Statistical Learning - module III

Aldo Solari aldo.solari@unimib.it

Course webpage: https://aldosolari.github.io/SL3/

Contents

Chapter 1

Sample splitting for high dimensional variable selection

> the simple idea of splitting a sample in two and then developing the hypothesis on the basis of one part and testing it on the remainder may perhaps be said to be one of the most seriously neglected ideas in statistics (Barnard 1974, p. 133).

Consider the gaussian linear model

$$y = X\beta + \varepsilon \tag{1.1}$$

where $\varepsilon \sim N(0, \sigma^2 I)$ and $X = [X_1, \ldots, X_p] \in \mathbb{R}^{n \times p}$ is the design matrix. The *support set* (or active set) is

$$S = \{j : \beta_j \neq 0\}$$

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

$$N = \{j : \beta_j = 0\}$$

The main goal is the construction of valid p-values and confidence intervals for individual regression parameters β_j. We allow for potentially high-dimensional designs, that is, $p \gg n$. This makes statistical inference very challenging.

1.1 Single split

The single-split approach proposed by Wasserman and Roeder (2009) splits the data into two parts L and I of equal sizes $n/2$:

- based the first half of the sample L, reduce the dimensionality of predictors to a manageable number of predictors, thereby reducing the effect of multiplicity;
- based on the second half of the sample I, compute p-values using classical least squares estimation and make a final selection.

Let $\hat{S}$ be a variable selection or screening procedure that estimates the set of active predictors S. Abusing notation slightly, we also denote by $\hat{S}$ the index set of selected predictors.
Use $\hat{S}$ on the L portion of the data (X^L, y^L), obtaining

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

and use the I portion (X^I, y^I) for constructing p-values

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

where p_j^I is the p-value testing the null hypotheses $H_j : \beta_j = 0$ in the linear model including only the selected variables, i.e. we regress the reduced $n_I \times 1$ response y^I on the reduced $n_I \times |\hat{S}^L|$ matrix $X^I_{\hat{S}^L}$.
If we wish to control the familywise error rate, adjust the p-values for their multiplicity $|\hat{S}^L|$, by e.g. Bonferroni

$$\tilde{p}_j^I = |\hat{S}^L| \cdot p_j^I \wedge 1$$

The single-split procedure is summarized by Algorithm 1.1.

Algorithm 1 Single-split procedure

Require: y, X, $\alpha \in (0, 1)$ and a variable selection procedure $\hat{S}$

1: Partition $\{1, \ldots, n\}$ into portions L and I of size $n_L = \lfloor n/2 \rfloor$ and $n_I = n - n_L$
2: Using L only, select the predictors $\hat{S}^L \subseteq \{1, \ldots, p\}$
3: Based on linear regression of y^I on $X^I_{\hat{S}^L}$, compute the p-value p_j^I testing $H_j : \beta_j = 0$ for $j \in \hat{S}^L$.
4: Bonferroni-adjustement: $\tilde{p}_j = \min(|\hat{S}^L| \cdot p_j^I, 1)$ for $j \in \hat{S}^L$. Assign $\tilde{p}_j = 1$ for $j \notin \hat{S}^L$.
5: Selected predictors: $\tilde{S} = \{j \in \hat{S}^L : \tilde{p}_j \leq \alpha\}$

Theorem 1.1.1. *Assume that*

1. *the linear model* $y \sim N(X\beta, \sigma^2 I)$ *is the true model;*

2. *the variable selection procedure* $\hat{S}$ *satisfies the screening property for the first half of the sample, i.e.*

$$\mathbb{P}(\hat{S}^L \supseteq S_0) \geq 1 - \delta$$

for some $\delta \in (0, 1)$.

3. *The reduced design matrix for the second half of the sample satisfies* $\text{rank}(X^I_{\hat{S}^L}) = |\hat{S}^L|$.

Then the single-split procedure yields familywise error control against inclusion of null predictors up to the additional (small) value δ*, i.e.*

$$\mathbb{P}(\tilde{S} \cap N_0 \neq \emptyset) \leq \alpha + \delta \tag{1.2}$$

Proof. Let $E = \{\hat{S}^L \supseteq S_0\}$ with $\mathbb{P}(E^c) \leq \delta$ by assumption. Suppose E happens. Then p^I_j is a valid p-value, i.e. $\mathbb{P}(p^I_j \leq u) \leq u$ for $j \in N_0 \cap \hat{S}^L$. Then

$$\begin{aligned}
\mathbb{P}(\bigcup_{j \in N_0 \cap \hat{S}^L} \mathbb{1}\{\tilde{p}_j \leq \alpha\}) &= \mathbb{P}(\bigcup_{j \in N_0 \cap \hat{S}^L} \mathbb{1}\{|\hat{S}^L| p^I_j \leq \alpha\} | E)\mathbb{P}(E) \\
&\quad + \mathbb{P}(\bigcup_{j \in N_0 \cap \hat{S}^L} \mathbb{1}\{\tilde{p}_j \leq \alpha\} | E^c)\mathbb{P}(E^c) \\
&\leq \alpha + \delta
\end{aligned}$$

□

Multi sample-splitting

Multiple sample-splitting has been proposed by Meinshausen, Meier, and Bühlmann (2009) and Dezeure, Bühlmann, Meier, and Meinshausen (2015).

Theorem 1.1.2. *Assume that*

1. *the linear model* $y \sim N(X\beta, \sigma^2 I)$ *is the true model;*

2. *the variable selection procedure* $\hat{S}$ *satisfies the screening property for the first half of the sample, i.e.*

$$\mathbb{P}(\hat{S}^{L^b} \supseteq S_0) \geq 1 - \delta$$

for some $\delta \in (0, 1)$.

Algorithm 2 Multi sample-splitting procedure

Require: y, X, $\alpha \in (0,1)$, $B \in \mathbb{N}$ and a variable selection procedure $\hat{S}$

1: **for** $b = 1, \ldots, B$ **do**
2: Apply the single-split procedure (L^b, I^b) to obtain $\tilde{p}_j^b$, $j = 1, \ldots, p$
3: **end for**
4: Aggregate the p-values as

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

5: Selected predictors: $\bar{S} = \{j \in \{1, \ldots, p\} : \bar{p}_j \leq \alpha\}$

3. *The reduced design matrix for the second half of the sample satisfies* $\text{rank}(X_{\hat{S}^L}^{I^b}) = |\hat{S}^{L^b}|$.

Then the single-split procedure yields familywise error control against inclusion of null predictors up to the additional (small) value $B\delta$, i.e.

$$\mathbb{P}(\bar{S} \cap N_0 \neq \emptyset) \leq \alpha + B\delta \tag{1.3}$$

Proof. See Dezeure et al. (2015). □

Confidence intervals can be constructed based on the duality with the p-values.

Chapter 2

Stability Selection

A problem of many variable selection procedures is that null predictors might be erroneously selected (and non-null predictors might be erroneously deselected). Let

$$V = |\hat{S} \cap N_0|$$

be the number of wrongly selected predictors. To improve the selection process and to obtain an error control for V, Meinshausen and Bühlmann (2010) proposed stability selection, which was later enhanced by Shah and Samworth (2013).

Stability selection is not a new variable selection technique. Its aim is rather to enhance and improve existing methods. It is a versatile approach, which can be combined with high-dimensional variable selection procedures.

A particularly attractive feature of Stability Selection is the control of the *per-family error rate* provided by an upper bound on the expected number of selected null predictors $\mathbb{E}(V)$.

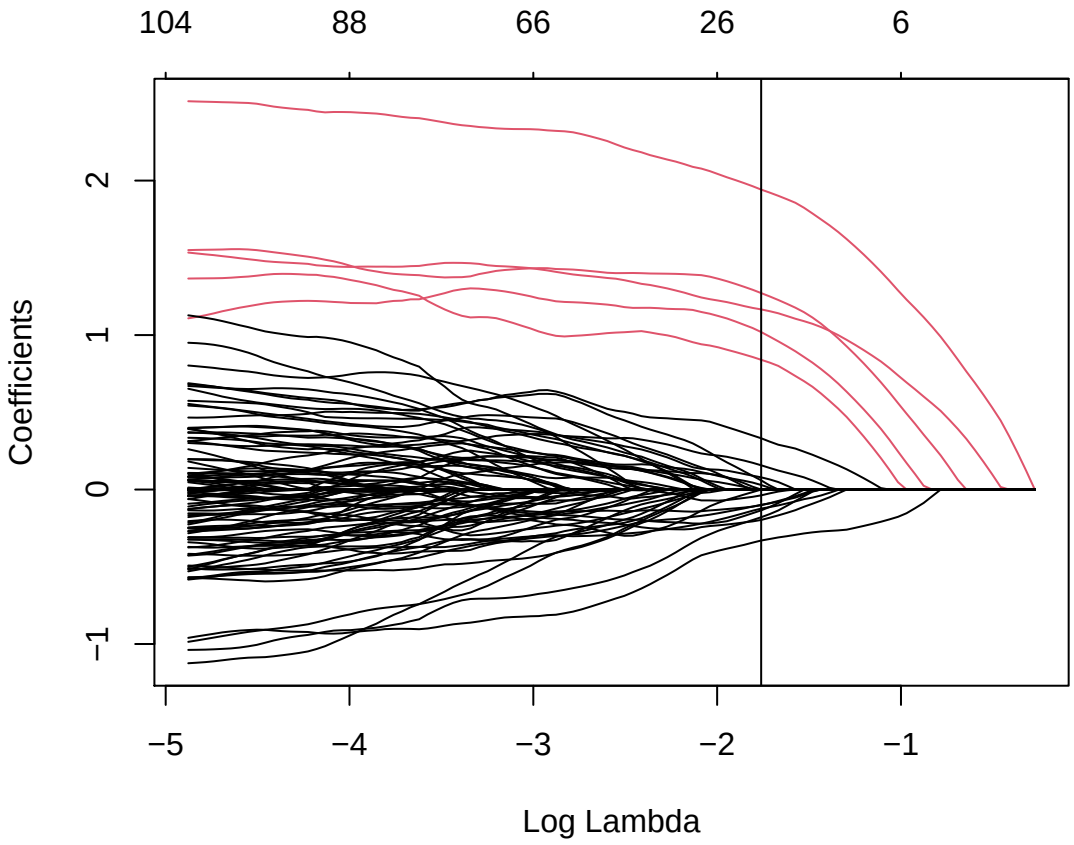
2.1 Stability path

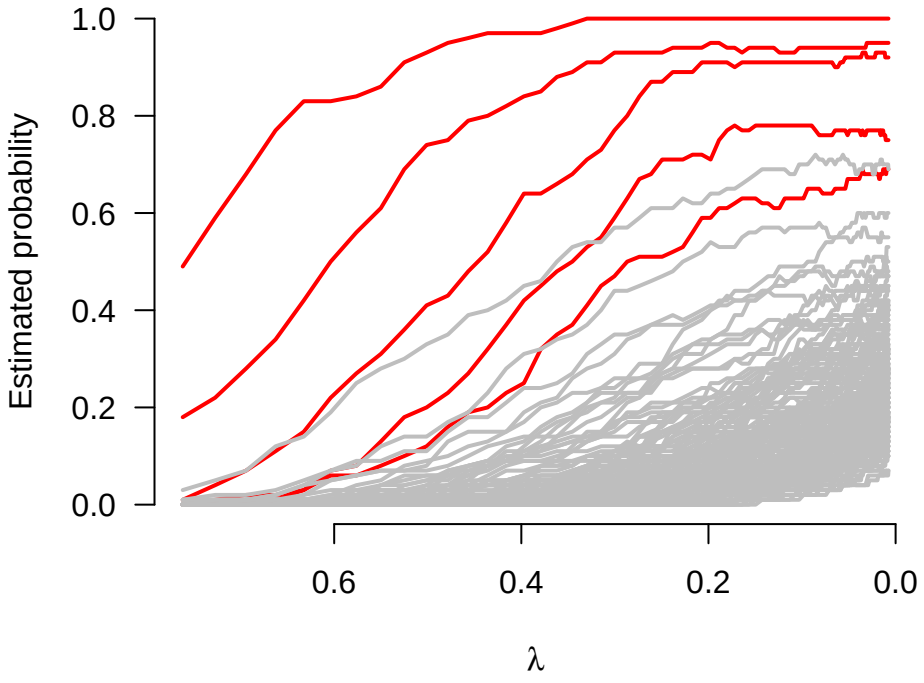
A *regularisation path* is given by the coefficient value of each variable over all regularisation parameters

$$\{\hat{\beta}_j(\lambda), \lambda \in \Lambda, j = 1, \ldots, p\}$$

See Figure 2.1.

Figure 2.1: Regularization path (upper figure) and stability path (lower figure), where red lines correspond to predictors with non-zero coefficients.





A *stability path* is, in contrast, the estimated probability for each predictor to be selected when randomly resampling from the data. See Figure 2.1.

The construction of the stability path by using the Lasso is described by Algorithm 2.1, where for simplicity we suppose that n is an even number.

Algorithm 3 Stability Path Algorithm with the Lasso

Require: $B \in \mathbb{N}$, $\Lambda \subset [0, \infty)$, $\tau \in [0, 1]$

1: **for** $b = 1, \ldots, B$ **do**

2: Randomly select $n/2$ indices from $\{1, \ldots, n\}$ without replacement;

3: Run the lasso algorithm on the $n/2$ observations obtained in the previous step to select a subset of predictors:

$$\hat{S}_{n/2}(\lambda) = \{j : \hat{\beta}_j(\lambda) \neq 0\}$$

for each $\lambda \in \Lambda$;

4: **end for**

5: Compute the relative selection frequencies:

$$\hat{\pi}_j(\lambda) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\{j \in \hat{S}_{n/2}(\lambda)\}$$

6: The set of *stable predictors* is given by

$$\hat{S}_{\text{stab}} = \{j : \max_{\lambda \in \Lambda} \hat{\pi}_j(\lambda) \geq \tau\}$$

Meinshausen and Bühlmann (2010) show that this selection procedure controls the PFER. Let $\hat{S}_{n/2}(\Lambda) = \bigcup_{\lambda \in \Lambda} \hat{S}_{n/2}(\lambda)$ is the set of selected predictors if varying the regularization λ in Λ.

Theorem 2.1.1. *If we assume that*

1. $\{\mathbb{1}\{j \in \hat{S}_{n/2}(\lambda)\}, j \in N_0\}$ *is exchangeable for all* $\lambda \in \Lambda$*;*
2. *The variable selection procedure is not worse than random guessing, i.e.*

$$\frac{\mathbb{E}(|\hat{S}_{n/2}(\Lambda) \cap S_0|)}{\mathbb{E}(|\hat{S}_{n/2}(\Lambda) \cap N_0|)} \geq \frac{|S_0|}{|N_0|}$$

then for $\tau \in (1/2, 1]$

$$\mathbb{E}(|\hat{S}_{\text{stab}} \cap N_0|) \leq \frac{1}{2\tau - 1} \frac{q^2}{p} \tag{2.1}$$

where $q = \mathbb{E}(|\hat{S}_{n/2}(\Lambda)|)$.

Proof. See Theorem 1 in Meinshausen and Bühlmann (2010). □

2.2 Complementary Pairs Stability Selection

A modification of stability selection was introduced by Shah and Samworth (2013). One major difference to the original stability selection approach is that instead of using B independent subsamples of the data, Shah and Samworth use $2B$ complementary pairs: one draws B subsamples of size $n/2$ from the data and uses, for each subsample, the remaining observations as a second complementary subsample.

More importantly, the error bound is theoretically derived that holds without assuming exchangeability of the null predictors and without assuming that the selection procedure is not worse than random guessing. The drawback of being able to drop these assumptions is that the modified bounds do not control the per-family error rate, but the expected number of selected variables with low selection probability.

Complementary Pairs Stability Selection (CPSS) is described by Algorithm 2.2.

Algorithm 4 (Complementary Pairs) Stability Selection

Require: A variable selection procedure $\hat{S}_n$, $B \in \mathbb{N}$, $\tau \in [0, 1]$

1: **for** $b = 1, \ldots, B$ **do**

2: Partition $\{1, \ldots, n\}$ into (I^{2b-1}, I^{2b}) of size $n/2$, and for each get

$$\hat{S}_{n/2} \subseteq \{1, \ldots, p\}$$

3: **end for**

4: Compute the relative selection frequencies:

$$\hat{\pi}_j = \frac{1}{2B} \sum_{r=1}^{2B} \mathbb{1}\{j \in \hat{S}^r_{n/2}\}$$

5: The set of *stable predictors* is given by

$$\hat{S}_{\text{stab}} = \{j : \hat{\pi}_j \geq \tau\}$$

Note that the relative selection frequency $\hat{\pi}_j$ is an unbiased estimator of $\pi_j^{n/2} = \mathbb{P}(j \in \hat{S}_{n/2})$ but, in general, a biased estimator of

$$\pi_j^n = \mathbb{P}(j \in \hat{S}_n) = \mathbb{E}(\mathbb{1}\{j \in \hat{S}_n\})$$

The key idea of stability selection is to improve on the simple estimator $\mathbb{1}\{j \in \hat{S}_n\}$ of π_j^n through subsampling. By means of averaging involved in $\hat{S}_{\text{stab}}$, we hope that $\hat{\pi}_j$ will have reduced variance compared to $\mathbb{1}\{j \in \hat{S}_n\}$ and this increased stability will more than compensate for the bias incurred.

For $\theta \in [0, 1]$, let

$$L_\theta = \{j : \pi_j^{n/2} \leq \theta\}$$

denote the set of variables that have low selection probability under $\hat{S}_{n/2}$.

Theorem 2.2.1. *If $\tau \in (1/2, 1]$, then*

$$\mathbb{E}(|\hat{S}_{\text{stab}} \cap L_\theta|) \leq \frac{\theta}{2\tau - 1}\mathbb{E}(|\hat{S}_{n/2} \cap L_\theta|) \tag{2.2}$$

If we assume that

1. *$\{\mathbb{1}\{j \in \hat{S}_{n/2}\}, j \in N_0\}$ is exchangeable;*
2. *The variable selection procedure is not worse than random guessing, i.e.*

$$\frac{\mathbb{E}(|\hat{S}_{n/2} \cap S_0|)}{\mathbb{E}(|\hat{S}_{n/2} \cap N_0|)} \geq \frac{|S_0|}{|N_0|}$$

then for $\tau \in (1/2, 1]$

$$\mathbb{E}(|\hat{S}_{\text{stab}} \cap N_0|) \leq \frac{1}{2\tau - 1}\frac{q^2}{p} \tag{2.3}$$

where $q = \mathbb{E}(|\hat{S}_{n/2}|)$.

Proof. See Theorem 1 in Shah and Samworth (2013). □

Remarks on the assumptions

If the exchangeability assumption holds and the selection procedure is not worse than random guessing, then all null predictors have a "below average" selection probability, i.e.

$$N_0 = L_{q/p}$$

.

Assumption 1. essentially means that each null predictor has the same selection probability. This assumption is very strong.

Assumption 2. essentially means that signal variables should be selected with higher probability than noise variables. This assumption is usually not very restrictive as we would expect it to hold for the lasso.

With additional assumptions we get much tighter error bounds: see Shah and Samworth (2013).

Choice of parameters

Some practical remarks:

- The choice of the number of subsamples B is of minor importance.
- One can regard the choice of $q = \mathbb{E}(|\hat{S}_{n/2}|)$ as part of the choice of the variable selection procedure. For the Lasso, one option is to fix q by varying λ at each evaluation of the selection procedure until it selects q variables. However, if cross-validation is used to choose λ at each iteration, then q can be estimated by $\sum_{j=1}^{p} \hat{\pi}_j$.
- If one fixes q, it should be chosen so high that in theory all signal variables S_0 can be chosen. If q is too small, one would inevitably select only a subset of the signal variables as

$$|\hat{S}_{\text{stab}}| \leq |\hat{S}_{n/2}| = q$$

- For a fixed value q, we can easily vary the desired error bound by varying the threshold τ accordingly:

$$\overline{\mathbb{E}(V)} = \frac{q^2}{(2\tau - 1)p}$$

An alternative is to turn things around and decide on a value of τ and $\overline{\mathbb{E}(V)}$, and then use

$$q = \sqrt{\overline{\mathbb{E}(V)}(2\tau - 1)}.$$

Chapter 3

The knockoff filter

There are two main approaches:

- *Fixed-X knockoffs*
  The original paper by Barber, Candès, et al. (2015) generates knockoffs without any assumptions on X; however this approach works if X is full rank with $n \geq 2p$.

- *Model-X knockoffs*
  Candes, Fan, Janson, and Lv (2018) extended this idea to the $p > n$ case, although we need to make assumptions about X.

For more variants, see `https://web.stanford.edu/group/candes/knockoffs/index.html`

3.1 Fixed-X knockoffs

Consider the gaussian linear model

$$y = X\beta + \varepsilon \tag{3.1}$$

where $\varepsilon \sim N(0, \sigma^2 I)$ and $X = [X_1, \ldots, X_p] \in \mathbb{R}^{n \times p}$ is a fixed design matrix of full rank with $n \geq 2p$. Let $N = \{j : \beta_j = 0\}$ be the index set of the null features.

The knockoff features

The basic idea of the knockoff filter is that for each feature X_j we construct a *knockoff copy* $\tilde{X}_j$.

Let $X^\mathsf{T} X = \Sigma$. Suppose without loss of generality that the features are normalized such that $\Sigma_{jj} = \|X_j\|_2^2 = X_j^\mathsf{T} X_j = 1$ for all j.

The property that a knockoff copy $\tilde{X}_j$ must satisfy is

$$\begin{aligned}[X\ \tilde{X}]^\mathsf{T}[X\ \tilde{X}] &= \begin{bmatrix} X^\mathsf{T}X & X^\mathsf{T}\tilde{X} \\ \tilde{X}^\mathsf{T}X & \tilde{X}^\mathsf{T}\tilde{X} \end{bmatrix} \\ &= \begin{bmatrix} \Sigma & \Sigma - \mathrm{diag}\{s\} \\ \Sigma - \mathrm{diag}\{s\} & \Sigma \end{bmatrix} = G \end{aligned} \tag{3.2}$$

where $s = (s_1, \ldots, s_p)$ is a p-dimensional nonnegative vector.

In words, $\tilde{X}$ exhibits the same covariance structure as the original design X, i.e.

$$X^\mathsf{T}X = \tilde{X}^\mathsf{T}\tilde{X} = \Sigma$$

but in addition, the correlations between distinct original and knockoff variables are the same as those between the originals (because Σ and $\Sigma - \mathrm{diag}(s)$ are equal on off-diagonal entries), i.e.

$$\tilde{X}_j^\mathsf{T}X_k = X_j^\mathsf{T}X_k \quad \forall\, k \neq j$$

However, comparing a feature X_j to its knockoff $\tilde{X}_j$, we see that

$$\tilde{X}_j^\mathsf{T}X_j = 1 - s_j$$

while

$$X_j^\mathsf{T}X_j = \tilde{X}_j^\mathsf{T}\tilde{X}_j = 1.$$

We should choose the entries of s as large as possible so that a variable X_j is not too similar to its knockoff X.

A strategy for constructing $\tilde{X}$ is to choose s satisfying

$$\mathrm{diag}(s) \preceq 2\Sigma$$

and construct $\tilde{X}$ as

$$\tilde{X} = X(I - \Sigma^{-1}\mathrm{diag}\{s\}) + \tilde{U}C \tag{3.3}$$

where $\tilde{U} \in \mathbb{R}^{n\times p}$ is an orthonormal matrix whose column space is orthogonal to X so that $\tilde{U}^\mathsf{T}X = 0$ and $C^\mathsf{T}C = 2\mathrm{diag}\{s\} - \mathrm{diag}\{s\}\Sigma^{-1}\mathrm{diag}\{s\}$ is a Cholesky decomposition.

In this construction, we would like to have X_j and $\tilde{X}_j$ to be as orthogonal as possibile, i.e. we would like to have $\tilde{X}_j^\mathsf{T}X_j = 1 - s_j$ as close to zero as possible (the s_j's are near 1).

For any subset $J \subset \{1, \ldots, p\}$, let $[X\ \tilde{X}]_{\mathrm{swap}(J)}$ be the matrix obtained from $[X\ \tilde{X}]$ by swapping the columns X_j and $\tilde{X}_j$ for each $j \in J$.

Lemma 3.1.1. *For any subset $J \subset \{1, \ldots, p\}$, we have*

$$[X\ \tilde{X}]^{\mathsf{T}}_{\mathrm{swap}(J)}[X\ \tilde{X}]_{\mathrm{swap}(J)} = [X\ \tilde{X}]^{\mathsf{T}}[X\ \tilde{X}] = G$$

Proof. See Lemma 2 in Barber et al. (2015). □

Lemma 3.1.2. *For any $J \subseteq N$, we have*

$$[X\ \tilde{X}]^{\mathsf{T}}_{\mathrm{swap}(J)}y \stackrel{d}{=} [X\ \tilde{X}]^{\mathsf{T}}y \tag{3.4}$$

Proof. See Lemma 3 in Barber et al. (2015). □

The knockoff statistics

With the knockoffs constructed, the next step is to fit (e.g.) Lasso to the *augmented* $n \times 2p$ design matrix $[X\ \tilde{X}]$.

Let $[\hat{\beta}\ \tilde{\beta}] \in \mathbb{R}^{2p}$ denote the resulting coefficient estimates. For each feature X_j and each knockoff $\tilde{X}_j$, record the first time that this feature or its knockoff enters the Lasso path, that is, the largest penalty parameter value λ such that $\hat{\beta}_j(\lambda) \neq 0$ or $\tilde{\beta}_j(\lambda) \neq 0$.

Let $Z_j = \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}$ and $\tilde{Z}_j = \sup\{\lambda : \tilde{\beta}_j(\lambda) \neq 0\}$, then set

$$W_j = \max(Z_j, \tilde{Z}_j) \times \begin{cases} +1 & \text{if } X_j \text{ enters first } (Z_j > \tilde{Z}_j) \\ -1 & \text{if } \tilde{X}_j \text{ enters first } (Z_j < \tilde{Z}_j) \end{cases}$$

and $W_j = 0$ in case $Z_j = \tilde{Z}_j$. A large positive W_j indicates that X_j was strongly preferred over $\tilde{X}_j$, and bears evidence against the null hypothesis $H_j : \beta_j = 0$.

Note that $W = (W_1, \ldots, W_p)^{\mathsf{T}}$ obeys the *antisymmetry property*

$$W_j([X\ \tilde{X}]_{\mathrm{swap}(J)}, y) = W_j([X\ \tilde{X}], y) \times \begin{cases} +1 & j \notin J \\ -1 & j \in J \end{cases}$$

and the *sufficiency property*

$$W = f([X\ \tilde{X}]^{\mathsf{T}}[X\ \tilde{X}], [X\ \tilde{X}]^{\mathsf{T}}y)$$

because the Lasso solution is equivalent to

$$\text{minimize} \quad \frac{1}{2}b^{\mathsf{T}}X^{\mathsf{T}}Xb - b^{\mathsf{T}}X^{\mathsf{T}}y + \lambda\|b\|_1$$

and thus depends upon the data (y, X) through $X^{\mathsf{T}}X$ and $X^{\mathsf{T}}y$ only.

Lemma 3.1.3. *Let $\pi \in \{\pm 1\}^p$ be a sign sequence independent of $W = (W_1, \ldots, W_p)^\mathsf{T}$, with $\pi_j = +1$ for all $j \in N^c$ and $\pi_j \overset{IID}{\sim} \pm 1$ for $j \in N$.*

Then

$$(W_1, \ldots, W_p)^\mathsf{T} \overset{d}{=} (W_1 \cdot \pi_1, \ldots, W_p \cdot \pi_p)^\mathsf{T}$$

Proof. For any $J \subset \{1, \ldots, p\}$, let $W_{\text{swap}(J)}$ be the statistic we would get if we had replaced $[X\ \tilde{X}]$ with $[X\ \tilde{X}]_{\text{swap}(J)}$. Then

$$W_{\text{swap}(J)} = (W_1 \cdot \pi_1, \ldots, W_p \cdot \pi_p)^\mathsf{T}$$

by the antisymmetry property where $\pi_j = 1$ for $j \notin J$ and $\pi_j = -1$ for $j \in J$.

Now let $\pi_j = +1$ for all $j \in N^c$ and $\pi_j \overset{IID}{\sim} \pm 1$ for $j \in N$. Take $J = \{j : \pi_j = -1\}$. Since $J \subseteq N$, Lemma 1 and 2 give

$$\begin{aligned} W_{\text{swap}(J)} &= f([X\ \tilde{X}]^\mathsf{T}_{\text{swap}(J)}[X\ \tilde{X}]_{\text{swap}(J)}, [X\ \tilde{X}]^\mathsf{T}_{\text{swap}(J)}y) \\ &\overset{d}{=} f([X\ \tilde{X}]^\mathsf{T}[X\ \tilde{X}], [X\ \tilde{X}]^\mathsf{T}y) = W \end{aligned} \tag{3.5}$$

□

Lemma 3 implies that

$$\#\{j \in N : W_j \leq -t\} \overset{d}{=} \#\{j \in N : W_j \geq t\} \tag{3.6}$$

Suppose we select those features for which $W_j \geq t$. Then by using (3.6) we can estimate the false discovery proportion at the threshold t as

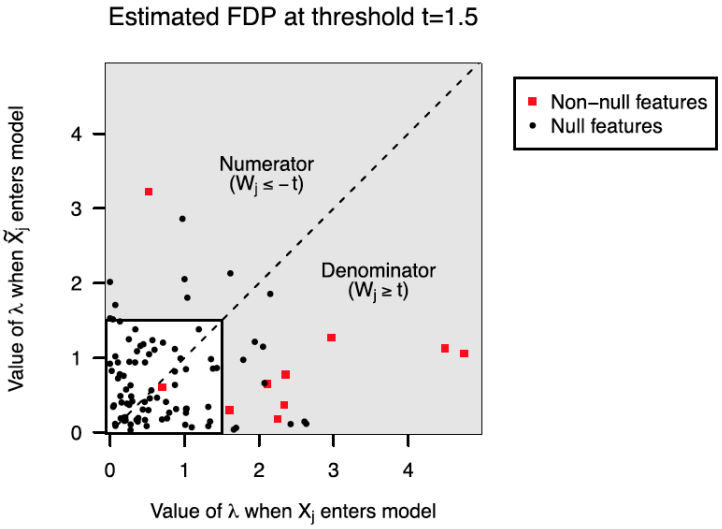
$$\begin{aligned} \text{FDP}(t) &= \frac{\#\{j \in N : W_j \geq t\}}{1 \vee \#\{j : W_j \geq t\}} \approx \frac{\#\{j \in N : W_j \leq -t\}}{1 \vee \#\{j : W_j \geq t\}} && (3.7) \\ &\leq \frac{1 + \#\{j : W_j \leq -t\}}{1 \vee \#\{j : W_j \geq t\}} = \widehat{\text{FDP}}(t) && (3.8) \end{aligned}$$

If we select those features for which $W_j \geq t$, we have a knockoff estimate of the false discovery proportion

$$\widehat{\text{FDP}}(t) = \frac{1 + \#\{j : W_j \leq -t\}}{1 \vee \#\{j : W_j \geq t\}}$$

See Figure 3.1.

Figure 3.1: Figure 1 of Barber et al. (2015). Representation of the knockoff procedure. Black dots correspond to $j \in N$ while red squares are $j \in N^c$. Setting $t = 1.5$, the number of points in the shaded region below the diagonal is equal to $\#j : W_j \geq t$, the number of selected variables at this threshold, while the number of points in the shaded region above the diagonal is equal to $\#j : W_j \leq t$. Observe that the true signals (red squares) are primarily below the diagonal, indicating $W_j > 0$, while the null features (black dots) are roughly symmetrically distributed across the diagonal.



The knockoff+ procedure chooses a data-dependent threshold $T > 0$

$$T = \min\left\{t > 0 : \widehat{\mathrm{FDP}}(t) \leq \alpha\right\}$$

with the convention that $T = +\infty$ if no such t exists.

Theorem 3.1.4. *For any $\alpha \in (0, 1)$, the knockoff+ procedure selects*

$$\hat{S} = \{j : W_j \geq T\}$$

with the guarantee that

$$\mathrm{FDR}(\hat{S}) = \mathbb{E}\left(\frac{|N \cap \hat{S}|}{1 \vee |\hat{S}|}\right) \leq \alpha$$

where the expectation is taken over the Gaussian noise ε in model (3.1) while treating X and $\tilde{X}$ as fixed.

Proof. See Theorem 2 in Barber et al. (2015). □

3.2 Model-X knockoffs

Extending the idea to $p > n$ situations requires to treat $X = (X_1, \dots, X_p)^\mathsf{T}$ as random and to model its distribution.
The main idea is to *shift the burden of knowledge*:

- In the classical setup, $Y|X$ is modelled and X is not.
- In the Model-X setup we do exactly the opposite: assume that we know everything about the distribution of X but require no assumptions on the conditional distribution of $Y|X$.

We have a random pair (X, Y) and ask the following question: the conditional distribution $Y|X$ depends on X through which variables?

Mathematically speaking, we are looking for the *Markov blanket* S, i.e. the 'smallest' subset S such that, conditionally on $\{X_j, j \in S\}$, Y is independent of all other variables, i.e.

$$Y \perp\!\!\!\perp \{X_j, j \notin S\} | \{X_j, j \in S\}$$

For almost all joint distributions of (X, Y), there is a unique Markov blanket but there are pathological cases where it does not.

Under weak regularity conditions, the Markov blanket coincides with the (unique) set of relevant variables defined by *pairwise conditional independence*:

$$N = \{j \in \{1, \dots, p\} : Y \perp\!\!\!\perp X_j | X_{-j}\}$$

where $X_{-j} = \{X_1, \dots, X_p\} \setminus \{X_j\}$.

Note that if the likelihood of Y follows a GLM, then $Y \perp\!\!\!\perp X_j | X_{-j}$ if and only if $\beta_j = 0$, thus

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

A variable X_j is said to be *null* if and only if $j \in N$ and *non-null* or relevant if $j \notin N$. The goal is to discover as many relevant variables as possible while keeping the FDR under control, i.e.

$$\mathrm{FDR}(\hat{S}) = \mathbb{E}\left(\frac{|\hat{S} \cap N|}{|\hat{S}|}\right)$$

Suppose that we have n IID samples from a population, each of the form (X, Y) , where $X = (X_1, \dots, X_p)^\mathsf{T} \in \mathbb{R}^p$ and $Y \in \mathbb{R}$.

Model-X knockoffs

A Model-X knockoff copy $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)^\mathsf{T}$ must satisfy two properties :

1. *Pairwise exchangeability*: for any $J \subseteq \{1, \ldots, p\}$, the distribution of $[X\ \tilde{X}]$ is invariant to any change of original and knockoff features

$$[X\ \tilde{X}]_{\text{swap}(J)} \stackrel{d}{=} [X\ \tilde{X}]$$

2. *Conditional independence*: $\tilde{X} \perp\!\!\!\perp Y|X$

To give an example of Model-X knockoffs, suppose that $X \sim N(0, \Sigma)$. Then a joint distribution obeying Property 1 is

$$[X\ \tilde{X}] \sim N(0, G) \quad \text{with } G = \begin{bmatrix} \Sigma & \Sigma - D \\ \Sigma - D & \Sigma \end{bmatrix}$$

where D is a diagonal matrix with entries $\{D\}_{jj} = d_j$ chosen such that the covariance matrix of $[X\ \tilde{X}]$ is positive definite.

For constructing knockoff variables having observed X, a possibility is to sample the knockoff vector $\tilde{X}$ from the conditional distribution

$$\tilde{X}|X \stackrel{d}{=} N(\mu, V)$$

with $\mu = \mathbb{E}(\tilde{X}|X) = X - X\Sigma^{-1}D$ and $V = \mathbb{V}\text{ar}(\tilde{X}|X) = 2D - D\Sigma^{-1}D$.

In many real applications, the true Σ may not be known exactly, forcing the user to estimate it from the available data. In this case we generate knockoffs for Gaussian variables but, instead of using the true covariance matrix Σ, we use its estimate $\hat{\Sigma}$.

References

Barber, R. F., Candès, E. J., et al. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, *43*(5), 2055–2085.

Candes, E., Fan, Y., Janson, L., & Lv, J. (2018). Panning for gold:?model-x?knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *80*(3), 551–577.

Dezeure, R., Bühlmann, P., Meier, L., & Meinshausen, N. (2015). High-dimensional inference: Confidence intervals, p-values and r-software hdi. *Statistical science*, 533–558.

Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(4), 417–473.

Meinshausen, N., Meier, L., & Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, *104*(488), 1671–1681.

Shah, R. D., & Samworth, R. J. (2013). Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *75*(1), 55–80.

Wasserman, L., & Roeder, K. (2009). High dimensional variable selection. *Annals of statistics*, *37*(5A), 2178.