Covariate-Adaptive Method in Multiple Testing: AdaFDR

Kyurhi Kim

Seoul National University

kyurhi99@snu.ac.kr

January 19, 2023

Overview

1. Introduction
2. Previous Methods
3. AdaFDR
   - Scheme
   - Definitions and Notations
   - Multiple Testing via AdaFDR
   - Algorithm
   - Optimization
   - FDP control
4. Results On The Paper
5. Simulation

Introduction

- In multiple testing problem, general goal is to maximize the number of discoveries while controlling False Discovery.
- Well-known standard multiple testing procedures such as the BH method are based only on the p-values.
- However, they fail to utilize additional information (i.e. covariates) that is often available but not directly captured by the p-value.

Conventional Methods

Benjamini-Hochberg (BH)

- Order the original p-values then find $i_{max} = \left\{ i : p_i \leq \frac{i\alpha}{N} \right\}$.
- Then reject $p_1, ..., p_{i_{max}}$.

Storey-BH (SBH)

- Measure q-value: $\Pr(H_{0i} \text{ is true } | z \geq z_0)$
- Reject all hypotheses with a q-value that is less than or equal to the cutoff value for the false discovery rate.

- These methods use only p-values and have the same p-value threshold for all hypotheses with respect to covariates.

Covariate Adaptive Methods

IHW

- Group hypotheses into a prespecified number of groups and applies a constant threshold for each group.
- Only supports for univariate covariates and uses a stepwise-constant function for the threshold, which limits detection power.

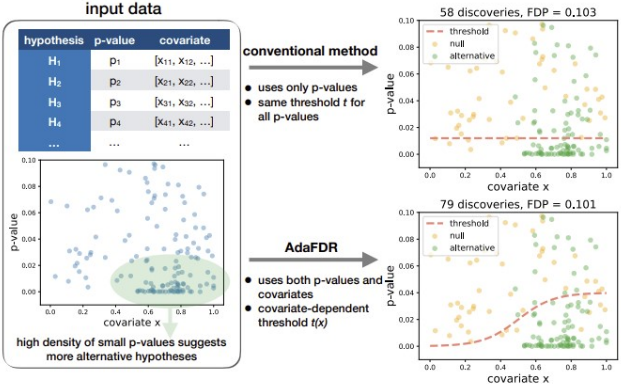
Boca and Leek

- A regression framework to estimate the null proportion conditional on the covariate, and weight the BH-adjusted p-values by $\pi_0(x_i)$.
- Not using covariate-dependent alternative distribution information.

AdaPT

- Mask p-values to control FDR in iterative method.
- Able to use the entire data, but computationally expensive due to many iterations of optimization in p-value masking procedure.

Scheme of AdaFDR



input data

hypothesis	p-value	covariate
H_1	p_1	$[x_{11}, x_{12}, \dots]$
H_2	p_2	$[x_{21}, x_{22}, \dots]$
H_3	p_3	$[x_{31}, x_{32}, \dots]$
H_4	p_4	$[x_{41}, x_{42}, \dots]$
...	...	...

high density of small p-values suggests more alternative hypotheses

conventional method

- uses only p-values
- same threshold *t* for all p-values

58 discoveries, FDP = 0.103

AdaFDR

- uses both p-values and covariates
- covariate-dependent threshold *t(x)*

79 discoveries, FDP = 0.101

- Input: hypotheses, each with a p-value and a covariate vector
- Ouput: a set of rejected hypotheses

AdaFDR

- Data may have an enrichment of small p-values for certain values of the covariate, which suggests an enrichment of alternative hypotheses around these covariate values.
- AdaFDR learns the covariate-dependent threshold by fitting a mixture model using an EM algorithm.
    - The mixture model is a combination of a generalized linear model and Gaussian mixtures.
- Then it produces local adjustments in the p-value threshold by optimizing for more discoveries.

Definitions and Notations

- There are N hypothesis tests where each of them with p-value P_i, a d-dimensional covariate x_i, and an indicator variable h_i, with $h_i = 1$ for the hypothesis to be true alternative.
- Set of true null hypothesis : $\mathcal{H}_0 \overset{\text{def}}{=} \{i : i \in \{1, 2, .., N\}, h_i = 0\}$
  Set of true alternative hypothesis : $\mathcal{H}_1 \overset{\text{def}}{=} \{i : i \in \{1, 2, .., N\}, h_i = 1\}$
- Reject ith null hypothesis if $P_i \leq t(x_i)$, for threshold function $t(x)$.
- The number of discoveries : $\mathrm{D(t)} \overset{\text{def}}{=} \sum_{i \in \{1,\ldots,N\}} \mathbb{I}_{\{P_i \leq t(x_i)\}}$
- $\mathrm{FDP(t)} \overset{\text{def}}{=} \frac{\mathrm{FD(t)}}{\mathrm{D(t)} \vee 1}$ where $\mathrm{FD(t)} \overset{\text{def}}{=} \sum_{i \in \mathcal{H}_0} \mathbb{I}_{\{P_i \leq t(x_i)\}}$
- $\mathrm{FDR} \overset{\text{def}}{=} \mathbb{E}(\mathrm{FDP})$

Multiple Testing via AdaFDR

- **Assumption**: null p-values follows uniform regardless of the covariate value. (alternative p-values and likelihood for hypotheses to be true null/alternative may have dependencies on the covariate.)
- **Aim**: Optimize over a set of decision rules $t(x) \in \mathcal{T}$(set of decision threhsolds) to maximize the number of discoveries, with constraint that the FDP is less than α. (i.e. $\underset{t \in \mathcal{T}}{\mathit{maximize}}$ D(t) s.t. FDP(t) $\leq \alpha$)
- Challenges in this optimization
  1. $\mathcal{T}$ needs to be parametrized in such a way that captures the covariate information and scales well with the covariate dimension.
  2. The actual FDP is not available from the data.
  3. Overfitting that might lead to fail FDR control.

Multiple Testing via AdaFDR

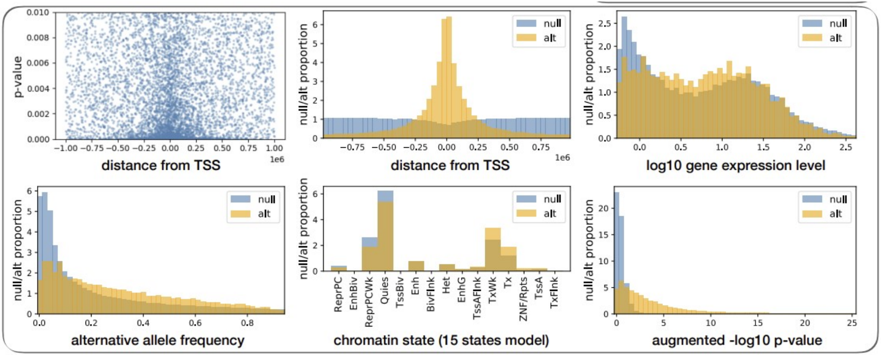
- **The idea for the first challenge** : The decision threshold should have large values where the alternative hypotheses are enriched
    - Such enrichment pattern usually consists of local bumps at certain covariate locations and a global slope that represents generic monotonic relationships.
- AdaFDR addresses these structures by using a mixture of GLM and K-component Gaussian mixture with diagonal covariance matrices.

$$t(x) = exp(a^T x + b) + \sum_{k=1}^{K} exp[w_k - (x - \mu_k)^T diag(\sigma_k)(x - \mu_k)]$$

- The set of parameters to optimize:

$$a \in \mathbb{R}^d, b \in \mathbb{R}, \left\{ w_k \in \mathbb{R}, \mu_k \in \mathbb{R}^d, \sigma_k \in \mathbb{R}^d \right\}_{k=1}^{K}$$

Relation between p-values and covariates



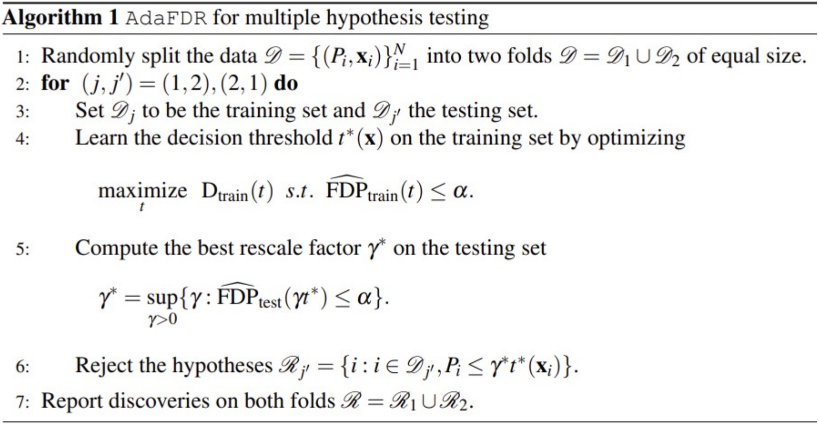
Multiple Testing via AdaFDR

- **For the second challenge,** use mirror statistic to estimate the number of False Discoveries:

$$\widehat{\mathsf{FD}(t)} \overset{\text{def}}{=} \sum_{i=1}^{n} \mathbb{I}_{\{P_i \geq 1-t(x_i)\}} \ , \ \widehat{\mathsf{FDP}(t)} = \frac{\widehat{\mathsf{FD}(t)}}{\mathrm{D(t)}}$$

- **For the third challenge,** AdaFDR controls FDP via hypotheses splitting.
  - The hypotheses are randomly split into two folds.
  - A separate decision threshold is learned on each fold and applied to the other.
  - Covered in later Theorem.

Algorithm of AdaFDR



Algorithm 1 `AdaFDR` for multiple hypothesis testing

```
Randomly split the data $\mathscr{D} = \{(P_i, \mathbf{x}_i)\}_{i=1}^N$ into two folds $\mathscr{D} = \mathscr{D}_1 \cup \mathscr{D}_2$ of equal size.
for $(j, j') = (1,2), (2,1)$ do
    Set $\mathscr{D}_j$ to be the training set and $\mathscr{D}_{j'}$ the testing set.
    Learn the decision threshold $t^*(\mathbf{x})$ on the training set by optimizing
        $$\underset{t}{\text{maximize}} \ \ \mathrm{D}_{\text{train}}(t) \ \ s.t. \ \ \widehat{\mathrm{FDP}}_{\text{train}}(t) \leq \alpha.$$
    Compute the best rescale factor $\gamma^*$ on the testing set
        $$\gamma^* = \sup_{\gamma > 0}\{\gamma : \widehat{\mathrm{FDP}}_{\text{test}}(\gamma t^*) \leq \alpha\}.$$
    Reject the hypotheses $\mathscr{R}_{j'} = \{i : i \in \mathscr{D}_{j'}, P_i \leq \gamma^* t^*(\mathbf{x}_i)\}$.
Report discoveries on both folds $\mathscr{R} = \mathscr{R}_1 \cup \mathscr{R}_2$.
```

Optimization

- In the previous slide, the aim was written as $\underset{t \in \mathcal{T}}{\mathit{maximize}}\ \mathrm{D(t)}$ s.t. $\mathrm{FDP(t)} \leq \alpha$.
- Note that optimization is conducted soley on the training set $\mathcal{D}_{train}$, and FDP is replaced by mirror estimate.
- Then the optimization problem can be rewritten as

$$\underset{t \in \mathcal{T}}{\mathit{maximize}}\ \mathrm{D}_{train}(\mathrm{t}) \text{ s.t.} \frac{\widehat{\mathrm{FD}}_{train}(\mathrm{t})}{\mathrm{D}_{train}(\mathrm{t})} \leq \alpha$$

- To achieve this, first need to compute a good initialization point and then perform optimization (by gradient descent in the paper).

i. Initialization

- **Idea**: It is intuitive to let threshold t(x) $\propto \frac{\pi_1(x)}{\pi_0(x)}$, since the threshold should be large when the number of alternative hypotheses is high and the number of null hypotheses is low.

  ($\pi_0(x)$, $\pi_1(x)$: covariate distribution for the null and alternative hypotheses respectively)

- **Process**
  1. Treat $\{x_i : i \in \mathcal{D}_{train}, P_i \geq 0.75\}$ and $\{x_i : i \in \mathcal{D}_{train}, P_i \leq t_{BH}\}$ as an approximate ensemble of the null hypotheses and alternative hypotheses respectively.
  2. Mixture model is fitted on the null ensemble using an EM algorithm, resulting in an estimate of the null hypothesis distribution $\widehat{\pi_0}(x)$.
  3. Each point in the alternative ensemble receives a sample weight $\frac{1}{\widehat{\pi_0}(X)}$.
  4. Mixture model is fitted on the weighted alternative ensemble using an EM algorithm to obtain the final initialization threshold.

ii. Optimization

1. $\underset{t\in\mathcal{T}}{minimize}[-\mathrm{D}_{train}(t)+\left\{\lambda_1(\widehat{\mathrm{FD}}_{train}(t)-\alpha\mathrm{D}_{train}(t))\vee 0\right\}]$, where λ_1 is chosen heuristically to be $\frac{10}{\alpha}$.
2. The sigmoid function is used to control the discontinuity of the indicator functions in $\mathrm{D}_{train}(t)$ and $\widehat{\mathrm{FD}}_{train}(t)$:

$$\mathrm{D}_{train}(t)=\sum_{i\in\mathcal{D}_{train}}\mathbb{I}_{\{P_i\leq t(x_i)\}}\approx\sum_{i\in\mathcal{D}_{train}}S[\lambda_0(t(x_i)-P_i)]$$

$$\widehat{\mathrm{FD}}_{train}(t)=\sum_{i\in\mathcal{D}_{train}}\mathbb{I}_{\{P_i\geq 1-t(x_i)\}}\approx\sum_{i\in\mathcal{D}_{train}}S[\lambda_0(P_i-1+t(x_i))]$$

   where sigmoid function $S(\cdot)=\frac{1}{1+e^{-x}}$, and λ_0 is automatically chosen at the beginning of the optimization.
3. Adam optimizer is used for gradient descent.

FDP control

- When the number of rejections is small (< 100), the result should be treated with precaution.
- For the theoretical result, it is require that for each fold, the best scale factor γ^* should have a number of discoveries exceeding $c_0 N$ for some pre-specified small proportion c_0.
- If this condition is not satisfied, there will be no rejection in the fold.
- i.e. γ^* in algorithm is substituted to a modified version as follows:

$$\gamma^* = \underset{\gamma \geq 0}{sup} \left\{ \gamma : \widehat{\mathrm{FDP}}_{test}(\gamma t^*) \leq \alpha, \mathrm{D}_{test}(\gamma t^*) \geq c_0 N \right\} \cup \{0\}$$

Theorem1

Theorem (FDP control)

Assume that all null p-values $P_i \in \mathcal{H}_0$, conditional on the covariates, are independently and identically distributed following Unif[0,1].
Then with probability at least $1 - \delta$, **AdaFDR** *with the modificated scale factor γ^* controls FDP at level $(1 + \epsilon)\alpha$, where $\epsilon = O(\sqrt{(log\frac{1}{\delta})/(\alpha N)})$.*

- The assumption can be easily relaxed to the assumptin that the null p-values, conditional on the covariates, are independently distributed and stochastically greater than Unif[0,1].
- Proof for above Thm1 is avilable in [Zhang, 2019]

Proof

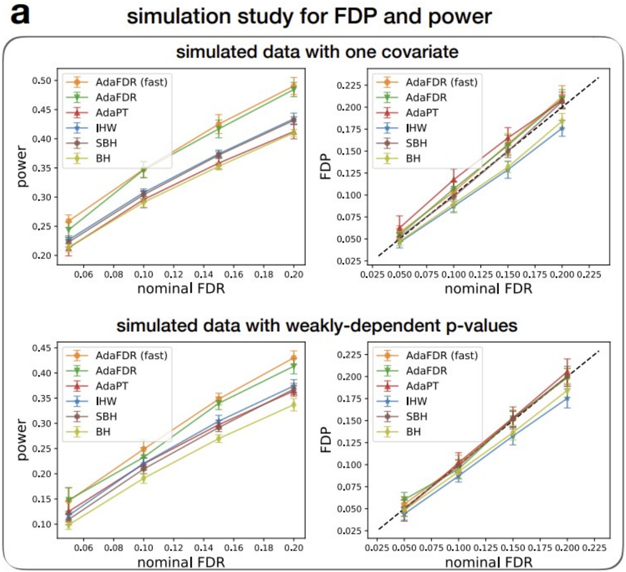
This is overall step of the proof.

step1. It suffices to show that $\mathbb{P}(\mathrm{FDP}_2 \geq (1+\epsilon)\alpha) \leq \frac{\delta}{2}$

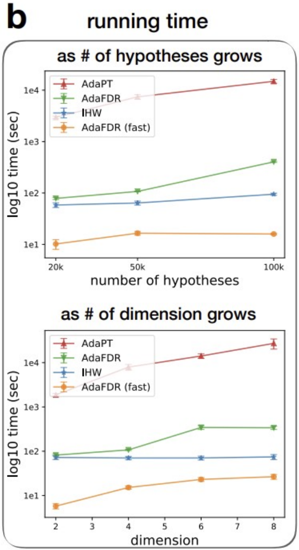
step2. Covert the above probability to some analyzable stochastic process by introducing the set of random variables to condition on.
The random variable set might include hypotheses splitting, all covariates, the type of hypotheses, and the alternative p-values.

step3. Get upper bound of the probability.

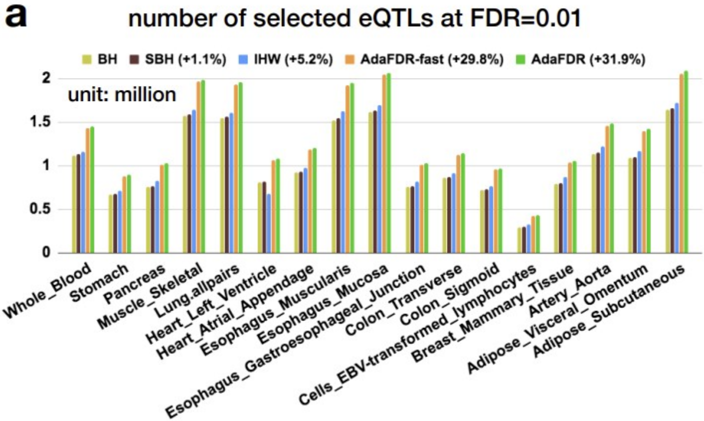
Results


a
simulation study for FDP and power
simulated data with one covariate
AdaFDR (fast)
AdaFDR
AdaPT
IHW
SBH
BH
power
FDP
nominal FDR
0.20 0.25 0.30 0.35 0.40 0.45 0.50
0.06 0.08 0.10 0.12 0.14 0.16 0.18 0.20
0.025 0.050 0.075 0.100 0.125 0.150 0.175 0.200 0.225
simulated data with weakly-dependent p-values
0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45

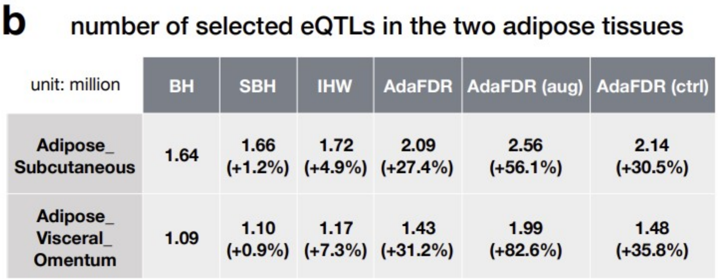
Results


b
running time
as # of hypotheses grows
AdaPT
AdaFDR
IHW
AdaFDR (fast)
log10 time (sec)
1e4
1e3
1e2
1e1
20k
50k
100k
number of hypotheses
as # of dimension grows
AdaPT
AdaFDR
IHW
AdaFDR (fast)
log10 time (sec)
1e4
1e3
1e2
1e1
2
3
4
5
6
7
8
dimension

Results


a
number of selected eQTLs at FDR=0.01
BH
SBH (+1.1%)
IHW (+5.2%)
AdaFDR-fast (+29.8%)
AdaFDR (+31.9%)
unit: million
2
1.5
1
0.5
0
Whole_Blood
Stomach
Pancreas
Muscle_Skeletal
Lung.allpairs
Heart_Left_Ventricle
Heart_Atrial_Appendage
Esophagus_Muscularis
Esophagus_Mucosa
Esophagus_Gastroesophageal_Junction
Colon_Transverse
Colon_Sigmoid
Cells_EBV-transformed_lymphocytes
Breast_Mammary_Tissue
Artery_Aorta
Adipose_Visceral_Omentum
Adipose_Subcutaneous

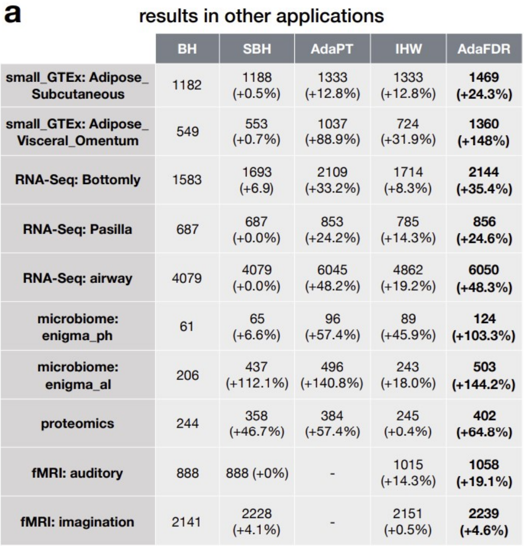
Results



b number of selected eQTLs in the two adipose tissues

unit: million	BH	SBH	IHW	AdaFDR	AdaFDR (aug)	AdaFDR (ctrl)
Adipose_ Subcutaneous	1.64	1.66 (+1.2%)	1.72 (+4.9%)	2.09 (+27.4%)	2.56 (+56.1%)	2.14 (+30.5%)
Adipose_ Visceral_ Omentum	1.09	1.10 (+0.9%)	1.17 (+7.3%)	1.43 (+31.2%)	1.99 (+82.6%)	1.48 (+35.8%)

Results



a

results in other applications

	BH	SBH	AdaPT	IHW	AdaFDR
small_GTEx: Adipose_ Subcutaneous	1182	1188 (+0.5%)	1333 (+12.8%)	1333 (+12.8%)	**1469 (+24.3%)**
small_GTEx: Adipose_ Visceral_Omentum	549	553 (+0.7%)	1037 (+88.9%)	724 (+31.9%)	**1360 (+148%)**
RNA-Seq: Bottomly	1583	1693 (+6.9)	2109 (+33.2%)	1714 (+8.3%)	**2144 (+35.4%)**
RNA-Seq: Pasilla	687	687 (+0.0%)	853 (+24.2%)	785 (+14.3%)	**856 (+24.6%)**
RNA-Seq: airway	4079	4079 (+0.0%)	6045 (+48.2%)	4862 (+19.2%)	**6050 (+48.3%)**
microbiome: enigma_ph	61	65 (+6.6%)	96 (+57.4%)	89 (+45.9%)	**124 (+103.3%)**
microbiome: enigma_al	206	437 (+112.1%)	496 (+140.8%)	243 (+18.0%)	**503 (+144.2%)**
proteomics	244	358 (+46.7%)	384 (+57.4%)	245 (+0.4%)	**402 (+64.8%)**
fMRI: auditory	888	888 (+0%)	-	1015 (+14.3%)	**1058 (+19.1%)**
fMRI: imagination	2141	2228 (+4.1%)	-	2151 (+0.5%)	**2239 (+4.6%)**

References

Zhang, M. J., Xia, F., Zou, J. (2019, April)
AdaFDR: A Fast, Powerful and Covariate-Adaptive Approach to Multiple Hypothesis Testing
In RECOMB (pp. 330-333).