Journal of Big Data

METHODOLOGY

Open Access

A non-parametric maximum for number of selected features: objective optima for FDR and significance threshold with application to ordinal survey analysis

Amir Hassan Ghaseminejad Tafreshi[1,2,3*]

*Correspondence:
aghasemi@capilanou.ca
[3] School of Business, Capilano University, 2055 Purcell Way, North Vancouver, BC V7J 3H5, Canada
Full list of author information is available at the end of the article

Abstract

This paper identifies a criterion for choosing an optimum set of selected features, or rejected null hypotheses, in high-dimensional data analysis. The method is designed for dimension reduction with multiple hypothesis testing used in filtering process of big data, and in exploratory research, to identify significant associations among many predictor variables and few outcomes. The novelty of the proposed method is that the selected p-value threshold will be insensitive to dependency within features, and between features and outcome. The method neither requires predetermined thresholds for level of significance, nor uses presumed thresholds for false discovery rate. Using the presented method, the optimum p-value for powerful yet parsimonious model is chosen, then for every set of rejected hypotheses, the researcher can also report traditional measures of statistical accuracy such as the expected number of false positives, and false discovery rate. The upper limit for number of rejected hypotheses (or selected features) is determined by finding the maximum difference between expected true hypotheses and expected false hypotheses among all possible sets of rejected hypotheses. Then, many methods of choosing an optimum number of selected features such as piecewise regression are used to form a parsimonious model. The paper reports the results of implementation of proposed methods in a novel example of non-parametric analysis of high-dimensional ordinal survey data.

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

Introduction

High-dimensionality is one of the attributes of big data in many fields. As stated by Dutheil and Hobolth [7], the shift from genetics to genomics brings to new challenges in data analysis. For example, when tests are performed, the global false discovery rate (FDR) has to be properly controlled for (p. 310). According to Kim and Halabi [12], a vital step in model building is dimension reduction. For example, in clinical studies, it is assumed that there are several variables that are associated with the outcome in the large dimensional data. The main purpose of the variable selection is to identify only those variables which are related to the response. They have identified two steps in variable selection: screening and model building. The screening step is to reduce the number of variables while maintaining most of the variables relevant to the response, and the model building step is to develop a best model (p. 1).

Fan et al. [8] argue that the complexity of big data often makes dimension reduction techniques necessary before conducting statistical inference. Principal component analysis (PCA), the goal of which is to find a lower dimensional subspace that captures most of the variation in the dataset, has become an essential tool for multivariate data analysis and unsupervised dimension reduction (p. 1). But Kim et al. [13] have argued that although PCA and partial least squares (PLS) methods have been widely used for feature selection in nuclear magnetic resonance (NMR) spectra, extracting meaningful features from the reduced dimensions obtained through PCA or PLS is complicated because, in both PCA and PLS, reduced dimensions are linear combinations of a large number of the original features. The authors show that successful implementation of feature selection, based on their proposed multiple testing procedure of controlling FDR, is an efficient method for feature selection in NMR spectra that improves classification ability and simplifies the entire modeling process; thus, reduces computational and analytical efforts (p. 1).

Miao and Niu [14] define feature selection, as a dimensionality reduction technique which aims to "choosing a small subset of the relevant features from the original features by removing irrelevant, redundant or noisy features." They also point out that, considering the increase in both number of samples and dimensionality of data used in many machine learning applications such as text mining, computer vision and biomedical, feature selection can lead to better learning performance, higher learning accuracy, lower computational cost, and better model interpretability (p. 919). According to Bolón-Canedo et al. [6], due to the appearance of datasets containing hundreds of thousands of variables, feature selection has been one of the high activity research areas (p. ix). For example, contemporary biological technologies produce extremely high-dimensional data sets with limited samples which demands feature selection in classifier design [10]. Fiori et al. [9] consider feature selection as one of the popular and recent data mining techniques applied to microarray data (p. 29).

Shmueli et al. [20] define a predictor is a variable, used as an input into a predictive model, also called a feature, input variable, independent variable, or from a database perspective, a field (p. 10). The general research question is how many predictors, should be included in the model being constructed.

In high-dimensional data analysis, multiple simultaneous hypothesis testing arises because we need to identify which null hypotheses, among many, should be reasonably rejected [15]. A significant finding (discovery) is a hypothesis that is rejected based on statistical evidence. Rejecting a null hypothesis, about the relevance of a predictor to an outcome, is in fact selecting the predictor as a feature that will be included in the model.

As explained by Ochoa et al. [17], each test "yields a score s and a p-value, defined as the probability of obtaining a score equal to or larger than s if the null hypothesis holds." While a p-value threshold like 0.05 is acceptable to declare a single test significant, this is inappropriate for a large number of tests. Some studies are based on the E-value defined as: $E = pN$, where N is the number of tests, and yields the expected number of false positives at this p-value threshold. E-values are not very meaningful when millions of positives are obtained, and a relatively larger number of false positives might be tolerated. FDR, however, is an appealing alternative approach (p. 2–3).

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 0.05, 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 0.05 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.

Iterson et al. [11] argue that thresholds for significance yielded by multiple testing methods decrease as the number of hypotheses tested increases (p. 2). The hypotheses with very low p-values are the hypotheses that we might be inclined to declare as rejected null hypotheses, selected features, or significant discoveries; however, if we simply choose hypotheses with (for example) p-value < 0.05, 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 reduce our p-value threshold 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 significance threshold (alpha), we are accepting to have more false negatives or in other words more hypotheses which should be rejected but are not.

By choosing a rejection threshold much lower than alpha, that is less than or equal to alpha/N, the probability of making one or more false discoveries will be less than or equal to alpha [22].

Dealing with N p-values, applying Bonferroni correction limits the false positive rate to less than or equal to alpha

If
Bonferroni corrected threshold $\leq \frac{\propto}{N}$ is used
Then
Expected number of false positives with Bonferroni corrected threshold $\leq \frac{\propto}{N}N \leq \propto$

Bonferroni correction 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 is 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 few more false-positives, if that will prevent many false negatives. Not detecting many important associations may be more harmful than probability of a few false-negative among many significant findings.

Benjamini and Hochberg [5] suggested that, instead of classical approach of using the FWER in the strong sense, we can control the FDR. FDR is defined as expected number of false discoveries (false positives among rejected hypotheses) divided by total number of rejected hypotheses [15]. They proved that their way of determining p-value threshold controls the FDR at a certain level when the p-values corresponding to true null hypotheses are independent and identically distributed with a uniform distribution [23]. As emphasised by Storey and Tibshirani [22], the false positive rate and FDR are different. "Given a rule for calling features significant, the false positive rate is the rate that truly null features are called significant. The FDR is the rate that significant features are truly null."

In many situations, a p-value of 0.05 may lead to a big FDR. Several algorithms have been proposed to consider FDR in the process of selecting significant findings. Holms has proposed a sequential step-down algorithm which is shown to be "uniformly more powerful than Bonferroni's simple procedure." Also, Hochberg has suggested a step-up procedure which is very similar to Holm's proposed method [1].

Iterson et al. [11] explain that statistical analysis of high dimensional data, occurs when the number of parameters is much larger than the number of samples. It often involves testing of multiple hypotheses in which p-values must be corrected. The larger the number of hypotheses tested, the stronger the correction for multiple testing must be in order to keep the error rate acceptably low. To decrease this penalty, and improve power, some studies select some features prior to the data analysis. But this selecting procedure, called "filtering process" can leave some features out of the analysis. Also, inevitably some non-features may be selected by these filters. In absence of proper filtering out of the entire range of p-values the result will be a biased multiple testing correction (p. 1). They conclude that: to avoid filtering-induced FDR-bias, Alternatives, for any generic filter and test, should adapt the multiple testing correction methods that relax the assumption of uniform distribution for the null features in a way that filtering-induced bias is avoided (p. 10).

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 [4]. Plug-in procedures use an estimate of the proportion of true null hypotheses [15]. Thresholding-based multiple testing procedures, reject hypotheses with p-values less than a threshold [15]. Storey and Tibshirani [22] have proposed a strategy that assigns each hypothesis an individual measure of significance in terms of expected FDR called q-value. Most q-value based strategies rely on some estimate of the proportion of true null hypotheses.

Storey [21] has argued that two steps that are involved in any multiple-testing procedure. In the first step one must rank the tests from most significant to least significant. In the second step one must choose an appropriate significance cut-off. Storey focuses on performing the first step optimally, given a certain significance framework for the second step. Story cites Shaffer [19] identifying the goal to be estimating the reasonable cut-off resulting a particular error rate. Storey 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 [16] have proposed balanced probability analysis (BPA) based on three variables: (i) the total number of true positives (TTP); (ii) the aggregate chance that any gene listed is truly not changing and is, thus, on the list by statistical accident (iii) 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 non-discovery rate (the ratio of hypotheses that should truly be rejected but are not discovered to the number of un-rejected hypothesis) are difficult to understand for those who are not expert statisticians. They calculate the FNR, by using resampling to estimate the null and alternate distributions, directly from the data. Their procedure a model-dependent step to optimize a single parameter.

As Norris and Kahn [16] 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 [22] they estimate the TTPs. They estimated FDR and then they estimated FNR based on their estimates of FDR and TTP. In his dissertation, Benditkis [2] has shown that for some classes of step-down procedures the expected number of false rejections is controlled under martingale dependence. Benditkis et al. [3] have presented a rapid approach to the step up and step-down tests.

According to Park and Mori [18] the *FDR* method is perhaps the most popularly used multiple comparison procedures (MCP) in microarrays. Kim and Halabi [12] have proposed the use of FDR as a screening method to reduce the high dimension to a lower dimension as well as controlling the FDR with other variable selection methods such as least absolute shrinkage and selection operator (LASSO), and smoothly clipped absolute deviation (SCAD) (p. 1). In our example, which is in the context of high dimensional ordinal analysis of survey data, we will compare the results of our proposed method with FDR results.

One of the difficulties with targeting an FDR such as 0.05 is that when predictors, or features, are dependent to each other and to the outcome, the p-values of null hypotheses tested about their association with outcome will be similarly small. These p-values will inflate the FDR while their selection does not contribute to the number of differentiable constructs in the model. The method that will be proposed is insensitive to the number of highly correlated features that are selected. Thus, it can improve the feature selection power.

Methods

A non-parametric maximum for reasonable number of rejected hypotheses or number of selected features

This article, is concerned about choosing an appropriate significance threshold after we have ordered the hypotheses based on their p-values without knowing or estimating the total number of true positives or total number of true negatives.

There are research questions where possibility of even one false discovery (existence of one false positive among all the rejected hypothesis) is not desirable. For such research a Bonferroni corrected threshold is necessary. But, when identification of contributing variables is the goal, and having some falsely rejected hypothesis or falsely chosen features is not prohibitive, the researcher may choose the p-value threshold based on an expected FDR. 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 of reasonable findings a researcher should report. For some researchers, who accept an FDR of 5%; it might be also reasonable to accept and report a model that includes features in which the FDR is 6%, specially if this increase will add a group of items to selected features, or rejected nulls, that are mostly true discoveries.

In many situations, "it is reasonable to assume that larger p-values are more likely to correspond to true null hypotheses than smaller ones" [15], which means smaller p-values are less likely to correspond to true null hypotheses (if rejected, they are more likely to be true discoveries). With no objective reason to accept 5% FDR and not 6% FDR, in some situations, grounded on observed data we can identify an objective upper bound for "level of significance and FDR" that is reasonable for the researcher to report, beyond which the resulting model is not parsimonious. To find such maximum, we tabulate the p-values resulted from hypotheses testing into sorted classes (from smallest to largest p-value). Then, we choose the smallest p-value; we count the number of hypotheses that have the same p-value and we put them in set 1. Then we choose the next smallest p-value; we count the number of hypotheses that have the same p-value and we put them in set 2. We continue to the biggest p-value. We will have the frequency of each observed p-value. But, 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) are in set S_1 in which will have f_1 members ($f_1 \geq 1$).

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 ($f_2 \geq 1$). For each one of k observed p-values there will be corresponding frequency and a set of hypotheses.

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

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).

$$\mathrm{EFD_1 = p_1 * N}$$

Therefore, from the first set we expect to have:

$$\mathrm{ETD_1 = f_1 - \left(p_1 * N\right)}$$

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 p_2 and the expected false discoveries in rejected set S_1 and S_2 will be p_2*N. R_2 is the set total discoveries including all the features selected so far.

$$\mathrm{R_2 \ = \ S_1 U \, S_2}$$

$p_2 * \mathrm{N}$ is always bigger than $p_1 * \mathrm{N}$. $p_2 * \mathrm{N}$ will be the cumulative expected false discoveries (CEFD) in R_2:

$$\mathrm{CEFD}_2 = p_2 * \mathrm{N}$$

Therefore, from the first two sets we expect to have cumulative expected false discoveries (CETD) in R_2 as:

$$\mathrm{CETD}_2 = (\mathrm{f}_1 + \mathrm{f}_2) - (p_2 * \mathrm{N})$$

Therefore, cumulative expected true discoveries CETD_2 from S_1 and S_2, will be bigger than ETD_1. The series of cumulative expected false discoveries: CEFD_1, CEFD_2, CEFD_3, … is usually increasing because the p-values are getting bigger. And the series of cumulative expected true discoveries is each set: CETD_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 ($\mathrm{j} > \mathrm{i}$) 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 (which means selecting all possible features), p_i goes toward 1 (which means selecting features that have no significance).

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

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

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

If we define delta:

$$\delta_\mathrm{i} = \mathrm{CETD}_\mathrm{i} - \mathrm{CEFD}_\mathrm{i}$$

The δ_1 is always positive, and δ_N is always negative. 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 $\mathrm{S}_{\mathrm{max}+1}$ (setting alpha at $\mathrm{p}_{\mathrm{max}+1}$) 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. $\mathrm{FDR}_\mathrm{max}$ 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 $\delta_{\max}$ 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 $\leq p_{max}$.

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

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

$$R_{max} = \sum_{1}^{m} f_i$$

Maximum ECTD can be calculated based on the following formula:

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

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

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 cumulative expected true hypotheses and cumulative expected false hypotheses. The significance threshold is reported (not assumed) and is not subjectively selected. The significance threshold and resulting FDR are 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.

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 may have a smooth decrease of δ. The value of "$\mathrm{CETD}_i - \mathrm{CEFD}_i$" sometimes changes relatives slowly around R_{max}. Then, we have a peak and a slow reversal in trend for δ. Thus, 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 does not 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

Table 1 The format of a table that will be used to analyze p-values 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)	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	$-\mathrm{N}$
	$\sum_{i=1}^{k} (p_i) = 1$	$\sum_{i=1}^{k} (f_i) = \mathrm{N}$				

point (R_{max}, p_{max}). In some situations, the resulting set of discoveries is not very sensitive to the selection of piecewise regression method.

Results and discussion

As an example, we will use the dataset resulted from multiple hypotheses testing that was performed analyzing a survey regarding variables that influence citizen engagement in mediated democracies. The example included 1045 ordinal Likert scale questions. The answers to a question were cross tabulated with all the other 1044 questions. The null hypothesis was that the observed association in each pairwise cross tabulation is accidental. To choose the significant associations, Fisher's exact test was performed and the p-values for each test was recorded. Sommer's D was used to measure the extent of association.

Table 2 shows the sorted p-values and associated δs. The first column is the set number. The second column is an ascending sorted list of all p-vales observed. The third column is the frequency of hypotheses with the p-value. The forth column is the cumulative frequency. The fifth column is the cumulative expected number of false discoveries. If we decide that the set in a row is the last set of rejected hypotheses, the cumulative expected number of false discoveries is the resulting expected number of false positives. In other words, if the p-value of the set in a row is considered as the rejection threshold, the cumulative expected number of false discoveries in the fifth column of that row is the resulting expected number of false discoveries. For example, we observe that if we reject 111 hypotheses, sets S_1 to S_{105}, the expected number of false positives will be 11.4485 leading to an FDR of 0.10314. Chosen p-value threshold is 0.010966 and FDR will provide a measure of statistical accuracy for the choice we make for our rejection threshold.

The sixth column is the difference between rows in the fifth column, in other words it is the contribution of each row to expected number of false discoveries. Column seven is the FDR if we consider the p-value of the row as rejection threshold. Column eight is the expected cumulative number of true discoveries if this set is rejected; it is calculated by subtracting expected number of false discoveries from cumulative number of rejected hypotheses. Column nine is the difference between rows in column nine. The last column is δ.

If we rely on 0.05 rule of thumb for rejection threshold, 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 un-rejected. We have seven 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 1044*0.000001 = 0.001044. Rejecting the hypotheses in sets S_1 and S_2, we are in fact declaring the rejection threshold is 0.000001, cumulative expected false discoveries will be 0.001044, FDR will be 0.000131 (0.001044/8).

Bonferroni's correction for p-value = 0.05 would suggest a threshold of rejection of < 0.0000485 which means we can conservatively reject merely 16 null hypotheses. If we reject all the hypotheses in sets S_1 to S_{36}, we will have 42 hypotheses in our set of

Table 2 Sorted p-values and associated δ_S, R_{max} and breakpoints

Column:	2	3	4	5	6	7	8	9	10
Set	p-value of the set (alpha if this set is rejected)	Frequency in set	Less than cumulative frequency hypotheses in rejected sets	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	$\delta = \text{CETD} - \text{CEFD}$
1	0	7	7	0	0	0	7	7	7
2	0.000001	1	8	0.001044	0.001044	0.000131	7.998956	0.998956	7.997912
3	0.000005	1	9	0.00522	0.004176	0.00058	8.99478	0.995824	8.98956
4	0.000008	1	10	0.008352	0.003132	0.000835	9.991648	0.996868	9.983296
5	0.000009	1	11	0.009396	0.001044	0.000854	10.9906	0.998956	10.98121
6	0.000015	1	12	0.01566	0.006264	0.001305	11.98434	0.993736	11.96868
7	0.000021	1	13	0.021924	0.006264	0.001686	12.97808	0.993736	12.95615
8	0.000022	1	14	0.022968	0.001044	0.001641	13.97703	0.998956	13.95406
9	0.000034	1	15	0.035496	0.012528	0.002366	14.9645	0.987472	14.92901
10	0.000037	1	16	0.038628	0.003132	0.002414	15.96137	0.996868	15.92274
11	0.0001	1	17	0.1044	0.065772	0.006141	16.8956	0.934228	16.7912
…	…	…	…	…	…	…	…	…	…
34	0.00183	1	40	1.91052	0.061596	0.047763	38.08948	0.938404	36.17896
35	0.001841	1	41	1.922004	0.011484	0.046878	39.078	0.988516	37.15599
36	*0.001944*	*1*	*R @ FDR 0.05 42*	*2.029536*	*0.107532*	*0.048322*	*39.97046*	*0.892468*	*37.94093*
37	0.002081	1	43	2.172564	0.143028	0.050525	40.82744	0.856972	38.65487
38	0.00213	1	44	2.22372	0.051156	0.050539	41.77628	0.948844	39.55256
39	0.002369	1	45	2.473236	0.249516	0.054961	42.52676	0.750484	40.05353
…	…	…	…	…	…	…	…	…	…
104	0.010835	1	110	11.31174	0.363312	0.102834	98.68826	0.636688	87.37652
105	*0.010966*	*1*	*R @ Break Point 111*	*11.4485*	*0.136764*	*0.10314*	*99.5515*	*0.863236*	*88.10299*
106	0.011041	1	112	11.5268	0.0783	0.102918	100.4732	0.9217	88.94639
107	0.01112	1	113	11.60928	0.082476	0.102737	101.3907	0.917524	89.78144
108	0.011198	1	114	11.69071	0.081432	0.10255	102.3093	0.918568	90.61858

Table 2 (continued)

Column:	2	3	4	5	6	7	8	9	10
Set	**p-value of the set (alpha if this set is rejected)**	**Frequency in set**	**Less than cumulative frequency hypotheses in rejected sets**	**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**	**$\delta = \text{CETD} - \text{CEFD}$**
109	0.011256	1	115	11.75126	0.060552	0.102185	103.2487	0.939448	91.49747
110	0.01126	1	116	11.75544	0.004176	0.10134	104.2446	0.995824	92.48912
111	0.012177	1	117	12.71279	0.957348	0.108656	104.2872	0.042652	91.57442
112	*0.012225*	*1*	*R @ Break Point 118*	*12.7629*	*0.050112*	*0.10816*	*105.2371*	*0.949888*	*92.4742*
113	0.014141	1	119	14.7632	2.000304	0.124061	104.2368	-1.0003	89.47359
114	0.014273	1	120	14.90101	0.137808	0.124175	105.099	0.862192	90.19798
115	0.014334	1	121	14.9647	0.063684	0.123675	106.0353	0.936316	91.07061
…	…	…	…	…	…	…	…	…	…
174	0.038832	1	181	40.54061	1.092024	0.223981	140.4594	-0.09202	99.91878
175	0.039077	1	182	40.79639	0.25578	0.224156	141.2036	0.74422	100.4072
176	0.039224	1	183	40.94986	0.153468	0.22377	142.0501	0.846532	101.1003
177	*0.0393*	*1*	*R_{max} 184*	*41.0292*	*0.079344*	*0.222985*	*142.9708*	*0.920656*	*101.9416*
178	0.04194	1	185	43.78536	2.75616	0.236678	141.2146	-1.75616	97.42928
179	0.042014	1	186	43.86262	0.077256	0.235821	142.1374	0.922744	98.27477
180	0.042584	1	187	44.4577	0.59508	0.237742	142.5423	0.40492	98.08461
181	0.042642	1	188	44.51825	0.060552	0.236799	143.4818	0.939448	98.9635
…	…	…	…	…	…	…	…	…	…

If we set p-value threshold at 0.001944, we reject 42 null hypotheses, and we can report FDR less than 0.05

If we set p-value threshold at 0.010966, we reject 111 null hypotheses, and we can report FDR less than 0.11

If we set p-value threshold at 0.012225, we reject 118 null hypotheses, and we can report FDR less than 0.11

If we set p-value threshold at 0.0393, we reject 184 null hypotheses, and we can report FDR less than 0.23

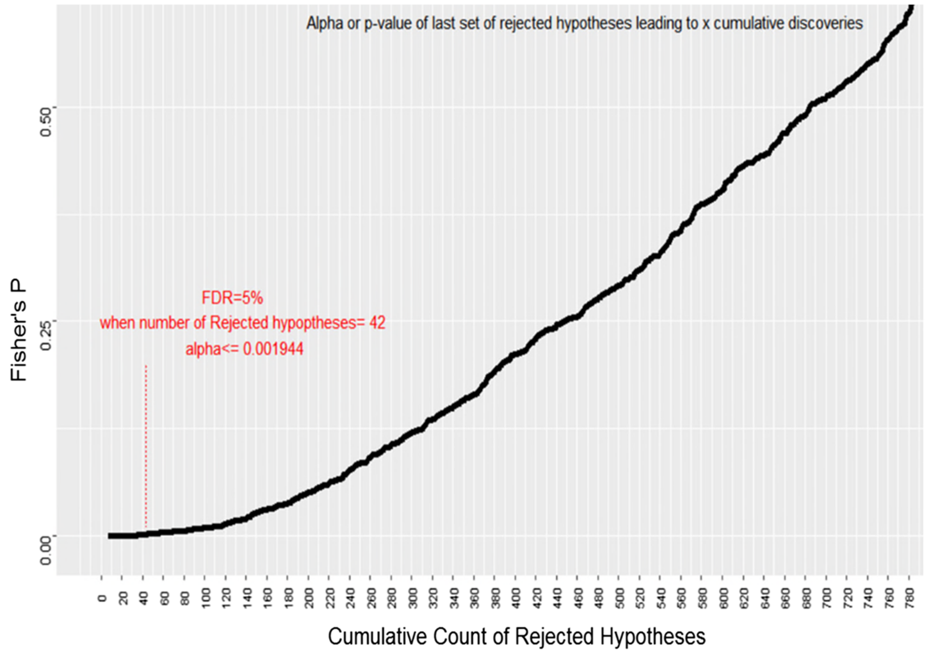

Fig. 1 All p-values for 1031 hypotheses tested

rejected hypotheses R_{36} and FDR will be 0.048322. Like many researchers who will not reject set S_{37}, we can define our p-values threshold of rejection to be 0.001944. This is more powerful than Bonferroni's correction. But we expect 2.029536 of 42 discoveries to be false. Our set S_{177} has the 184th p-value at 0.0393. The resulting set of rejected hypotheses from S_1 to S_{177} is expected to have 41.0292 to be false discoveries and 142.9708 true discoveries. The expected FDR as the result of increasing alpha to 0.0393 will be 0.222985.

The p-value of each set can be observed in second column of Fig. 1. Since we have sorted our hypotheses based on their p-values, as we include more sets of hypotheses to our rejected set, the reasonable alpha (threshold p-value) increases. Depicted in red, we see that at FDR of 0.05 we can select 42 features, or we can reject 42 null hypotheses.

Figure 2, 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 monotonic p-value of the last class rejected chosen as threshold, CEFD is an increasing entity. The purple curve, FDR in percentage form, is also generally increasing even though one may find local fluctuations in its values.

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

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

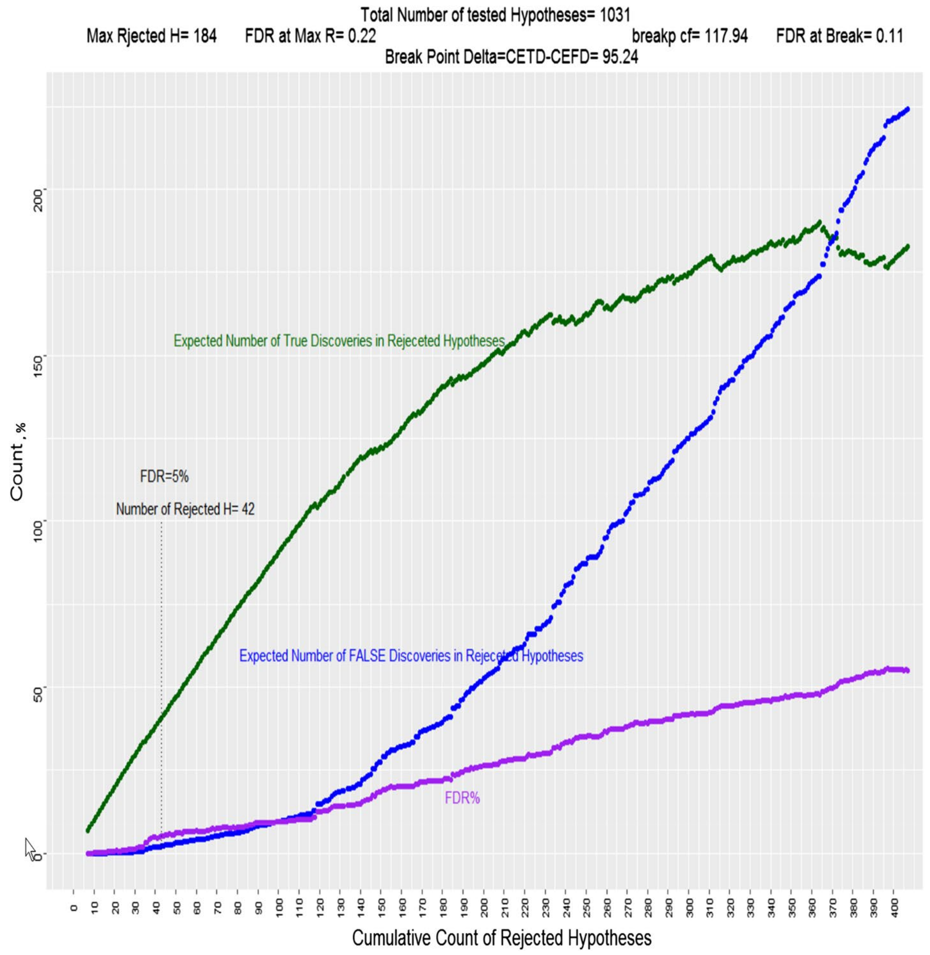

Fig. 2 CETD, CEFD and FDR for different number rejected hypotheses

The green line depicts the CETD. The p-value of first sets is usually 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 0.05. 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 does not 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 closer to 1. If we look at the difference CETD − CEFD shown in the last column of Table 2, we see that it has a maximum at set 177 above which rejecting a set of hypotheses will contribute more to CEFD than CETD and the difference will start to decline.

In Fig. 3, 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

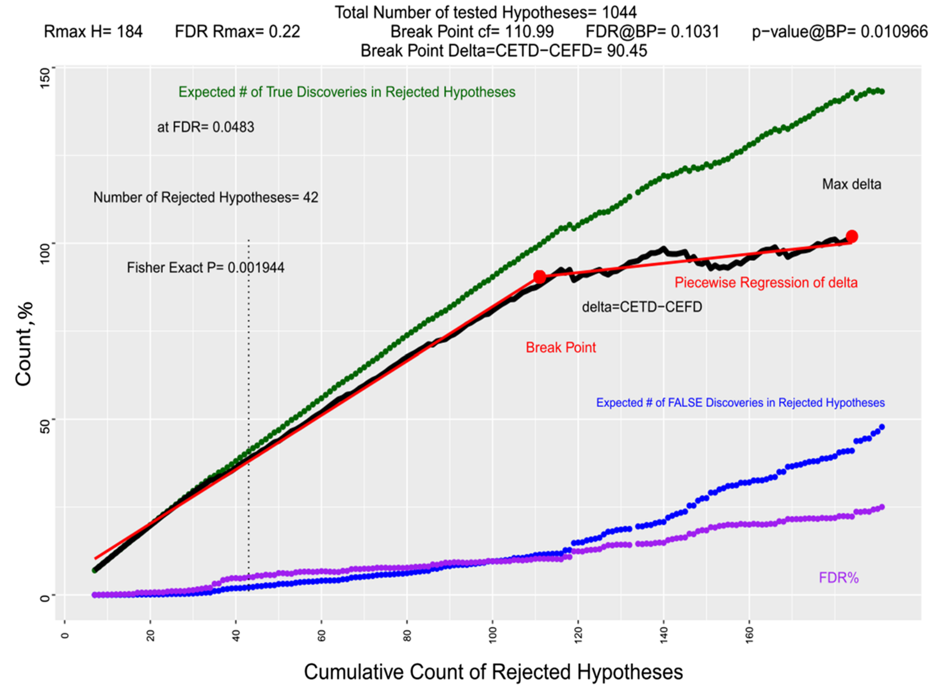

Fig. 3 Maximum δ and the breakpoint of piecewise regression

new rejected hypothesis, 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 our model with more contribution to expected false discoveries than expected true discoveries. Table 2 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 the maximum number of rejected hypotheses (features included in the model) which our data can justify. It will dictate a maximum for acceptable significance level alpha considering the data we have. In this example, R_{max} does not appear at 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} leading to more parsimonious models

If we use piecewise regression to identify two regression line-segments, that will mimic the data up to R_{max}, the breakpoint is found at set S_{105}. If we reject set S_{105}, or reject 111 hypotheses with lowest p-value, we will have a $\delta_{105} = 88.10299$ vs $\delta_{max=}101.9416$ at R_{max}. Our FDR will be $FDR_{S105} = 0.10314$; about half of $FDR_{max} = 0.222985$. As shown in Table 2, the p-value of set S_{105} is $p_{105} = 0.010966$, about three times less than the p-value for $p_{max} = 0.0393$. At breakpoint, we select 69 (111–42) more features of which 9.418964 (11.4485–2.029536) are expected to be false discoveries. The selection process based on p-value threshold of 0.010966 with 111 selected features is more powerful than the model with 42 features based on FDR = 0.05. At the same time, it is more parsimonious than a model with 184 features

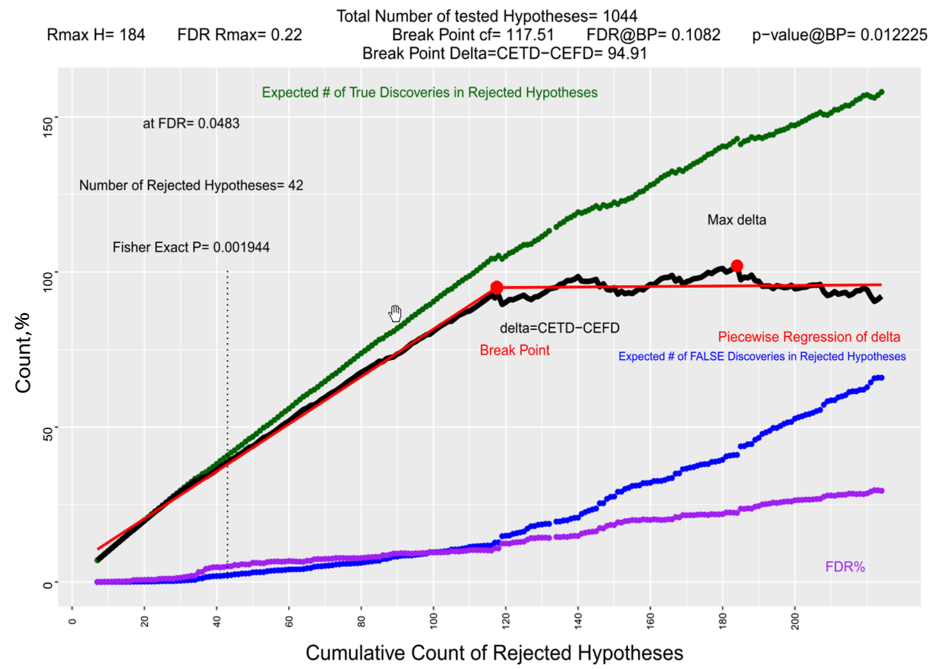

Fig. 4 Maximum δ and the breakpoint of piecewise regression with horizontal piece

suggested by maximum delta, with 72 (184–111) more variables including 29.5807 (41.0292–11.4485) more false discoveries.

Figure 4 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 $\mathrm{p}_{\mathrm{max}}$. The breakpoint is at p_{118}. If we reject set S_{112}, or reject 118 hypotheses, we will have a $\delta_{112} = 92.4742$ close to $\delta_{\mathrm{max=}}$101.9416 at R_{max}, with an $\mathrm{FDR}_{112} = 0.10816$ about half of $FDR_{max} = 0.222985$. As shown in Table 2, the p-value of set S_{112} is $\mathrm{p}_{112} = 0.012225$, about three times less than the p-value for $\mathrm{p}_{\mathrm{max}} = 0.0393$.

Using segmented regression is just one of many ways the researcher can include the information about $\mathrm{R}_{\mathrm{max}}$. 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; and, should report the resulting alpha and FDR instead of assuming them.

In the example shown above, the optimum (breakpoint of piecewise regression) is not very sensitive to the method of conducting regression or identification of breakpoint. Either way, it suggests a threshold that corresponds to an FDR between 10 and 11%. At this neighborhood of FDR, 111 or 118 hypotheses could be rejected (111, or 118 features could be selected); while based on FDR = 0.05 criterion, 42 hypotheses could be rejected (42 features could be selected); nevertheless, the proposed method increases the power of selection process. Resulting model is much more parsimonious than selecting 184 features suggested by $\mathrm{R}_{\mathrm{max}}$ (absolute maximum reasonable number of rejected hypotheses).

It important to notice that if a number of predictors, or features, which are dependent to each other and associated with the outcome exist, the p-values of null hypotheses tested about their association with outcome will be similarly small. These p-values will inflate the FDR and may exclude some eligible features from the model, but their contribution to ECTD and ECFD will be similar and will not affect the delta; therefore, maximum delta and the delta at breakpoint are not affected by the number of features in each set with similar p-values and the proposed method of selecting p-value threshold is insensitive to the number highly correlated features that my be selected.

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 (selected features) had a level of association whose confidence intervals for the extent of association were on one side of zero.

Conclusion

In exploratory research, or when a few more possible false positives among many truly rejected hypotheses or selected features 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 presented method is a proposal for an objective threshold for level of significance, largest p-value and consequently number of selected features or rejected null hypothesis for parsimonious yet powerful model grounded on data.

The following steps present the algorithm to identify the biggest reasonable set of features, that data can afford:

1. Choose the test method for example t-test or Fisher's exact test depending on the data;
2. Obtain p-values by performing the chosen test on all hypotheses;
3. Sort p-values from smallest to largest;
4. Find the smallest p-value;
5. Count the number of hypotheses with the p-value found in step 4 and put them in a set;
6. Continue steps 4 and 5 until the hypotheses with biggest p-values are in the last set;
7. Tabulate the hypotheses to classes of observed p-values;
8. Reject the set of hypotheses with the least p-value (the first set is called S_1);
9. Calculate cumulative expected false discoveries for all the rejected hypotheses ($P_i \times N$);
10. Calculate $1 - \text{CEFD}$ for all the rejected hypotheses;
11. Calculate $\delta = \text{CETD} - \text{CEFD}$;
12. Record the results;
13. Repeat steps 2–7 for all the sets.
14. Find the set with maximum recorded δ (called δ_{max}) resulting from rejecting set S_{max};

15. The biggest reasonable set of rejected hypotheses $\mathrm{R}_{\mathrm{max}}$ will be

$$R_{max}(\textit{the biggest reasonable set of selected features}) = S_1 \cup S_2 \cup S_3 \cup \ldots \cup S_{max};$$

16. The p-value for set S_m is p_m which is the alpha that should be reported;
17. The FDR that should be reported for $\mathrm{R}_{\mathrm{max}}$ is:

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

18. When there is a slow reversal in trend for δ. The researcher can use different methods of piecewise regression on δ vs number selected features to identify an optimum number of rejected hypotheses, or selected features, which may be much less than R_{max} but much more than what would be dictated by a 0.05 FDR;
19. The break point of piecewise regression on δ vs number of selected features identifies the optimum number of selected features.

The process explained in this paper neither requires predetermined thresholds for level of significance, nor uses presumed thresholds for FDR. 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 model. Once the set of rejected hypotheses is determined a related significance level and FDR should be reported.

The paper presented methods that could identify an objective 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 cross tabulation analysis of an ordinal outcome variable with 1044 potential ordinal predictors in a large survey, regarding variables that influence citizen engagement, was used as a novel example of application of the method in social sciences.

Additional file

Additional file 1. The raw data file used in this study.

Authors' contributions

There is a single author for the whole study and paper. The author read and approved the final manuscript.

Author details

[1] School of Communication, Simon Fraser University, Burnaby, BC, Canada. [2] Institute for Canadian Urban Research Studies, Simon Fraser University, Burnaby, BC, Canada. [3] School of Business, Capilano University, 2055 Purcell Way, North Vancouver, BC V7J 3H5, Canada.

Acknowledgements

No other one than author has contributed to the study.

Competing interests

The author declares no competing interests.

Availability of data and materials

The data used as an example is attached and is available at Simon Fraser University institutional repository (Additional file 1).

Consent for publication

Hereby I declare my consent.

Ethics approval and consent to participate

The data used in hypothesis testing example has been collected with the approval of Office of Research Ethics at Simon Fraser University (Study Number: 2014s0383), and all the participants have consented to the approved consent form before participation.

Funding

This study was not funded.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 27 November 2017 Accepted: 19 May 2018
Published online: 31 May 2018

References

1. Austin SR, Dialsingh I, Altman N. Multiple hypothesis testing: a review. J Indian Soc Agric Stat. 2014;68:303–14.
2. Benditkis J. Martingale methods for control of false discovery rate and expected number of false rejections. Dissertation. Heinrich Heine University Duesseldorf. 2015. http://docserv.uni-duesseldorf.de/servlets/DocumentServlet?id=35438.
3. Benditkis J, Heesen P, Janssen A. The false discovery rate (FDR) of multiple tests in a class room lecture. 2015. arXiv preprint arXiv:1511.07050.
4. Blanchard G, Roquain E. Adaptive false discovery rate control under independence and dependence. J Mach Learn Res. 2009;10:2837–71.
5. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B. 1995;57(1):289–300.
6. Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A. Feature selection for high-dimensional data. 1st ed. Berlin: Springer; 2015.
7. Dutheil JY, Hobolth A. Ancestral population genomics. Methods Mol Biol (Clifton, N.J.). 2012;856:293–313. https://doi.org/10.1007/978-1-61779-585-5_12.
8. Fan J, Sun O, Zhou W, Zhu Z. Principal component analysis for big data. 2018. arXiv:1801.01602.
9. Fiori A, Grand A, Bruno G, Brundu FG, Schioppa D, Bertotti A. Information extraction from microarray data: a survey of data mining techniques. J Database Manag (JDM). 2014;25(1):29–58. https://doi.org/10.4018/jdm.2014010102.
10. Hua J, Tembe W, Dougherty ER. Feature selection in the classification of high-dimension data. In: 2008 IEEE international workshop on genomic signal processing and statistics; 2008. p. 1–2. https://doi.org/10.1109/gensips.2008.4555665.
11. Iterson M, Boer JM, Menezes RX. Filtering, FDR and power. BMC Bioinform. 2010;11(September):450. https://doi.org/10.1186/1471-2105-11-450.
12. Kim S, Halabi S. High dimensional variable selection with error control. Biomed Res Int. 2016. https://doi.org/10.1155/2016/8209453.
13. Kim SB, Chen VCP, Park Y, Ziegler TR, Jones DP. Controlling the false discovery rate for feature selection in high-resolution NMR spectra. Stat Anal Data Mining. 2008;1(2):57–66. https://doi.org/10.1002/sam.10005.
14. Miao J, Niu L. A survey on feature selection. Procedia computer science, promoting business analytics and quantitative management of technology: 4th international conference on information technology and quantitative management (ITQM 2016). 2016; 91(January): 919–26. https://doi.org/10.1016/j.procs.2016.07.111.
15. Neuvial P. Asymptotic results on adaptive false discovery rate controlling procedures based on kernel estimators. JMLR. 2013;14:1423–59.
16. Norris AW, Kahn CR. Analysis of gene expression in pathophysiological states: balancing false discovery and false negative rates. Proc Natl Acad Sci USA. 2006;103:649–53.
17. Ochoa A, Storey JD, Llinás M, Singh M. Beyond the E-value: stratified statistics for protein domain prediction. PLoS Comput Biol. 2015;11(11):e1004509. https://doi.org/10.1371/journal.pcbi.1004509.
18. Park BS, Mori M. Balancing false discovery and false negative rates in selection of differentially expressed genes in microarrays. Open Access Bioinformatics. 2010;2:1–9. https://doi.org/10.2147/OAB.S7181.
19. Shaffer J. Multiple hypothesis testing. Annu Rev Psychol. 1995;46:561–84.
20. Shmueli G, Bruce PC, Yahav I, Patel NR, Lichtendahl KC Jr. Data mining for business analytics: concepts, techniques, and applications in R. New York: Wiley; 2017.
21. Storey JD. The optimal discovery procedure: a new approach to simultaneous significance testing. J R Stat Soc Ser B Stat Methodol. 2007;69(3):347–68.
22. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci. 2003;100:9440–5.
23. Storey JD. False discovery rate. In: Lovric M, editor. International encyclopedia of statistical science. Heidelberg: Springer; 2011.