FDR control via Data Splitting

SHIN DO HYUP

Department of Statistics
Seoul National University

2022.07.28

Outline

FDR control via Data Splitting

- Review of Multiple testing
- Motivation of Data Splitting
- Data Splitting

Review of Multiple testing

- In High dimensional linear regression model, we explain variable selection procedure using the “FDR control via Data splitting” (C Dai, 2020).
- Before explaining the paper, we review the several multiple testing procedures

Multiple Testing

- The simultaneous testing of more than one hypothesis
- Having observed a large number N of test statistics, how should we decide which if any of the null hypotheses to reject
- Assume that there are N hypotheses testing

$$H_{0i} \text{ vs } H_{1i}, \quad i = 1, 2, \ldots, N$$

Multiple Testing

		Decision		
		Null	Non-Null	
Actual	Null	$N_0 - a$	a	N_0
	Non-Null	$N_1 - b$	b	N_1
		$N - R$	R	N

- Decision rule $\mathcal{D}$ has rejected R out of N null hypotheses. (N, R is known)
- a of these decisions were incorrect. (false discoveries)
- N_0, N_1, a, b are unknown random variables.

Family-Wise Error Rate(FWER)

- The FWER criterion aims to control the probability of making even one false rejection among N simultaneous hypothesis tests.
- The FWER is

$$FWER = Pr\{\text{ reject any true } H_{0i}\} = P(a \geq 1)$$

- There are several methods for control FWER at level α. (Bonferroni's procedure, Holm's procedure)
- This means we control $FWER = P(a \geq 1) \leq \alpha$.
- But FWER usually proved too conservative when N is large.

False Discovery Rates

- Define the false discovery proportion(Fdp) $Fdp(\mathcal{D}) = a/R$.
- Since Fdp is unobservable, we control the false discovery rate(fdr) at level q $(0 < q < 1)$ defined

$$FDR(\mathcal{D}) = E(Fdp(\mathcal{D})) \leq q$$

- How can we choose the decision rule $\mathcal{D}$?
- The idea of finding the decision rule is to order the observed p-values from smallest to largest.
- Let p_i be the p-value corresponding the H_{0i} for all $i = 1, 2, \ldots, N$ and $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(N)}$ be the ordering p-values.

Benjamini–Hochberg FDR Control

Theorem(Benjamini–Hochberg FDR Control, 1995)

Let $H_{0(i)}$ be the null hypothesis corresponding the ith ordering p-value $p_{(i)}$. Define $i_{max} = \max\{i : p_{(i)} \leq \frac{i}{N}q\}$ and let $\mathcal{D}_q$ be the rule that rejects $H_{0(i)}$ for $i \leq i_{max}$. If the p-values corresponding to valid null hypotheses are independent of each other, then

$$FDR(\mathcal{D}_q) = \pi_0 q \leq q; \quad \text{where } \pi_0 = N_0/N$$

False Discovery Rates

- The Benjamin Hochberg (BHq) procedure is the most commonly used basic method of FDR control
- But the BHq procedure requires the assumption of independence for all the p-values.
- Benjamini and Yekutieli(2001) generalized BHq to handle dependent p-values by using a shrinkage of the control level $\tilde{q} = \frac{q}{\sum_{j=1}^{N} 1/j}$.
- There are many methods of the FDR control.
    - Sarkar(2002), Storey (2004) and so on.

Outline

FDR control via Data Splitting

- Review of Multiple testing
- Motivation of Data Splitting
- Data Splitting

Notation and Definition

- Let $\mathbf{X}_{n\times p} = (X_1, X_2, \ldots, X_p)$ be the explanatory features with p being large.
- For each row of the design matrix $\mathbf{X}$ independently follows a p-dimensional distribution with a covariance matrix Σ.
- For each feature has been normalized with zero mean and unit variance.
- Let $Y = (y_1, \ldots, y_n)$ be the vector of n independent response variable. Consider the linear regression model

$$Y = X_{n\times p}\beta + \epsilon$$

Notation and Definition

- $S_0 = \{i : \beta_i = 0\}$: the index set of the null features
- $S_1 = \{i : \beta_i \neq 0\}$: the index set of the relevant features
- $p_0 = |S_0|$ and $p_1 = |S_1|$
- $\hat{S}$: the index set of the selected features (estimator of S_1)
- In this case, the hypotheses are

$$H_{0i} : \beta_i = 0 \text{ vs } H_{1i} : \beta_i \neq 0$$

Motivation of Data Splitting

- In High-Dimesional linear regression model, we can select the significant variable by applying the FDR control procedures.

$$FDR = \mathbb{E}[FDP], \quad FDP = \frac{\#\{j : j \in S_0, j \in \hat{S}\}}{\#\{j : j \in \hat{S}\} \vee 1}$$

- But it is difficult to calculate p-values and estimate the joint distribution of features.
- For these reasons, there is a limit to applying the BHq procedure.
- To solve these problems, we use the data-splitting

Outline

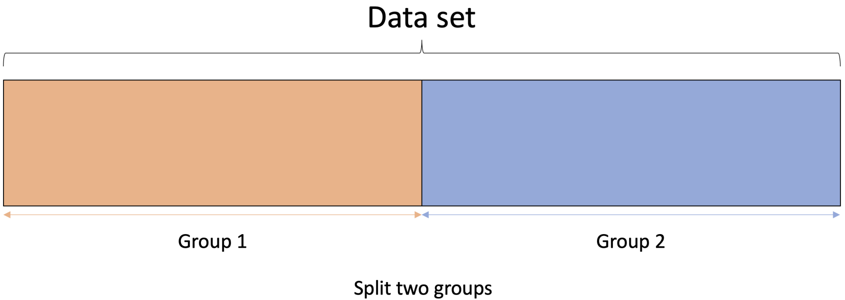
FDR control via Data Splitting

- Review of Multiple testing
- Motivation of Data Splitting
- Data Splitting

Single Data Splitting(DS)

- Features selection depend on $\hat{\beta}$. For example, $\hat{\beta}$ can be estimated via OLS or some shrinkage methods.
- In contrast to those commonly methods, we split the data into two groups of equal size denoted as $(\mathbf{y}^{(1)}, \boldsymbol{X}^{(1)})$, and $(\mathbf{y}^{(2)}, \boldsymbol{X}^{(2)})$.
- So we can estimate two independent regression coefficients denoted $\hat{\beta}^{(1)}, \hat{\beta}^{(2)}$ for each groups.
- To achieve FDR control under our data-splitting, two independent coefficients should satify the following assumption.

Single Data Splitting(DS)



Mirror Statistics

Assumption 1 (Symmetry)

For each null feature index $j \in S_0$, the sampling distribution of either $\hat{\beta}_j^{(1)}$ or $\hat{\beta}_j^{(2)}$ is symmetric about 0

- For $j \in S_0$, only one of $\hat{\beta}_j^{(1)}$ and $\hat{\beta}_j^{(2)}$ is symmetric about 0.
- Define the mirror statistics M_j by

$$M_j = \text{sign}(\hat{\beta}_j^{(1)}\hat{\beta}_j^{(2)})f(|\hat{\beta}_j^{(1)}|, |\hat{\beta}_j^{(2)}|)$$

  where the function $f(u, v)$ is non-negative, symmetric about u and v, and monotone increasing in both u and v.
- For a relevant feature, the corresponding mirror statistic tends to be positive and relatively large.

Mirror Statistics

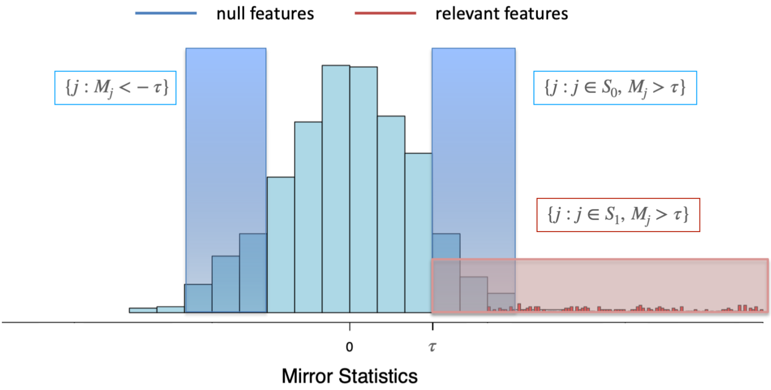
Lemma 1

Under Assumption 1, regardless of the data-splitting procedures, the sampling distribution of M_j is symmetric about 0 for $j \in S_0$

- We use mirror statistics as test statistics.
- The mirror statistics satisfy the following two properties.
  - (A1) A feature with a larger mirror statistic is more likely to be a relevant feature.
  - (A2) The sampling distribution of the mirror statistic of any null feature is symmetric about 0.

Mirror Statistics

1



Mirror Statistics

- There are several choice of $f(u, v)$. The optimal choice that obtains the highest power is $f(u, v) = u + v$.
- By using the Assumption 1, we can make an upper bound of the number of false positives

$$\#\{j \in S_0 : M_j > t\} \approx \#\{j \in S_0 : M_j < -t\} \leq \#\{j : M_j < -t\}, \ \forall t > 0.$$

- For given $t > 0$, we can define the FDP(t) of the selection $\hat{S}_t = \{j : M_j > t\}$ and the estimate of the FDP(t) as $\widehat{FDP}(t)$.

$$FDP(t) = \frac{\#\{j : M_j > t, j \in S_0\}}{\#\{j : M_j > t\} \vee 1}, \quad \widehat{FDP}(t) = \frac{\#\{j : M_j < -t\}}{\#\{j : M_j > t\} \vee 1}$$

Mirror Statistics

- Let $\forall q \in (0,1)$ be given FDR control level.
- Then, we can find the cutoff value τ_q as follows

$$\tau_q = \min\{t > 0 : \widehat{FDP}(t) \leq q\}$$

- Thus, we finally select $\hat{S}_{\tau_q} = \{j : M_j > \tau_q\}$ as a set of the index of relevant features.

Algorithm of FDR control via Single DS

Algorithm 1

1. Split the data into two groups, independent to the response vector y.
2. Estimate the "impact" coefficient $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ on each part of the data. The two estimation procedures can be potentially different.
3. Calculate the mirror statistics following

$$M_j = sign(\hat{\beta}_j^{(1)}\hat{\beta}_j^{(2)})(|\hat{\beta}_j^{(1)}| + |\hat{\beta}_j^{(2)}|)$$

4. Given a designated FDR level $q \in (0, 1)$, calculate the cutoff τ_q as :

$$\tau_q = \min\{t > 0 : \widehat{FDP}(t) = \frac{\#\{j : M_j < -t\}}{\#\{j : M_j > t\} \vee 1} \leq q\}$$

5. Select the features $\{j : M_j > \tau_q\}$

Single Data Splitting

- To obtain a good estimate of the number of false positives, the mirror statistics of the null features cannot be too correlated. So we require the following weak dependence assumption.

Assumption 2 (Weak dependence among the null features)

The mirror statistics $M_j's$ are continuous random variables, and there exist constants $c > 0$ and $\alpha \in (0, 2)$ such that

$$Var\left(\sum_{j \in S_0} 1(M_j > t)\right) \leq cp_0^{\alpha}, \ \forall t \in R, \text{ where } p_0 = |S_0|$$

- Additionally, we assume the variances of the mirror statistics are bounded.

Single Data Splitting

Proposition

For any designated FDR control level $q \in (0,1)$, assume that there exists a constant $t_q > 0$ such that $\mathbb{P}(FDP(t_q) \leq q) \to 1$ as $p \to \infty$. Then, under Assumption 1 and 2, the procedure in Algorithm 1 satisfies

$$FDP(\tau_q) \leq q + o_p(1) \quad \text{and} \quad \limsup_{p \to \infty} FDR(\tau_q) \leq q$$

- We note that the existence of $t_q > 0$ such that $\mathbb{P}(FDP(t_q) \leq q) \to 1$ essentially guarantees the asymptotic feasibility of FDR control based upon the rankings of features by their mirror statistics.

Multiple Data Splitting(MDS)

- There are two problems about DS.
- First, splitting the data into two halves inflates the variance of the estimated regression coefficient. So, DS can potentially suffer a power loss.
- Second, the selection result of DS may not be stable and can vary substantially across different sample splits.
- To solve this problem, we use a multiple data splitting procedure to aggregate the selection results obtained from independent replications of DS.

Multiple Data Splitting(MDS)

- Suppose we independently repeat DS m times with random sample splits.
- Each time the set of selected features is denoted as $\hat{S}^{(k)}$ for $k = 1, 2, \ldots, m$.
- For each feature X_j, we define the associated inclusion rate I_j and its estimate $\hat{I}_j$ as

$$I_j = \mathbb{E}\left[\frac{1(j \in \hat{S})}{|\hat{S}| \vee 1} | X, y\right], \quad \hat{I}_j = \frac{1}{m}\sum_{k=1}^{m}\frac{1(j \in \hat{S}^{(k)})}{|\hat{S}^{(k)}| \vee 1}$$

Multiple Data Splitting(MDS)

- This rate is an importance measurement of each feature relative to the DS selection procedure.
- MDS is most useful if the following statement is approximately true.
  - If a feature is selected less frequently in the repeated sample splitting, it is less likely to be a relevant feature.
- If this holds, we can choose a proper inclusion rate cutoff to control the FDR, and select those features with inclusion rates larger then cutoff.

Algorithm of aggregating selection results from multiple data splits

Algorithm 2

1. Sort the estimated inclusion rates : $0 \leq \hat{I}_{(1)} \leq \hat{I}_{(2)} \leq \cdots \leq \hat{I}_{(p)}$
2. Find the largest $l \in \{1, 2, \ldots, p\}$ such that $\hat{I}_{(1)} + \hat{I}_{(2)} + \cdots + \hat{I}_{(l)} \leq q$
3. Select the features $\hat{S} = \{j : \hat{I}_j > \hat{I}_{(l)}\}$

- The following proposition points out a key factor for MDS to achieve FDR control

Multiple Data Splitting(MDS)

Proposition

Suppose we can asymptotically control the FDP of DS for any designated level $q \in (0,1)$. Furthermore, we assume that with probability approaching 1, the power of DS is bounded below by some $\kappa > 0$. We consider the following two regimes with $n, p \to \infty$ at a proper rate.

(a) In the non-sparse regime where $\liminf p_1/p > 0$, we assume that the mirror statistics are consistent at ranking features, i.e., $\sup_{i \in S_1, j \in S_0} P(I_i < I_j) \to 0$.

(b) In the sparse regime where $\limsup p_1/p = 0$, we assume that the mirror statistics are strongly consistent at ranking features, i.e., $\sup_{i \in S_1} P(I_i < \max_{j \in S_0} I_j) \to 0$.

Then, for MDS (see Algorithm 2) in both the non-sparse and the sparse regimes, we have

$$FDP \leq q + o_p(1) \text{ and } \limsup_{n,p \to \infty} FDR \leq q$$

Application for Linear models

- Consider the process of above method applying to linear models.
- We proceed the data splitting by using a Lasso + OLS procedure.
- In detail, on the first half of the data $(\mathbf{y}^{(1)}, \boldsymbol{X}^{(1)})$, we apply Lasso for dimension reduction. Let $\hat{\boldsymbol{\beta}}^{(1)}$ be the estimated regression coefficients and denotes $\hat{S}^{(1)} = \{j : \hat{\beta}_j^{(1)} \neq 0\}$.
- Next, we restrict the features to $\hat{S}^{(1)}$ obtained above. Then, we run OLS using the second half of the data $(\mathbf{y}^{(2)}, \boldsymbol{X}^{(2)})$ to obtain the estimated coefficients $\hat{\boldsymbol{\beta}}^{(2)}$.
- So, we can make the mirror statistics by using $\hat{\boldsymbol{\beta}}^{(1)}$ and $\hat{\boldsymbol{\beta}}^{(2)}$.

Conclusion

- The main advantage of DS and MDS is that they do not require the p-values and the joint distribution of features.
- MDS stabilizes the result and improves the power of a single DS.
- DS and MDS control the FDR at the designated level in linear models.
- Both DS and MDS are conceptually simple and easy to implement based on exisiting softwares for high-dimensional regression methods.

Reference

- Chenguang Dai, Buyu Lin, Xin Xing, and Jun S.Liu(2020). FDR control via Data splitting. Journal of the American Statistical Association
- Yoav Benjamini and Yosef Hochberg (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical society
- Nicolai Meinshausen, Lukas Meier and Peter Bühlmann(2009). P-Values for High-Dimensional Regression. Journal of the American Statistical Association.
- Bradley Efron, Trevor Hastie, Computer Age Statistical Inference Algorithms, Evidence, and Data Science.
- Peter Bühlmann, Sara van de Geer. Statistics for High-Dimensional Data: Methods, Theory and Applications.