A Non-Parametric Maximum for Reasonable Number of Rejected Hypotheses: Objective Optima for False Discovery Rate and Significance Threshold in Exploratory Research with Application to Ordinal Survey Analysis

Amir H. Ghaseminejad

Simon Fraser University, School of Communication

Simon Fraser University, Institute for Canadian Urban Research Studies

Capilano University, School of Business

Vancouver, Canada

ahghasem@sfu.ca

24 March, 2017

Summary. This paper identifies a criterion for choosing the largest set of rejected hypotheses in high-dimensional data analysis where Multiple Hypothesis testing is used in exploratory research to identify significant associations among many variables. The method neither requires predetermined thresholds for level of significance, nor uses presumed thresholds for false discovery rate. The upper limit for number of rejected hypotheses is determined by finding maximum difference between expected true hypotheses and false hypotheses among all possible sets of rejected hypotheses. Methods of choosing a reasonable number of rejected hypotheses and application to non-parametric analysis of ordinal survey data are presented.

Keywords: High-dimensional data analysis, Multiple hypothesis testing, False discovery rate, Optimum significance threshold, Maximum for reasonable number of rejected hypotheses, Big data analysis

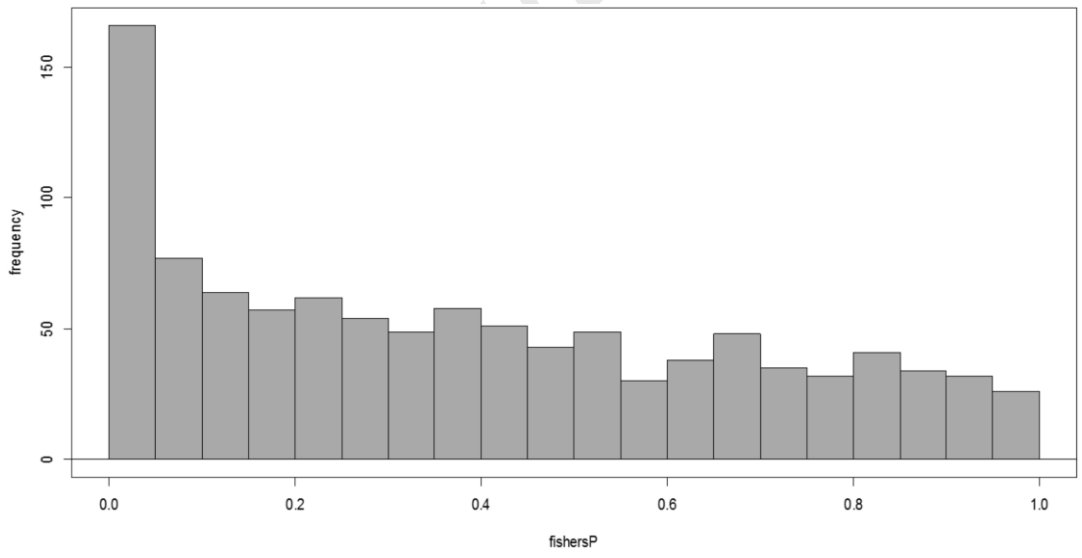
1. Introduction

In high-dimensional data analysis, multiple simultaneous hypothesis testing arises because we need to identify which null hypotheses, among many, should be reasonably rejected (Neuvial, 2013). A significant finding (discovery) is a hypothesis that is rejected based on statistical evidence.

When performing multiple hypothesis testing, as the number of hypotheses being tested (m) gets bigger and bigger, using a p-value threshold (alpha), such as .05, for rejecting hypothesis based on p-values becomes problematic. P-value is a measure of the probability of a rejected hypothesis to be a false positive. When number of hypothesis being tested is big, for example 1000, the expected number of false positives is (m*alpha). If alpha is .05 this means that the expected number of false positives among significant findings is less than or equal to 50.

It is anticipated that, when there are no expected true discoveries, the frequency distribution of p-values to be uniform. Which means that the proportion of tests resulting a p-value in any class should be the same. In many situations, "it is reasonable to assume that larger p-values are more likely to correspond to true null hypotheses than smaller ones" (Neuvial 2013, 1428) or smaller p-values are less likely to correspond to true null hypotheses. In many research situations, p-values have a frequency distribution like figure 1, where number of hypotheses with very low p-values are more than other p-values.

Figure 1, frequency distribution of p-values many expected true discoveries



The first column in figure 1 is presenting hypotheses with p-value<.05. These the hypotheses that we might be inclined to declare as rejected hypothesis or significant discoveries; however, the expected number of false positives in this subset can be very high. In other words, many of rejected hypothesis may be true nulls. Therefore, we may want to reduce our threshold (alpha) for rejection so fewer but more significant hypotheses

are rejected. Reduction of alpha, decreases the chance of false positives in our discovery set and thus leads to a smaller chance of false discoveries. Unfortunately, this may increase the false negatives. By decreasing Alpha, we are accepting to have more false negatives.

False discovery rate (FDR) is defined as expected number of false discoveries (false positives among rejected hypotheses) divided by total number of rejected hypotheses (Neuvial 2013, 2). In many situations, a p-value of .05 may lead to a big FDR. Several algorithms have been proposed to consider FDR in the process of selecting significant findings.

Bonferroni has suggested to consider a hypothesis significant when alpha is less than or equal to alpha/m (Storey and Tibshirani, 2003,2). By choosing a rejection threshold much lower than alpha, the probability of making one or more false discoveries will be less than or equal to alpha.

Bonferroni's suggestion guaranties a family-wise error rate (FWER) less than or equal to alpha; but this conservative measure can result in many false negatives. When the number of significant hypotheses are few this measure is appropriate; because even expectation of one false positive in the result set is damaging. In many studies, where number of significant findings are many, the researcher may be able to afford a higher FDR, if that will prevent many false negatives. Not detecting many important associations may be more harmful than probability of one false negative among many significant findings.

Many adaptive hypothesis testing procedures rely on estimates of the proportion of true null hypotheses in the initial pool using plugins, a single step, in multiple steps, or asymptotically (Blanchard and Roquain, 2009). Plug-in procedures use an estimate of the proportion of true null hypotheses (Neuvial, 2013). Thresholding-based multiple testing procedures, reject hypotheses with p-values less than a threshold (Neuvial, 2013).

Storey and Tibshirani (2003) have proposed a strategy that assigns each hypothesis an individual measure of significance in terms of expected false discovery rate called q-value. Most q-value based strategies rely on some estimate of the proportion of true null hypotheses. However, the choice of the threshold of q-values at which the researcher draws the line of significance remains subjective.

Storey (2007) has argued that two steps that are involved in any multiple-testing procedure. The first step is "determining the order in which the tests should be called significant" by "ranking the tests from most to least significant". The second step is "choosing an appropriate significance cut-off somewhere along this ordering". Storey focuses on performing the first step optimally, given a certain significance framework for the second step. He cites (Shaffer, 1995) defining the goal: "to estimate the appropriate cut-off to obtain a particular error rate, usually based on the familywise error rate or false discovery rate". Storey (2007) proposes an optimal discovery procedure based on maximizing Expected True Positives (ETP) for each Expected False Positive (EFP) among all Single Thresholding Procedures (STP).

Norris and Kahn (2006) have proposed balanced probability analysis (BPA) based on three variables: (i) The total number of true positives (TTP); (ii) The false discovery rate (FDR), defined as the aggregate chance that a true null hypothesis is rejected by statistical accident. (iii) The false negative rate (FNR), defined as the number of hypothesis that should truly be rejected but are missing from the significance list divided by the total number of hypothesis that should truly be rejected. They believe other definitions of type 2

error rates, such as the false nondiscovery rate (the ratio of hypotheses that should truly be rejected but are nondiscovered to the number of unrejected Hypothesis) "are difficult to intuit for the nonstatistician". They "calculate the FNR directly from the data, by using resampling to estimate the null and alternate distributions". Their "procedure weakly depends on the estimated FDR, and requires one model-dependent step to optimize a single parameter".

As Norris and Kahn (2006) have argued, the true FDR can be accurately determined only when the TTP is known. They used an adaptation of the algorithm by Storey and Tibshirani (2003) they estimate the TTPs. They estimated FDR and then they estimated FNR based on their estimates of FDR and TTP.

2. A Non-Parametric Maximum for Reasonable Number of Rejected Hypotheses

This article, is concerned about the second step mentioned above, "choosing an appropriate significance cut-off somewhere along this ordering", but we won't need to know or estimate the total number of true positives or total number of true Negatives.

Although Setting a subjective threshold for FDR (such as 0.05) can relax the extremely conservative suggestion by Bonferroni it can be a limitation which may unnecessarily limit the number reasonable findings a researcher should report. **I will show that, in some situations, grounded on observed data one can identify an objective upper bound for "level of significance and FDR" that is reasonable for the researcher to report.**

When we tabulate the p-values resulted from a study into sorted classes (from smallest to largest p-value), we will have the frequency of each observed p-value. We have a special interest in the set of smallest p-values; thus, the first class is the most valuable class for us. All the P-values with a value closest to Zero (or zero if such hypotheses exist) in set S_1 which will have will have f_1 members (f_1>=1).

The next smallest p-value will be p_2. Set 2, will contain all the hypotheses with a value of p_2. S_2 will have f_2 members (f_2>=1). For each one of k observed p-values there will be corresponding frequency and a set of hypotheses.

Total number of hypotheses tested= $N = \sum_{i=1}^{k}(f_i)$

In the equation above, f_i is the frequency of hypotheses in set Si. If we set the Alpha (rejection threshold) at p_1. We will have f_a rejected Hypotheses, of which $p_1 \times N$ are *expected to be false Discoveries* (EFD_1).

$$EFD_1 = p_1 * N$$

Therefore, from the first set we expect to have:

$$ETD_1 = f_1 - (p_1 * N)$$

ETD_1 is expected true discoveries if we reject hypotheses with p-value less than or equal to p_1. We may be interested in including the set of f_2 hypothesis S_2 in our discoveries, but the p-value of these Hypotheses is P2 and the expected false discoveries in rejected set S_1 and S_2 will be p_2*N.

$$R_2 = S_1 \cup S_2$$

p_2*N is always bigger than p_1*N. p_2*N will be the Cumulative Expected False Discoveries (CEFD) in R_2:

$$CEFD_2 = p_2 * N$$

Therefore, from the first two sets we expect to have Cumulative Expected False Discoveries (CETD) in R_2 as:

$$CETD_2 = (f_1 + f_2) - (p_2 * N)$$

Therefore, cumulative expected true discoveries $CETD_2$ from S_1 and S_2, will be bigger than ETD_1. The series of expected false discoveries in each set: EFD_1, EFD_2, EFD_3, …… is usually increasing because the p-values are getting bigger. And the series of expected true discoveries is each set: ETD_1, $CETD_2$, $CETD_3$, …. is usually increasing in the first sets. But because p-values are increasing and by adding each set to rejected set we are in fact increasing our Alpha, The proportion of false discoveries added by set S_j (j>i) to R_j is more than the contribution of false discoveries in set Si to Ri and contribution of true discoveries in from S_j to R_j is more than the contribution of true discoveries by S_i to R_i. When i goes toward N, pi goes toward 1

$$\lim_{i \to N} p_i = 1$$

$$\lim_{i \to N} CEFD_i = \lim_{i \to N} N * p_i = N$$

$$\lim_{i \to N} CETD_i = \lim_{i \to N} (R_i - CEFD_i) = 0$$

If we define delta:

$$\delta_i = CETD_i - CEFD_i$$

At some point δ_i must start to decrease and must have a maximum. The maximum number of rejected hypotheses happens at set S_{max} after which adding the hypotheses in the next set S_{max+1} (setting alpha at p_{max+1}) will contribute more to false discoveries than to true discoveries.

$$R_{max} = S_1 \cup S_2 \cup S_3 \cup \ldots \cup S_{max}$$

R_{max} is the largest set of rejected hypothesis that is reasonable to be reported. The largest alpha that is reasonable to be the threshold for rejecting hypotheses is P_{max}. FDR_{max} is the biggest reasonable FDR to be reported.

$$FDR_{max} = \frac{CEFD_{max}}{\sum_1^m f_i} = \frac{p_{max} \times N}{\sum_1^m f_i}$$

That is the point at which we have no incentive to add the set S_{max+1} to our discoveries. If we add set S_{max+1} to our set of rejected hypotheses, the difference between CETD and CEFD (δ) will start to decline. δ_{max} is an objective upper bound for the number of hypothesis we reject. If Maximum δ_{max} happens when we add S_{max} to set of rejected hypotheses, we have decided that the threshold alpha for rejecting null hypotheses is p_{max}, we will reject hypothesis with p-value<=p_{max}.

If we have k observed p-values $p_1 \leq p_2 \leq p_3 \leq \ldots \leq p_k$, related to sets of tested hypotheses S_1, S_2, S_3, … , S_k; δ_{max} happens when we add set S_m to our rejected hypotheses.

The number of rejected hypotheses, at level Alpha will be p_{max}, and the biggest reasonable set of rejected hypotheses R_{max} will be:

$$\mathrm{R}_{max} = \sum_{1}^{m} f_i$$

Maximum ECTD can be calculated based on the following formula:

$$\delta_{max} = Max(\mathrm{CETD}_i - \mathrm{CEFD}_i)$$

$$\delta_{max} = Max\left(\sum_{i=1}^{k}(f_i - CEFD_i) - \sum_{i=1}^{k} CEFD_i\right)$$

Table 1, summarizes what we discussed above. Notice that the upper limit for number of rejected hypotheses is determined based on maximization of difference between expected true hypotheses and False Hypotheses. Alpha is reported (not assumed) and is not subjectively selected. The FDR is dictated by data. If the researcher decides to add more sets to discoveries, he/she is accepting the cost of adding more false discoveries than true discoveries to the set of rejected hypotheses.

Table 1, CEFD , CETD and δ

Set of observations	Observed p-value In the set	Observed frequency of p-value	Set of rejected hypotheses	Cumulative Expected False Discoveries if set is rejected (CEFD)	Cumulative Expected TRUE Discoveries if set is rejected (CETD)	**Find the maximum in this column** δ=CETD-CEFD
S_1	p_1	f_1	$R_1 = S_1$	$N \times p_1$	$f_1 - N \times p_1$	$f_1 - N \times p_1 - N \times p_1$
S_2	p_2	f_2	$R_2 = S_1 \cup S_2$	$N \times p_2$	$f_1 + f_2 - N \times p_2$	$\sum_{i=1}^{2}(f_i) - 2 \times N \times p_2$
S_3	p_3	f_3	$R_3 = S_1 \cup S_2 \cup S_3$	$N \times p_3$	$f_1 + f_2 + f_3 - N \times p_3$	$\sum_{i=1}^{3}(f_i) - 2 \times N \times p_3$
…	…	…	…	…	…	
S_k	p_k	f_k	$R_k = S_1 \cup S_2 \cup S_3 \cup \ldots \cup S_k$	1	0	-N
	$\sum_{i=1}^{k}(p_i) = 1$	$\sum_{i=1}^{k}(f_i)$=N				

3. Objective Optima for False Discovery Rate and Significance Threshold

Making the set of Rejected hypotheses beyond R_{max} may increase CETD, but it will increase the CEFD even more; it will decrease the quality of discovery measured as δ. At R_{max} however, we don't have a sharp or sudden decrease of δ. Delta usually changes relatives slowly around R_{max} . We have a peak and a slow reversal in trend for δ. The researcher can use different ways of piecewise regression to identify an optimum number of rejected hypotheses much less than R_{max} but much more than $R_{FDR=0.05}$.

For example, piecewise regression of the p-values of hypotheses in sets S1 to S_{max} , and number of observations in R_1 to R_{max} , with one breakpoint can model the observations with two line segments. The breakpoint, where the slope of the two lines changes, is were the efficiency of adding more hypotheses to R changes. It is an objective threshold at which rejected hypotheses are less than R_{max}, while number of CETD is close to true discoveries at R_{max} resulting in a better FDR with little loss of CETD. Therefore, the number of rejected hypotheses at the break point, R_{bp} ,is an optimal number of hypotheses. It doesn't decrease the quality of our discovery, measured by δ very much.

A more computationally intensive piecewise regression of the p-values of hypotheses in sets S1 to $S_{max+\varepsilon}$ can be conducted such that the second segment is a horizontal line close to the point (R_{max}, p_{max}). The horizontal line can also be the one that passes the point (R_{max}, p_{max}). The resulting set of discoveries is not very sensitive to the selection of piecewise regression method.

4. Example: application of method to non-parametric analysis of survey data

Table 1 shows the sorted results of using the procedure when analyzing a large survey regarding "variables influencing citizen engagement in mediated democracies". Fisher's Exact test was used to check the significance of associations observed in cross-tabulated data, and Sommer's D was used to measure the extent of association. The null hypothesis was that the observed association in crosstabulation is accidental.

For one of the outcome variables, 1031 hypotheses regarding crosstabulations were tested. If we rely on 0.05 rule of thumb for alpha, too many hypotheses will be falsely rejected. If we rely on 0.05 rule of thumb for FDR, many potentially significant findings, may falsely remain unrejected. Notice that Bonferroni's correction for p-value=0.05 would suggest a threshold of rejection of 0.0000485 which means we can conservatively reject 16 hypotheses at FDR 0.002414.

We observe 7 Hypotheses with p-value of 0 in set S_1 which will be obviously rejected. If we decide to reject the hypothesis in the second set at p-value=0.000001, we will add 1 hypothesis to the set of rejected hypotheses. The single hypothesis that can be rejected contributes 0.998956 to the total expected true discoveries. Cumulative expected False discoveries will be 1031*0.000131=001044. Rejecting the hypotheses in sets S_1 and S_2, we are in fact declaring the threshold alpha is 0.000001, cumulative expected False discoveries will be 1031*0.000131=001044, FDR will be 0.000131.

We may be interested to reject more hypotheses in next sets. If we reject all the hypotheses in sets S_1 to S_{36}, we will have 42 hypotheses in our set of rejected hypotheses R_{36}. FDR will grow to 0.048322. Like many researchers who will not reject set S37, we can define our alpha to be 0.001944. In other words, we reject hypotheses with p-value less

than 0.001944. This is more powerful than Bonferroni's correction. But we expect 2.029536 of 42 discoveries to be false.

Our set S177 has the 184th p-value at 0.0393. The resulting set of rejected hypotheses from S1 to S177 is expected to have 41.0292 to be false discoveries and 142.9708 true discoveries. The expected false discovery rate as the result of increasing Alpha to 0.0393 will be 0.222985.

Table 1, Frequency distribution of p-values from 1031 tested hypotheses

Set	Frequency in set	Less than cumulative frequency hypotheses in rejected sets	p-value of the set (alpha if this set is rejected)	Cumulative Expected Number of False Discoveries CEFD if this set is rejected	Expected False Discovery contribution of this set	FDR	Expected Cumulative Number of True Discoveries CETD if this set is rejected	Expected True Discovery contribution of this set	δ = CETD-CEFD
1	7	7	0	0	0	0	7	7	7
2	1	8	0.000001	0.001044	0.001044	0.000131	7.998956	0.998956	7.997912
3	1	9	0.000005	0.00522	0.004176	0.00058	8.99478	0.995824	8.98956
4	1	10	0.000008	0.008352	0.003132	0.000835	9.991648	0.996868	9.983296
5	1	11	0.000009	0.009396	0.001044	0.000854	10.9906	0.998956	10.98121
6	1	12	0.000015	0.01566	0.006264	0.001305	11.98434	0.993736	11.96868
7	1	13	0.000021	0.021924	0.006264	0.001686	12.97808	0.993736	12.95615
8	1	14	0.000022	0.022968	0.001044	0.001641	13.97703	0.998956	13.95406
9	1	15	0.000034	0.035496	0.012528	0.002366	14.9645	0.987472	14.92901
10	1	16	0.000037	0.038628	0.003132	0.002414	15.96137	0.996868	15.92274
11	1	17	0.0001	0.1044	0.065772	0.006141	16.8956	0.934228	16.7912
…	…	…	…	…	…	…	…	…	…
34	1	40	0.00183	1.91052	0.061596	0.047763	38.08948	0.938404	36.17896
35	1	41	0.001841	1.922004	0.011484	0.046878	39.078	0.988516	37.15599
36	1	42	0.001944	2.029536	0.107532	0.048322	39.97046	0.892468	37.94093
37	1	43	0.002081	2.172564	0.143028	0.050525	40.82744	0.856972	38.65487
38	1	44	0.00213	2.22372	0.051156	0.050539	41.77628	0.948844	39.55256
39	1	45	0.002369	2.473236	0.249516	0.054961	42.52676	0.750484	40.05353
40	1	46	0.002387	2.492028	0.018792	0.054175	43.50797	0.981208	41.01594
41	1	47	0.002582	2.695608	0.20358	0.057353	44.30439	0.79642	41.60878
…	…	…	…	…	…	…	…	…	…
104	1	110	0.010835	11.31174	0.363312	0.102834	98.68826	0.636688	87.37652
105	1	111	0.010966	11.4485	0.136764	0.10314	99.5515	0.863236	88.10299
106	1	112	0.011041	11.5268	0.0783	0.102918	100.4732	0.9217	88.94639
107	1	113	0.01112	11.60928	0.082476	0.102737	101.3907	0.917524	89.78144
108	1	114	0.011198	11.69071	0.081432	0.10255	102.3093	0.918568	90.61858
109	1	115	0.011256	11.75126	0.060552	0.102185	103.2487	0.939448	91.49747
110	1	116	0.01126	11.75544	0.004176	0.10134	104.2446	0.995824	92.48912
111	1	117	0.012177	12.71279	0.957348	0.108656	104.2872	0.042652	91.57442
112	1	**118**	0.012225	12.7629	0.050112	0.10816	105.2371	0.949888	92.4742
113	1	119	0.014141	14.7632	2.000304	0.124061	104.2368	-1.0003	89.47359
114	1	120	0.014273	14.90101	0.137808	0.124175	105.099	0.862192	90.19798
115	1	121	0.014334	14.9647	0.063684	0.123675	106.0353	0.936316	91.07061
…	…	…	…	…	…	…	…	…	…
173	1	180	0.037786	39.44858	0.37584	0.219159	140.5514	0.62416	101.1028
174	1	181	0.038832	40.54061	1.092024	0.223981	140.4594	-0.09202	99.91878
175	1	182	0.039077	40.79639	0.25578	0.224156	141.2036	0.74422	100.4072
176	1	183	0.039224	40.94986	0.153468	0.22377	142.0501	0.846532	101.1003
177	1	184	0.0393	41.0292	0.079344	0.222985	142.9708	0.920656	101.9416
178	1	185	0.04194	43.78536	2.75616	0.236678	141.2146	-1.75616	97.42928
179	1	186	0.042014	43.86262	0.077256	0.235821	142.1374	0.922744	98.27477
180	1	187	0.042584	44.4577	0.59508	0.237742	142.5423	0.40492	98.08461
181	1	188	0.042642	44.51825	0.060552	0.236799	143.4818	0.939448	98.9635
182	1	189	0.043978	45.91303	1.394784	0.242926	143.087	-0.39478	97.17394
…	…	…	…	…	…	…	…	…	…

The p-value of each set can be observed in Figure 2. Since we have sorted our hypotheses based on their p-values, as we include more sets of hypotheses to our rejected set, the alpha (threshold p-value) increases. Depicted in red we see that FDR or .05 is allowing 42 or hypotheses to be rejected.

Figure 2, All p-values for 1031 hypotheses tested

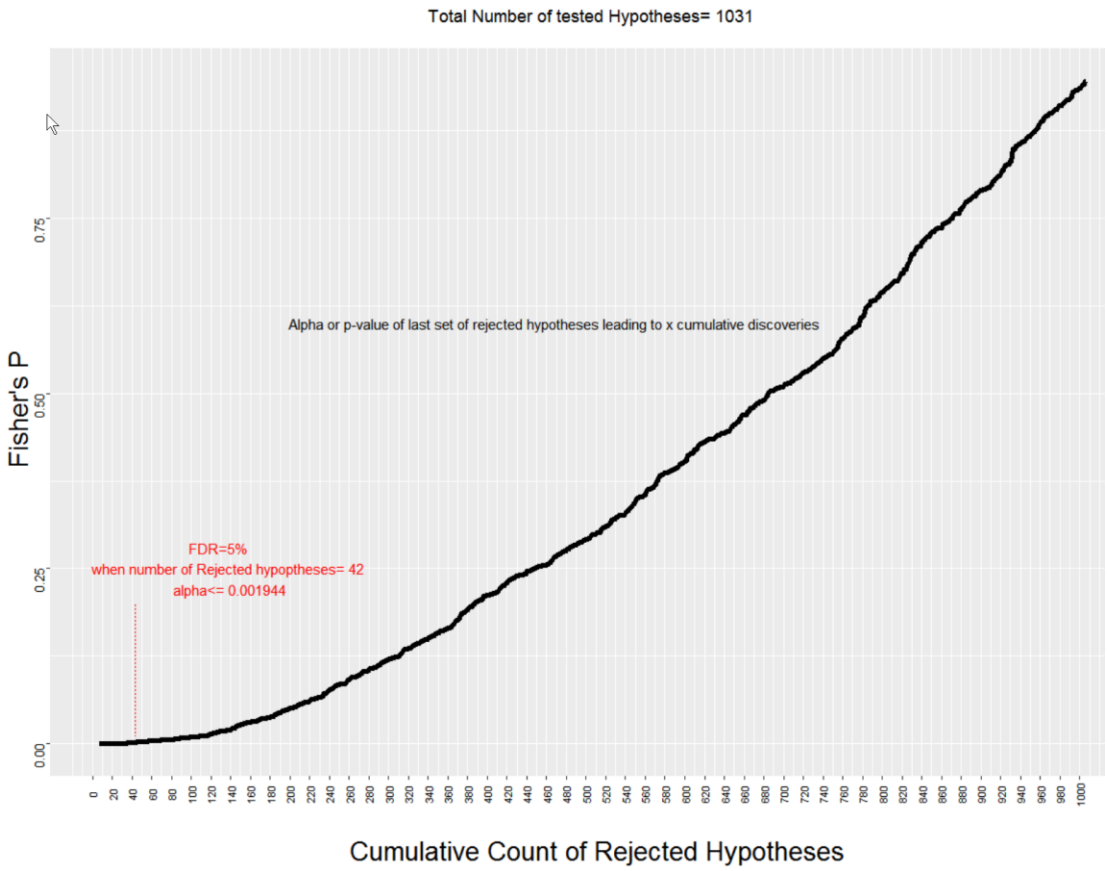


Figure 3, focuses on the first 400 lowest p-value hypotheses. The blue line is depicting the cumulative expected number of false discoveries among Rejected Hypotheses (false positives). Since CEFD is alpha*N, and alpha is the p-value of the last class rejected, CEFD is an increasing entity. The purple curve, FDR in percentage form, is also increasing even though one may find local fluctuations in its values.

$$\lim_{i \to N} CEFD_i = \lim_{i \to N} N \times p_i = N$$

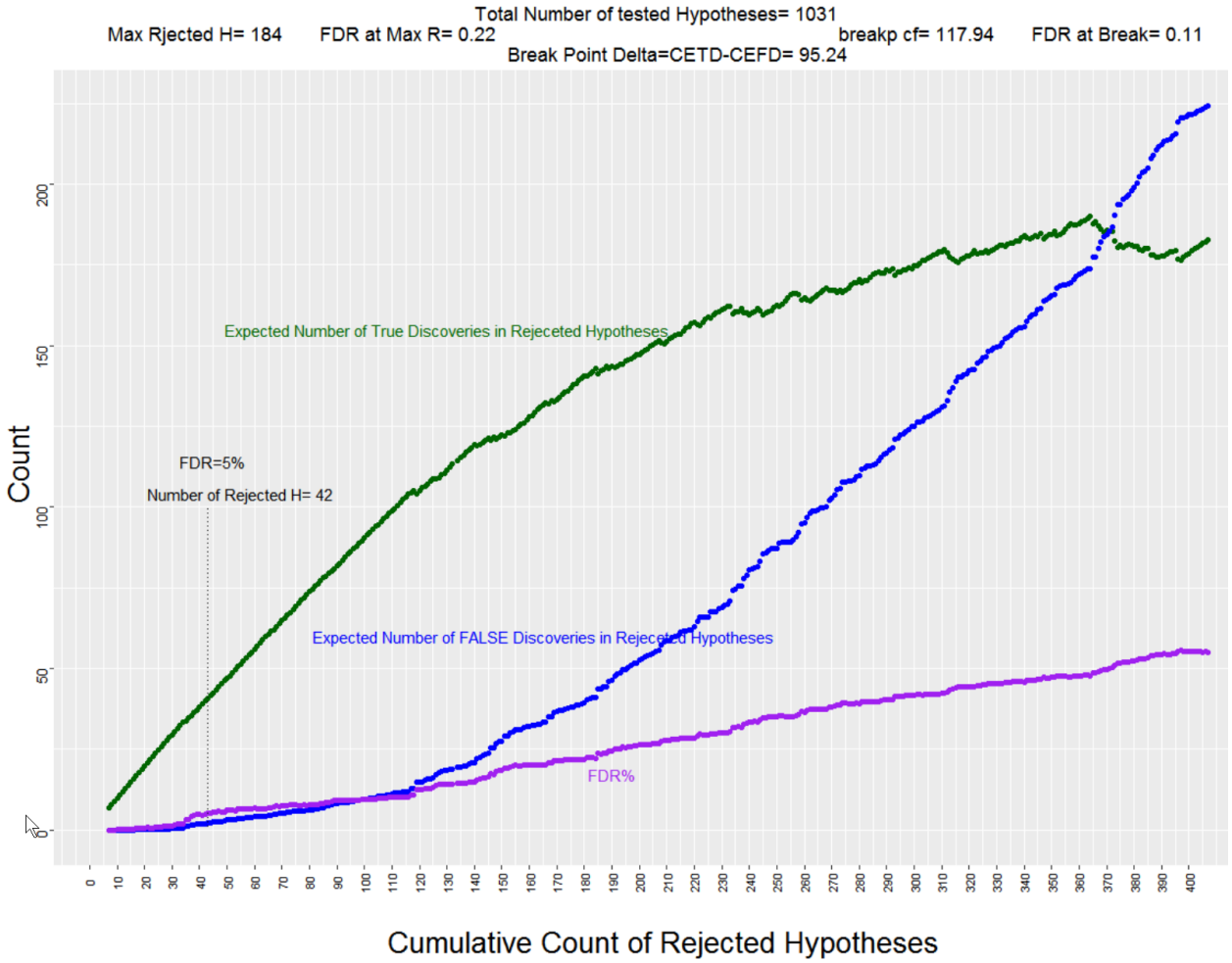
$$\lim_{i \to N} FDR_i = \lim_{p_i \to 1} FDRp_i = 1$$

The green line depicts the CETD. The p-value of first sets is very low, and these hypotheses are most likely to be true rejections, when we reject the first sets of hypotheses, CETD is growing very fast. Even when we pass the threshold of FDR=0.05 the p-values of

next sets are very low which keep FDRs close to .05. For example, in the study presented above, the hypothesis in set 37 has a p-values of 0.0001 and FDR_{37} is 0.050525.

If we add S_{37} to our rejected set R, our CETD will grow and CEFD will also grow, but the growth of CETD is much faster. This trend however doesn't last forever. As p-values get bigger, CEFD will grow faster and CETD will grow slower. If we continue rejecting hypotheses with big p-values CEFD will accelerate and will surpass CETD. CETD will start to decline when p-values included in rejection set get close to 1. If we look at the difference CETD-CEFD shown in the last column of table 1, we are sure that it has a maximum above which rejecting a set of hypotheses will contribute more to CEFD than CETD and the difference will start to decline.

Figure 3, CETD, CEFD and FDR for different number rejected hypotheses



In figure 4, δ, the difference between the expected true discoveries and expected false discoveries among rejected hypotheses, is depicted as a black line. As expected it has several local minima and maxima but it has a global maximum. Let us name the rejected

number of hypotheses at this point as R_{max}. FDR is always growing. By every new hypothesis we reject, we are increasing the proportion of false discoveries in the rejected set of hypotheses. Rejecting more hypotheses after we have reached R_{max}, will weaken the quality of discoveries in absolute sense. Table 1 shows that the p-values of set S_{177} is 0.0393. Rejecting hypothesis beyond R_{max} , for example rejecting set S_{178} which contains hypothesis 185, may increase the quantity CETD but it will increase the quantity of CEFD even more; it will decrease the quality of discovery because delta will go from 101.9416 to 97.42928.

R_{max} is a maximum for number of rejected hypotheses our data can justify. It will dictate a maximum for acceptable significance level alpha considering the data we have. In this data, R_{max} doesn't appear as a sharp peak at which we have a turn, it is a peak around which the trend has an slow reversal; therefore, we can use many methods that suggest a reasonable number of rejected hypotheses much lower than R_{max}.

If we use piecewise regression to identify two line segments, that will mimic the data upto R_{max}. The breakpoint is found at R_{105}. If we reject set S_{105}, or reject 110 hypotheses with lowest p-value, we will have a δ_{105}=88.10299 close to δ_{max}=101.9416 at R_{max}, with an FDR_{105} =0.10314 about half of FDR_{max} = 0.222985. As shown in table 1, the p-value of set S_{105} p_{105}=0.010966, about three times less than the p-value for p_{max}=0.0393.

Figure 4, maximum δ and the breakpoint of piecewise regression

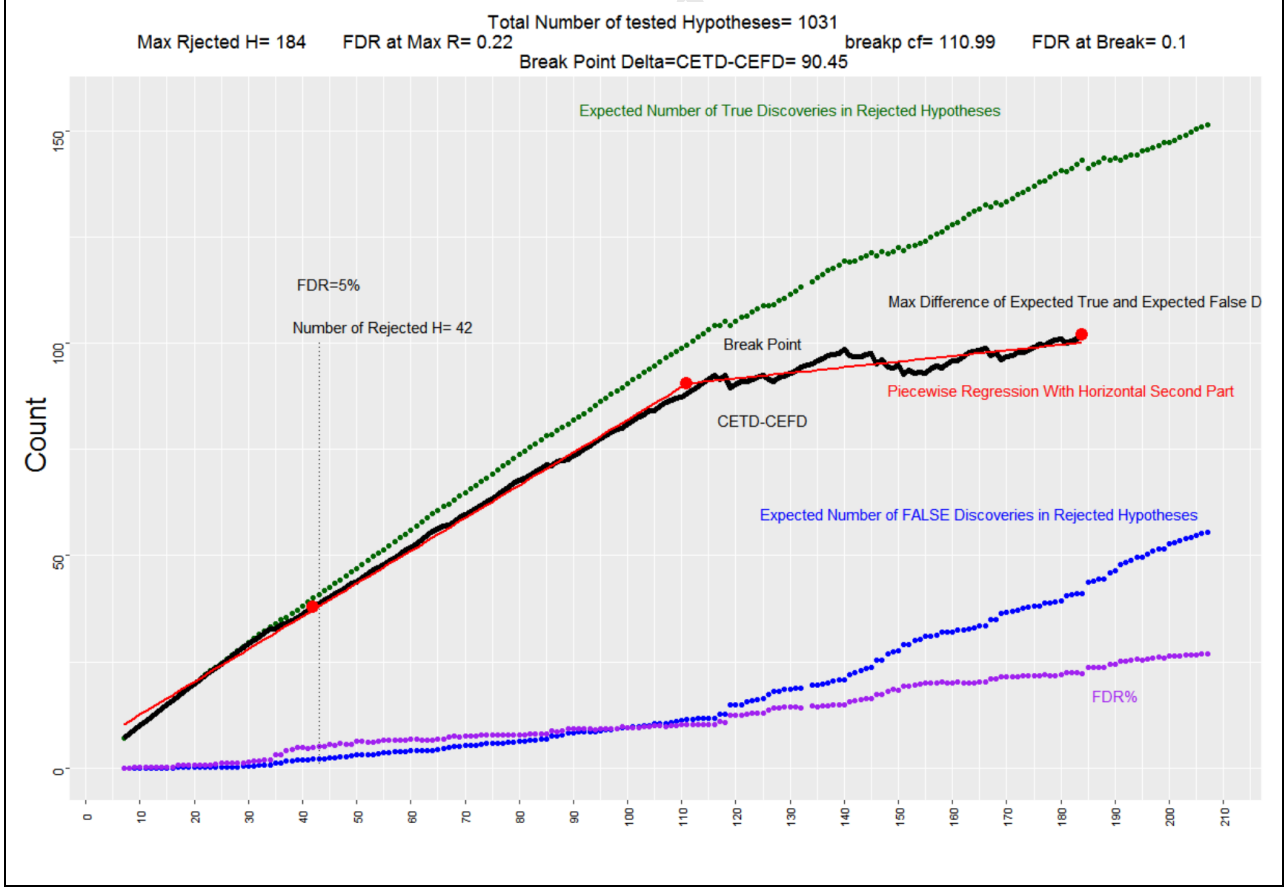
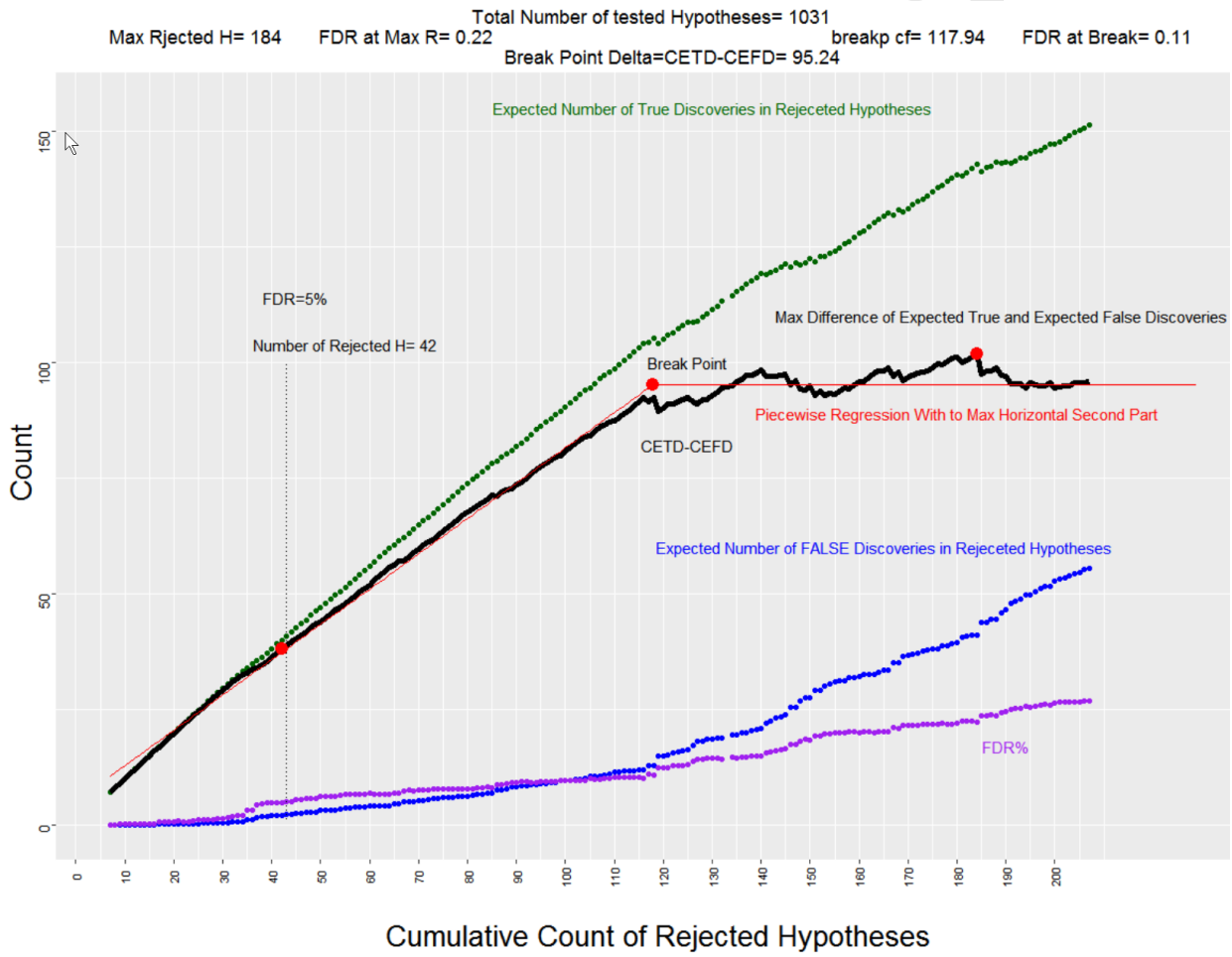


Figure 5 shows a slightly different strategy. If we use iterative piecewise regression to identify two line segments, one of which being a horizontal line that ends a few p-values after p_{max}. The breakpoint is at p_{118}. If we reject set S_{112}, or reject 118 hypotheses, we will have a δ_{112}=92.4742 close to δ_{max}=101.9416 at R_{max}, with an FDR_{112} =0.10816 about half of FDR_{max} = 0.222985. As shown in table 1, the p-value of set S_{112} p_{112}=0.012225, about three times less than the p-value for p_{max}=0.0393.

Using segmented regression is just one of many ways the researcher can include the information about R_{max}. The researcher can devise a more objective strategy to select the set of rejected hypothesis without relying on 0.05 or any other presumed thresholds for alpha or FDR. The researcher, should report the resulting alpha and FDR instead of assuming them.

Figure 5, maximum δ and the breakpoint of piecewise regression with horizontal piece



In the example shown above, the optimum (breakpoint of piecewise regression) is not very sensitive to the method of conducting regression. Either way, it suggests the about 10% of hypotheses which is much more than the number that could be rejected based on FDR=0.05 criterion and much less than absolute maximum Reasonable Number of Rejected Hypotheses.

In many exploratory researches the goal is to identify a set of significant associations. Many times, the extent of association (like slopes in linear regression) are more important for understanding the phenomena, or modeling the system, than the differences of FDRs associated with each p-value among significantly accepted alternatives. To test the quality of resulting set of rejected hypotheses, the non-parametric Sommer's D statistics for the extent of association for each comparison was calculated. It was observed that near all the rejected hypothesis had a level of association whose confidence intervals were on one side of Zero.

5. Discussion

In exploratory research, or for whom a few more possible false positives among many truly rejected hypotheses is not a sensitive issue, relying on predetermined threshold of 0.05 for FDR may be too limiting. But accepting larger and larger FDRs is not also a reasonable approach. The process explained in this paper neither requires predetermined thresholds for level of significance, nor uses presumed thresholds for false discovery rate. We observed a naturally occurring metric (for the quality of the set of rejected hypothesis), which has an upper bound. The researcher can rely on this maximum and devise methods to find an optimum that remains acceptable in terms of quality of discovery. Once the set of rejected hypotheses is determined a related significance level and FDR should be reported.

The paper presented methods that could identify optimum reasonable number of rejected hypotheses. The found optimum is in the range between most conservative selection criteria, such as what has been used in Bonferroni's procedure, and this identified upper bound.

The criterion and methods can be used in many fields of inquiry dealing with high-dimensional data, including genomics and survey analysis. The results of using the criterion in the pairwise crosstabulation analysis of an ordinal outcome variable with 1031 potential ordinal predictors in a large survey, regarding "variables influencing citizen engagement in mediated democracies", is used as an example of application of the method in social sciences.

One can follow the following steps to identify δ_{max} that data can afford.

1. start by sorting p-values from smallest to largest
2. tabulate the hypotheses to classes of observed p-values
3. reject the set of hypotheses with the least p-value (the first set is called S_1)
4. calculate cumulative expected false discoveries for all the rejected hypotheses ($P_i \times N$)
5. calculate 1-CEFD for all the rejected hypotheses
6. calculate δ=CETD-CEFD
7. record the results

8. repeat steps 2 to 7 for all the sets.
9. Find the set with maximum recorded δ called δ_{max} resulting from rejecting set S_{max}
10. The biggest reasonable set of rejected hypotheses R_{max} will be

$$R_{max} = S_1 \cup S_2 \cup S_3 \cup \ldots \cup S_{max}$$

11. The p-value for set S_m is p_m which is the alpha that should be reported
12. The FDR that should be reported is

$$FDR_{max} = \frac{p_{max} \times N}{\sum_1^m f_i}$$

References

Blanchard, G. and Roquain, E. (2009) *Adaptive False Discovery Rate Control under Independence and Dependence* Journal of Machine Learning Research 10:2837-2871.

Neuvial, P. (2013) *Asymptotic results on adaptive false discovery rate controlling procedures based on kernel estimators.* JMLR **14**, 1423-1459

Norris AW, Kahn CR. Analysis of gene expression in pathophysiological states: Balancing false discovery and false negative rates. Proc Natl Acad Sci U S A. 2006; 103:649–653. [PubMed: 16407153]

Park, B. S., & Mori, M. (2010). Balancing false discovery and false negative rates in selection of differentially expressed genes in microarrays. Open Access Bioinformatics, 2010(2), 1–9. http://doi.org/10.2147/OAB.S7181

Storey, John D. (2007). The optimal discovery procedure: A new approach to simultaneous significance testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **69**(3), 347–368.

Storey, J. and Tibshirani, R. (2003). *Statistical significance for genomewide studies*, Proceedings of the National Academy of Sciences 100-: 9440–9445.