in the paper by Storey (2002).[3] This is simply the expected false discovery rate conditional upon rejecting at least one hypothesis,

$$pFDR = E(V/R|R > 0)$$

For a small number of hypotheses this difference could be important, but in general this difference is inconsequential. If we consider our rejection region to be a set of p-values of the form $[0, \gamma]$, which is the form of most rejection procedures, it turns out that we can represent the pFDR of a procedure with rejection level γ as,

$$pFDR(\gamma) = \frac{\pi_0 Pr(P \leq \gamma|H = 0)}{Pr(P \leq \gamma)} = \frac{\pi_0 \gamma}{Pr(P \leq \gamma)},$$

where $\pi_0 = m_0/m$. The only difficulty with this formula is that we need to estimate π_0 and $Pr(P \leq \gamma)$. However it turns out a decent estimate of the number of null hypotheses in the sample is the number of p-values in the tail of the distribution, divided by the width of the tail,

$$\hat{m_0}(\lambda) = \frac{\#\{p_i > \lambda\}}{(1-\lambda)} = \frac{W(\lambda)}{(1-\lambda)}$$

Giving us an estimate for $\hat{\pi}(\lambda) = \hat{m_0}(\lambda)/m$. Similarly we can estimate the probaiblity of a p-value being less than γ by our observed proportion $R(\gamma)/m$. These two estimates give us a relatively simple formula for pFDR as a function of λ and γ,

$$\hat{Q}_\lambda(\gamma) = \frac{W(\lambda)\gamma}{(1-\lambda)R(\gamma)}.$$

References

[1] Benjamini, Yoav, and Yosef Hochberg. *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society. Series B (Methodological) (1995): 289-300.

[2] Hochberg, Yosef. *A sharper Bonferroni procedure for multiple tests of significance.* Biometrika 75.4 (1988): 800-802.

[3] Storey, John D. *A direct approach to false discovery rates.* Journal of the Royal Statistical Society: Series B (Statistical Methodology) , 64.3 (2002): 479-498.