# Data splitting for variable selection

Statistical Learning
CLAMSES - University of Milano-Bicocca

Aldo Solari

# References


- Dezeure, Buhlmann, Meier, Meinshausen (2015). High dimensional inference: Confidence intervals, $p$-values and r-software `hdi`. Statistical Science, 533–558

# High-dimensional inference

- Consider the gaussian linear model

$$y \sim N_n(1_n\beta_0 + X\beta, \sigma^2 I_n)$$

  with $n \times p$ design matrix $X$ and $p \times 1$ vector of coefficients $\beta$
- When $p \geq n$, classical approaches for estimation and inference of $\beta$ cannot be directly applied
- How to perform inference on $\beta$ (e.g. confidence intervals and $p$-values for individual regression parameters $\beta_j$, $j = 1, \ldots, p$) in a high-dimensional setting?

# Support set

- The *support set* is

$$S = \{j \in \{1, \ldots, p\} : \beta_j \neq 0\}$$

  with cardinality $s = |S|$, and its complement is the *null set*, i.e.

$$N = \{j \in \{1, \ldots, p\} : \beta_j = 0\}$$

- Let $\hat{S} \subseteq \{1, \ldots, p\}$ be an estimator of $S$. Then

$$|\hat{S} \cap N|$$

  is the number of the wrong selections (type I errors) and

$$|S \setminus \hat{S}|$$

  is the number of wrong deselections (type II errors)

# Error rates

- Define the *False Discovery Proportion* (FDP) by

$$\mathrm{FDP}(\hat{S}) = \frac{|\hat{S} \cap N|}{|\hat{S}|}$$

  with $\mathrm{FDP}(\emptyset) = 0$

- *FamilyWise Error Rate* (FWER)

$$\mathrm{P}(\mathrm{FDP}(\hat{S}) > 0) = \mathrm{P}(\hat{S} \cap N \neq \emptyset)$$

- *False Discovery Rate* (FDR)

$$\mathbb{E}(\mathrm{FDP}(\hat{S}))$$

# Error control

- We would like to *control* the chosen error rate at level $\alpha$, i.e.

$$P(\hat{S} \cap N \neq \emptyset) \leq \alpha \quad \text{or} \quad \mathbb{E}(\mathrm{FDP}(\hat{S})) \leq \alpha$$

  while maximizing some notion of power e.g. the average power

$$\mathrm{AvgPower} = \frac{\sum_{j \in S} P(\hat{S} \in j)}{|S|}$$

- We are dealing with the trade-off between type I and type II errors, and since FWER is more stringent than FDR, i.e.

$$\mathbb{E}(\mathrm{FDP}(\hat{S})) \leq P(\hat{S} \cap N \neq \emptyset)$$

  methods that control FWER are less powerful

Simulate data as described in Section 3.1 of Hastie et al. (2020)

Given $n$ (number of observations), $p$ (problem dimensions), $s$ (sparsity level), beta-type (pattern of sparsity), $\rho$ (predictor autocorrelation level), and $\nu$ (signal-to-noise ratio (SNR) level)

1. we define coefficients $\beta \in \mathbb{R}^p$ according to $s$ and the beta-type; e.g. beta-type 2: $\beta$ has its first $s$ components equal to 1, and the rest equal to 0
2. we draw the rows of the predictor matrix $X \in \mathbb{R}^{n \times p}$ i.i.d. from $N_p(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{p \times p}$ has entry $(i, j)$ equal to $\rho^{|i-j|}$ (Toeplitz matrix)
3. we draw the response vector $y \in \mathbb{R}^n$ from $N_n(X\beta, \sigma^2 I_n)$ with $\sigma^2$ defined to meet the desired SNR level, i.e. $\sigma^2 = \beta^t \Sigma \beta / \nu$

# Lasso active set

Lasso with $\lambda$ chosen by e.g. the 1-se rule

$$\hat{S} = \{j \in \{1, \dots, p\} : \hat{\beta}_j \neq 0\}$$

Simulated data with $n = 200$, $p = 1000$, $s = 10$, $\rho = 0$, $\nu = 2.5$:

| Size $\lvert\hat{S}\rvert$ | # Type I $\lvert\hat{S} \cap N\rvert$ | # Type II $\lvert S \setminus \hat{S}\rvert$ | FDP $\lvert\hat{S} \cap N\rvert / \lvert\hat{S}\rvert$ | Sensitivity $\lvert S \setminus \hat{S}\rvert / \lvert S\rvert$ |
|---|---|---|---|---|
| 23 | 13 | 0 | 56.5% | 100% |

100 replications

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Size | 23 | 20 | 13 | 25 | 23 | 21 | 11 |
| # Type I | 13 | 10 | 3 | 15 | 13 | 11 | 4 |
| # Type II | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| FDP | 0.57 | 0.50 | 0.23 | 0.60 | 0.57 | 0.52 | 0.36 |
| Sensitivity | 1 | 1 | 1 | 1 | 1 | 1 | 0.7 |

FWER = 99%, FDR = 54.2%, AvgPower = 99.6%

# Naïve two-step procedure

1. Perform the lasso in order to obtain the active set

$$\hat{M} = \{j \in \{1, \ldots, p\} : \hat{\beta}_j \neq 0\}$$

2. Use least squares to fit the submodel containing just the variables in $\hat{M}$, i.e. linear regression of the $n \times 1$ response $y$ on the reduced $n \times |\hat{M}|$ submatrix $X_{\hat{M}}$. Obtain

$$\hat{S} = \{j \in \hat{M} : p_j \leq \alpha\}$$

where $p_j$ is the $p$-value for testing the null hypothesis $H_j : \beta_j = 0$ in the linear model including only the selected variables

Simulation with $n = 200$, $p = 1000$, $s = 10$, $\rho = 0$, $\nu = 2.5$, $\alpha = 5\%$:

| Size $\lvert\hat{S}\rvert$ | # Type I $\lvert\hat{S} \cap N\rvert$ | # Type II $\lvert S \setminus \hat{S}\rvert$ | FDP $\lvert\hat{S} \cap N\rvert / \lvert\hat{S}\rvert$ | Sensitivity $\lvert S \setminus \hat{S}\rvert / \lvert S\rvert$ |
|---|---|---|---|---|
| 15 | 5 | 0 | 33.3% | 100% |

100 replications

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Size | 15 | 18 | 12 | 17 | 18 | 17 | 11 |
| # Type I | 5 | 8 | 2 | 7 | 8 | 7 | 4 |
| # Type II | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| FDP | 0.33 | 0.44 | 0.17 | 0.41 | 0.44 | 0.41 | 0.36 |
| Sensitivity | 1 | 1 | 1 | 1 | 1 | 1 | 0.7 |

FWER = 99%, FDR = 42.1%, AvgPower = 99.6%

- The main problem with the naïve two-step procedure is that it peeks at the data twice: once to select the variables to include in $\hat{M}$, and then again to test hypotheses associated with those variables
- Here $\hat{M}$ is a random variable (it is a function of the data), but inference for linear model assumes it fixed (given a priori)
- A secondary problem is the multiplicity of the tests performed
- A simple idea is to use data-splitting to break up the dependence of variable selection and hypothesis testing (Cox, 1975)

Data-split

The *single-split* approach (Wasserman and Roeder, 2009) splits the data into two parts $I$ and $L$ of equal sizes $n_I = n_L = n/2$:

1. Use variable selection on the $L$ portion $(X^L, y^L)$ to obtain

$$\hat{M}^L \subseteq \{1, \ldots, p\}$$

2. Use the $I$ portion $(X^I, y^I)$ for constructing $p$-values

$$p_j = \begin{cases} p_j^I & \text{if } j \in \hat{M}^L \\ 1 & \text{if } j \notin \hat{M}^L \end{cases}$$

   where $p_j^I$ is the $p$-value testing $H_j : \beta_j = 0$ in the linear model including only the selected variables, i.e. based on the linear regression of the reduced $n_I \times 1$ response $y^I$ on the reduced $n_I \times |\hat{M}^L|$ matrix $X^I_{\hat{M}^L}$

3. Adjust the $p$-values for their multiplicity $|\hat{M}^L|$, by e.g. Bonferroni

$$\tilde{p}_j = \min(|\hat{M}^L| \cdot p_j, 1), \quad j = 1, \ldots, p$$

4. Selected variables

$$\tilde{S} = \{j \in \hat{M}^L : \tilde{p}_j \leq \alpha\}$$

Theorem

*Assume that*

1. *the linear model* $y \sim N_n(1\beta_0 + X\beta, \sigma^2 I)$ *holds*
2. *the variable selection procedure satisfies the screening property for the first half of the sample, i.e.*

   $$\mathrm{P}(\hat{M}^L \supseteq S) \geq 1 - \delta$$

   *for some* $\delta \in (0, 1)$.
3. *The reduced design matrix for the second half of the sample satisfies* $\mathrm{rank}(X^I_{\hat{M}^L}) = |\hat{M}^L|$.

*Then the single-split procedure yields FWER control at* $\alpha$ *against inclusion of null predictors up to the additional (small) value* $\delta$, *i.e.*

$$\mathrm{P}(\tilde{S} \cap N \neq \emptyset) \leq \alpha + \delta$$

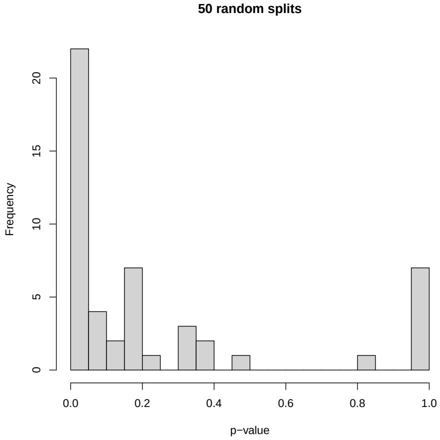
Proof.
Let $E = \{\hat{M}^L \supseteq S\}$ with $\mathrm{P}(E^c) \leq \delta$ by assumption. If $E$ happens, then $p_j^I$ is a valid $p$-value, i.e. $\mathrm{P}(p_j^I \leq u|E) \leq u$ for $j \in N \cap \hat{M}^L$. We have

$$\begin{aligned}
& \mathrm{P}(\tilde{S} \cap N \neq \emptyset) = \mathrm{P}(\bigcup_{j \in \hat{M}^L \cap N} \{\tilde{p}_j \leq \alpha\}) \\
= \quad & \mathrm{P}(\bigcup_{j \in \hat{M}^L \cap N} \{\tilde{p}_j \leq \alpha\}|E)\mathrm{P}(E) + \mathrm{P}(\bigcup_{j \in \hat{M}^L \cap N} \{\tilde{p}_j \leq \alpha\}|E^c)\mathrm{P}(E^c) \\
\leq \quad & \big[\sum_{j \in \hat{M}^L \cap N} \mathrm{P}(p_j^I \leq \frac{\alpha}{|\hat{M}^L|}|E)\big]\mathrm{P}(E) + \mathrm{P}(\bigcup_{j \in \hat{M}^L \cap N} \mathbb{1}\{\tilde{p}_j \leq \alpha\}|E^c)\mathrm{P}(E^c) \\
\leq \quad & |\hat{M}^L \cap N| \frac{\alpha}{|\hat{M}^L|} \cdot 1 + 1 \cdot \delta \\
\leq \quad & \alpha + \delta
\end{aligned}$$

$\square$

# P-value lottery

A major problem of the single data-splitting method is that different data splits lead to different *p*-values



# Multi-split

The *multi-split* approach (Meinshausen et al., 2009)

1. For $b = 1, \dots, B$
   apply the single-split procedure $(L^b, I^b)$ to obtain

$$\{\tilde{p}_j^b, j = 1, \dots, p\}$$

2. Aggregate the $p$-values as

$$\bar{p}_j = 2 \cdot \text{median}(\tilde{p}_j^1, \dots, \tilde{p}_j^B), \quad j = 1, \dots, p$$

3. Selected predictors:

$$\bar{S} = \{j \in \{1, \dots, p\} : \bar{p}_j \leq \alpha\}$$

# Simultaneous confidence intervals

$$P(\beta_j \in [\hat{L}_j, \hat{U}_j] \; \forall j \in \{1, \ldots, p\}) \geq 1 - \alpha$$

| $j$ | lower | upper |
|---:|---:|---:|
| 1 | 0.20 | 1.76 |
| 2 | 0.65 | 1.88 |
| 3 | 0.54 | 1.79 |
| 4 | 0.37 | 1.61 |
| 5 | 0.39 | 1.64 |
| 6 | 0.62 | 1.74 |
| 7 | 0.25 | 1.49 |
| 8 | 0.34 | 1.68 |
| 9 | 0.40 | 1.58 |
| 10 | 0.41 | 1.54 |
| 11 | -$\infty$ | $\infty$ |
| ... | | |