STK-IN4300
Statistical Learning Methods in Data Science

Riccardo De Bin

debin@math.uio.no

Outline of the lecture

- Feature Assessment when $p \gg N$
  - Feature Assessment and Multiple Testing Problem
  - The false discovery rate
- Stability Selection
  - Introduction
  - Selection probability
  - Stability path
  - Choice of regularization

Feature Assessment when $p \gg N$: multiple testing problem

In the previous lecture:

- talked about the $p \gg N$ framework;
- focused on the construction of prediction models.

More basic goal:

- assess the significance of the M variables;
  - in this lecture M is the number of variables (as in the book);
- e.g., identify the genes most related to cancer.

Feature Assessment when $p \gg N$: multiple testing problem

Assessing the significance of a variable can be done:

- as a by-product of a multivariate model,
  - selection by a procedure with variable selection property;
  - absolute value of a regression coefficient in lasso;
  - if and how fast a variable enter in a boosting model.
- evaluating the variables one-by-one:
  - univariate tests;

↓

multiple hypothesis testing

Feature Assessment when $p \gg N$: multiple testing problem

Consider the data from Rieger et al. (2004):

- study on the sensitivity of cancer patients to ionizing radiation treatment;
- oligo-nucleotide microarray data ($M = 12625$);
- $N = 58$:
  - 44 patients with normal reaction;
  - 14 patients who had a severe reaction.

Feature Assessment when $p \gg N$: multiple testing problem

TABLE 18.4. *Subset of the 12,625 genes from microarray study of radiation sensitivity. There are a total of 44 samples in the normal group and 14 in the radiation sensitive group; we only show three samples from each group.*

	Normal				Radiation Sensitive			
Gene 1	7.85	29.74	29.50	...	17.20	-50.75	-18.89	...
Gene 2	15.44	2.70	19.37	...	6.57	-7.41	79.18	...
Gene 3	-1.79	15.52	-3.13	...	-8.32	12.64	4.75	...
Gene 4	-11.74	22.35	-36.11	...	-52.17	7.24	-2.32	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Gene 12,625	-14.09	32.77	57.78	...	-32.84	24.09	-101.44	...

Feature Assessment when $p \gg N$: multiple testing problem

The simplest way to identify significative genes:

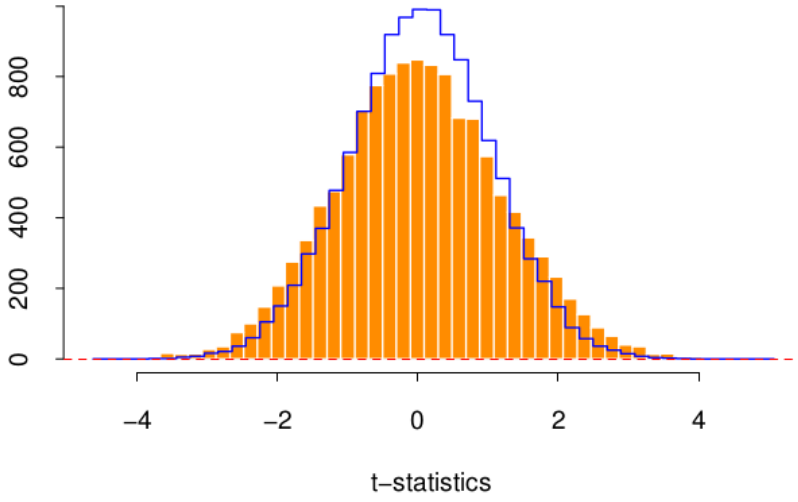
- two-sample t-statistic for each gene,

$$t_j = \frac{\bar{x}_{2j} - \bar{x}_{1j}}{se_j}$$

where

- $\bar{x}_{kj} = \sum_{i \in C_k} x_{kj}/N_k$;
- C_k are the indexes of the N_k observations of group k;
- $se_j = \hat{\sigma}_j \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$;
- $\hat{\sigma}_j^2 = \frac{1}{N_1+N_2-2} \left(\sum_{i \in C_1} (x_{ij} - \bar{x}_{1j})^2 + \sum_{i \in C_2} (x_{ij} - \bar{x}_{2j})^2 \right)$.

Feature Assessment when $p \gg N$: multiple testing problem



Feature Assessment when $p \gg N$: multiple testing problem

From the histogram (12625 t-statistics):

- the values range from -4.7 to 5.0;
- assuming $t_j \sim N(0,1)$, significance at 5% when $|t_j| \geqslant 2$;
- in the example, 1189 genes with $|t_j| \geqslant 2$.

However:

- out of 12625 genes, many are significant by chance;
- supposing (it is not true) independence:
  - expected falsely significant genes, $12625 \cdot 0.05 = 631.25$;
  - standard deviation, $\sqrt{12625 \cdot 0.05 \cdot (1 - 0.05)} \approx 24.5$;
- the actual 1189 is way out of range.

Feature Assessment when $p \gg N$: multiple testing problem

Without assuming normality, permutation test:

- perform $K = \binom{58}{14}$ permutations of the sample labels;
- compute the statistic $t_j^{[k]}$ for each permutation k;
- the p-value for the gene j is

$$p_j = \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}(|t_j^{[k]}| > |t_j|)$$

(not all $\binom{58}{14}$ are needed, random sample of $K = 1000$)

Feature Assessment when $p \gg N$: multiple testing problem

For $j \in 1, \dots, M$ test the hypotheses:

H$_{0j}$: treatment has no effect on gene j

H$_{1j}$: treatment has an effect on gene j

H$_{0j}$ is rejected at level α if $p_j < \alpha$:

- α is the type-I error;
- we expect a probability of falsely rejecting H$_{0j}$ of α.

Feature Assessment when $p \gg N$: family-wise error rate

Define $A_j = \{H_{0j}$ is falsely rejected$\} \longrightarrow Pr(A_j) = \alpha$.

The **family-wise error rate** (FWER) is the probability of at least one false rejection,

$$Pr(A) = Pr(\bigcup_{j=1}^{M} A_j)$$

- for p large, $Pr(A) \gg \alpha$;
- it depends on the correlation between the test;
- if tests independent, $Pr(A) = 1 - (1 - \alpha)^M$;
- test with positive dependence, $Pr(A) < 1 - (1 - \alpha)^M$;
  - positive dependence is typical in genomic studies.

Feature Assessment when $p \gg N$: family-wise error rate

The simplest approach to correct the p-value for the multiplicity of the tests is the **Bonferroni method**:

- reject H_{0j} if $p_j < \alpha/M$;
- it makes the individual test more stringent;
- controls the FWER
  - it is easy to show that FWER $\leqslant \alpha$;
- it is very (too) conservative.

In the example:

- with $\alpha = 0.05$, $\alpha/M = 0.05/12635 \approx 3.9 \times 10^{-6}$;
- no gene has a p-value so small.

Feature Assessment when $p \gg N$: the false discovery rate

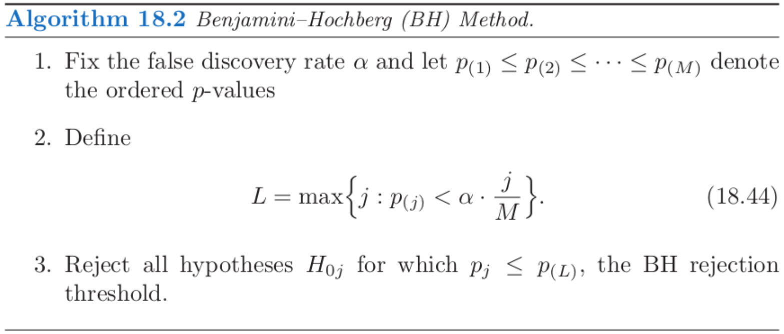
Instead of FWER, we can control the **false discovery rate** (FDR):

- expected proportion of genes incorrectly defined significant among those selected as significant,

	Called Not Significant	Called Significant	Total
H_0 True	U	V	M_0
H_0 False	T	S	M_1
Total	$M - R$	R	M

- in formula, FDR $= E[V/R]$;
- procedure to have the FDR smaller than an user-defined α.

Feature Assessment when $p \gg N$: the false discovery rate



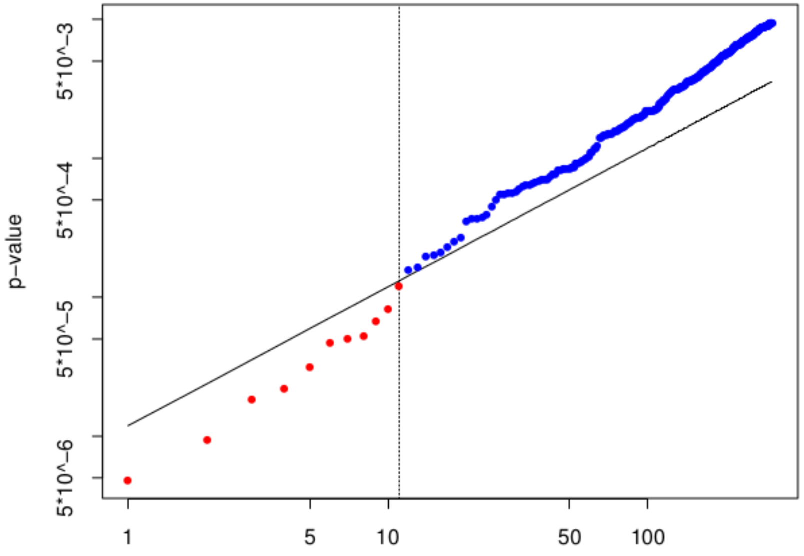
Algorithm 18.2 *Benjamini–Hochberg (BH) Method.*

1. Fix the false discovery rate α and let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(M)}$ denote the ordered p-values
2. Define

$$L = \max\left\{j : p_{(j)} < \alpha \cdot \frac{j}{M}\right\}. \quad (18.44)$$

3. Reject all hypotheses H_{0j} for which $p_j \leq p_{(L)}$, the BH rejection threshold.

Feature Assessment when $p \gg N$: the false discovery rate



Feature Assessment when $p \gg N$: the false discovery rate

In the example:

- $\alpha = 0.15$;
- the last p_j under the line $\alpha \cdot (j/M)$ occurs at $j = 11$;
- the smallest 11 p-values are considered significative;
- in the example, $p_{(11)} = 0.00012$;
- the corresponding t-statistic is $|t_{(11)}| = 4.101$;
- a gene is relevant if the corresponding t-statistics is in absolute value larger than 4.101.

Feature Assessment when $p \gg N$: the false discovery rate

It can be proved (Benjamini & Hochberg, 1995) that

$$\text{FDR} \leqslant \frac{M_0}{M}\alpha \leqslant \alpha$$

- regardless the number of true null hypotheses;
- regardless the distribution of the p-values under H_1;
- suppose independent test statistics;
- in case of dependence, see Benjamini & Yekutieli (2001).

Stability Selection: introduction

In general:

- the L_1-penalty is often use to perform model selection;
- no oracle property (strict conditions to have it);
- issues with selecting the proper amount of regularization;

Meinshausen & Bühlmann (2010) suggested a procedure:

- based on subsampling (could work with bootstrapping as well);
- determines the amount of regularization to control the FWER;
- new structure estimation or variable selection scheme;
- here presented with L_1-penalty, works in general.

Stability Selection: introduction

Setting:

- β is a p-dimensional vector of coefficients;
- $S = \{j : \beta_j \neq 0\}$, $|S| < p$;
- $S^C = \{j : \beta_j = 0\}$;
- $Z^{[i]} = (X^{[i]}, Y^{[i]})$, $i = 1, \dots, N$, are the i.i.d. data,
  - univariate response Y;
  - $N \times p$ covariate matrix X.
- consider a linear model

$$Y = X\beta + \epsilon$$

with $\epsilon = (\epsilon_1, \dots, \epsilon_N)$ with i.i.d. components.

Stability Selection: introduction

The goal is to infer S from the data. We saw that lasso,

$$\hat{\beta}^\lambda = \text{argmin}_{\beta \in \mathbb{R}^p}\left(||Y - X\beta||_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|\right)$$

provides an estimate of S, $S^\lambda = \{j : \hat{\beta}_j \neq 0\} \subseteq \{1, \dots, p\}$.

Remember:

- $\lambda \in \mathbb{R}^+$ is the regularization factor;
- $||X_j||_2^2 = \sum_{i=1}^N (x_j^{[i]})^2 = 1$;

Stability Selection: selection probability

Stability selection is built on the concept of **selection probability**,

Definition 1: Let I be a random subsample of $\{1, \dots, N\}$ of size $\lfloor N/2 \rfloor$ drawn without replacement. We define selection probability the probability for a variable X_j of being in $S^\lambda(I)$,

$$\hat{\Pi}_j^\lambda = Pr^*[j \subseteq S^\lambda(I)]$$

Note:

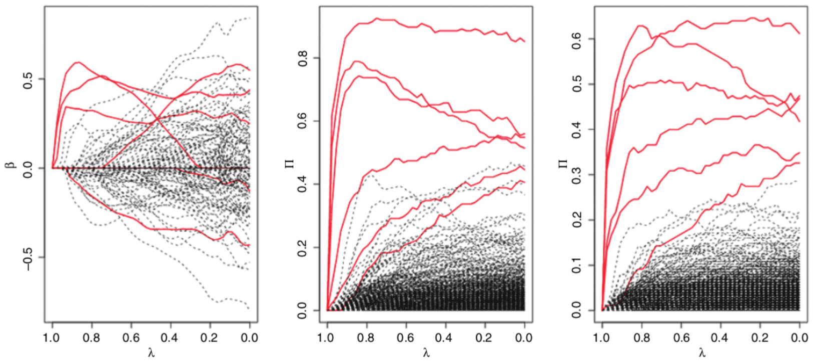
- Pr^* is with respect of both the random subsampling and other sources of randomness if S^λ is not deterministic;
- $\lfloor N/2 \rfloor$ is chosen for computational efficiency.

Stability Selection: stability path

Once we have the selection probability, we can define the **stability path**, as the evolution of $\hat{\Pi}_j^\lambda$ when $\lambda \in \Lambda$ varies,

- similar to the learning path plot of lasso;
- it shows the selection probabilities for all variables;
- it is very useful for improved variable selection, especially in high-dimensional cases.

Stability Selection: stability path



- left: lasso learning path;
- center: stability path of the lasso;
- right: stability path of the randomized lasso.

Stability Selection: stability path

Normally we would choose a specific λ:

- it is a single element of the set $\hat{S}^{\lambda}, \lambda \in \Lambda$;
- S might not be a member of the set;
- even if it is, it is hard to find the right λ high-dimensions.

With **stability selection**:

- we do not simply select one model in $\hat{S}^{\lambda}, \lambda \in \Lambda$;
- the data are perturbed (e.g. by subsampling) many times;
- we choose all variables that occur in a large fraction of the resulting selection sets.

Stability Selection: stability selection

Definition 2: For a cut-off π_{thr}, with $0 < \pi_{thr} < 1$, and a set of regularization parameters Λ, the set of stable variables is defined as

$$\hat{S}^{\text{stable}} = \left\{ j : \max_{\lambda \in \Lambda}(\Pi_j^{\lambda}) \geqslant \pi_{\text{thr}} \right\}.$$

In this way:

- we keep variables with a high selection probability;
- we disregard those with low selection probabilities;
- the exact cut-off π_{thr} is a tuning parameter;
- the results vary surprisingly little for sensible choices of π_{thr};
- results do not strongly depend on the choice of λ or Λ.

Stability Selection: choice of regularization

Let:

- $S^{\Lambda} = \bigcup_{\lambda \in \Lambda} \hat{S}^{\lambda}$ be the set of selected variables $\forall \lambda \in \Lambda$;
- $q_{\Lambda} = E[|\hat{S}^{\Lambda}(I)|]$ be the average number of selected variables;
- $V = |S^C \bigcap \hat{S}^{\text{stable}}|$ the number of falsely selected variables with stability selection.

Theorem (Meinshausen & Bühlmann, 2010): *Assuming that the distribution of $\{\mathbb{1}_{j \in \hat{S}^{\lambda}}\}$ is exchangeable $\forall \lambda \in \Lambda$ and the procedure is not worse than a random guess, then*

$$E[V] \leqslant \frac{1}{2\pi_{\text{thr}} - 1} \frac{q_{\Lambda}^2}{p}$$

Stability Selection: choice of regularization

Therefore:

- π_{thr} is a tuning parameter whose influence is very small;
  - sensible values are in $(0.6, 0.9)$;
- once decided π_{thr}, Λ is determined by the error control desired;
- specifically for $\pi_{\text{thr}} = 0.9$,
  - $\Lambda : q_{\Lambda} = \sqrt{0.8p} \quad \longrightarrow \quad E[V] \leqslant 1;$
  - $\Lambda : q_{\Lambda} = \sqrt{0.8\alpha p} \quad \longrightarrow \quad Pr[V > 0] \leqslant \alpha;$
- i.e., we need to find Λ that gives a specific q_{Λ},
  - q is given by the number of variables which enter in the model;
  - for lasso, find $\lambda_{\text{min}} : |\bigcup_{\lambda_{\text{max}} \geqslant \lambda \geqslant \lambda_{\text{min}}} \hat{S}^{\lambda}| \leqslant q$

Stability Selection: choice of regularization

Final remarks:

- without stability selection, λ depends on the unknown noise level of the observations;
- the advantages of stability selection are:
  - exact error control is possible;
  - the method works fine even though the noise level is unknown;
- real advantage when $p \geqslant N$ (hard to estimate the noise level);

- consistency can be proved (see Meinshausen & Bühlmann, 2010, for the proof for randomized lasso);
- exchangeability in Theorem 1 is only need for the proof.

References I

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* , 289–300.

Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29**, 1165–1188.

Meinshausen, N. & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 417–473.

Rieger, K. E., Hong, W.-J., Tusher, V. G., Tang, J., Tibshirani, R. & Chu, G. (2004). Toxicity from radiation therapy associated with abnormal transcriptional responses to dna damage. *Proceedings of the National Academy of Sciences* **101**, 6635–6640.