Multiple testing problems in linear models: An introduction with the Knockoff filter

August 23, 2021

Industries have now well understood the interest to use statistical and machine learning tools to develop and improve their activity. Data being the fuel of such techniques, people put a lot of effort in gathering as much as information as possible on their business. In this context, engeniors often have to deal with high-dimensional datasets where a lot of features are not relevant to address the question at stake. One big challenge for them is to identify the features that are essential to explain the studied phenomena. To cope with this issue, *variable selection* has emerged as an active field of research.
In this homework, we shed light on a recent method to control the False Discovery Rate (FDR) for variable selection in linear models.

1 Preliminaries

Let us consider a design matrix (also called matrix of features) $X \in \mathbb{R}^{n \times d}$. We consider the gaussian model where we observe a vector $y \in \mathbb{R}^n$ given by

$$y \sim \mathcal{N}(X\beta^*, \sigma^2 \mathrm{Id}_n),$$

for some $\beta^* \in \mathbb{R}^p$ which is unknown and some noise variance $\sigma^2 > 0$.

When the number of features n is significantly smaller than d or/and when we have the prior knowledge that only a small number of predictors are relevant to explain the response variable, it is common to consider the Lasso problem

$$\hat{\beta}_\lambda \in \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}.$$

This problem has been extensively studied from a theoretical point of view, and one can mention for example that if the entries of the predictor matrix X are drawn from a continuous probability distribution, then the lasso solution is unique with probability one (regardless of the size of n and p). It is now well-known that optimal solutions of the Lasso problem are sparse, meaning that they have some zero coefficients. Actually the larger λ, the larger the number of zeros in $\hat{\beta}_\lambda$. Hence, varying the regularization parameter λ, we obtain different models in which more or less variables have a non-zero coefficient. Intuitively, it seems reasonable to select variables whose fitted coefficient is (in absolute value) above some significance threshold. However, it is not an easy task to choose the threshold (or the value of λ) in such a way as to control the Type-I error.

The difficulty arises from the distribution of the estimated coefficients for the null variables being unknown, but at least some of them most likely being non-zero. Moreover, the fitted coefficients are correlated among each other and an incorrect threshold can yield either a very high proportion of false discoveries (if too low) or very low power (if too high).

2 The Candès' approach

To overcome the above mentioned difficulties Candès and Barber in [1] introduced the *knockoff filter*. The method consists in constructing dummies covariates $\widetilde{X}_j$ that mimic the correlation structure of the true features. The so-called *knockoffs* features $\widetilde{X}_j$ should be understood as a **negative control group**. This imitation needs to be carefully designed to ensure a procedure controlling the FDR.

Considering the features X_j have been normalized – meaning that $\Sigma_{j,j} = \|X_j\|_2 = 1$ for $\Sigma = X^\top X$ – a $n \times p$ matrix $\widetilde{X}$ give valid knockoffs if

(i) $\widetilde{X}^\top \widetilde{X} = \Sigma$.

(ii) $X^\top \widetilde{X} = \Sigma - \text{diag}(s)$ where s is a p-dimensional nonnegative vector.

Condition (i) ensures that $\widetilde{X}$ has the same covariance structure as the original matrix X. Condition (ii) states that the correlations between distinct original and knockoff variables are the same as those between the originals. However, comparing a feature X_j to its knockoff $\widetilde{X}_j$, we see that

$$\widetilde{X}_j^\top X_j = \Sigma_{j,j} - s_j = 1 - s_j.$$

The role of s is to make $\widetilde{X}$ as uncorrelated with X as possible in order to increase statistical power during variable selection.

Define $Z_j = \sup\{\lambda \,:\, \hat{\beta}(\lambda) \neq 0\}$ for $j = 1, \dots, 2p$ where $\hat{\beta}(\lambda)$ is the solution to the augmented Lasso model regressing y on $[X\ \widetilde{X}]$. We define $W_j = (Z_j \vee Z_{j+p})\text{sign}(Z_j - Z_{j+p})$. Note that Z_j corresponds to the point λ on the Lasso regularization path at which feature X_j first enters the model. We then expect that Z_j is large for most of the signals, and small for most of the null variables. Hence, a large positive value of W_j indicates that variable X_j enters the Lasso model early (at some large value of λ) and that it does so before its knockoff copy $\widetilde{X}_j$. Hence, this is an indication that this variable is a genuine signal and belongs in the model.

We wish to select variables such that W_j is large and positive, i.e. such that $W_j \geq t$ for some $t > 0$. Letting q be the target FDR, define a data-dependent threshold T as:

$$T = \min\{t > 0 \,:\, \frac{1 + \#\{j \,:\, W_j \leq -t\}}{\#\{j \,:\, W_j \geq t\} \vee 1} \leq q\},$$

(or $T = +\infty$ if this set is empty).

1. Justify that when $T < +\infty$, T belongs to the set $\mathscr{W} = \{|W_j| \,:\, j = 1, \dots, p\} \backslash \{0\}$.

In the following, we consider that $n > 2p$.

3 Construction of Knockoffs

2. Give a lower and an upper bound on the $(s_j)_j$ to ensure that $\widetilde{X}$ exists. You may use the fact that for any symmetric matrix, $M = \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix}$ with D a (square) positive definite matrix, M is positive semidefinite if and only if the Schur complement of the block D given by $A - BD^{-1}B^\top$ is positive semidefinite.

3. Show that for s satisfying the condition of the previous question, $\widetilde{X} = X(I - \Sigma^{-1}\text{diag}(s)) + \widetilde{U}C$ leads to valid knockoffs where $\widetilde{U}$ is an $n \times p$ orthonormal matrix that is orthogonal to the span of the features X, and $C^\top C = 2\text{diag}(s) - \text{diag}(s)\Sigma^{-1}\text{diag}(s)$ is a Cholesky decomposition.

4. Explain why this approach does not work anymore if $2p > n$.

5. Actually a simple trick can be used to extend the previous knockoffs construction to the case where $p \leq n < 2p$. If the noise level $\sigma^2 > 0$ is known, one can simply draw $y' \sim \mathcal{N}(0, \sigma^2 \text{Id}_{2p-n})$. Then by independence of the gaussian vector y and y' it holds

$$\begin{bmatrix} y \\ y' \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} X \\ 0 \end{bmatrix}\beta, \sigma^2 \text{Id}\right),$$

and we can apply the previous procedure to this row-augmented data.
In the case where σ^2 is unknown, propose a modification of this approach to obtain knockoffs. Give the details for a practical implementation of this method.

4 Exchangeability results

6. Pairwise exchangeability for the features.
   Prove that for any subset $S \subset \{1,\dots,p\}$,
   $$[X\ \widetilde{X}]_{swap(S)}^\top [X\ \widetilde{X}]_{swap(S)} = [X\ \widetilde{X}]^\top [X\ \widetilde{X}]_{swap(S)}.$$
   That is, the Gram matrix of $[X\ \widetilde{X}]$ is unchanged when we swap X_j and $\widetilde{X}_j$ for each $j \in S$.

7. Pairwise exchangeability for the response.
   Prove that for any subset S of nulls,
   $$[X\ \widetilde{X}]_{swap(S)}^\top y \overset{d}{=} [X\ \widetilde{X}]^\top y.$$
   That is, the distribution of the product $[X\ \widetilde{X}]^\top y$ is unchanged when we swap X_j and $\widetilde{X}_j$ for each $j \in S$, as long as none of the swapped features appear in the true model.

8. i.i.d. signs for the nulls.
   Let $\varepsilon \in \{\pm 1\}^p$ be a sign sequence independent of W, with $\varepsilon_j = +1$ for all non-null j and $\varepsilon_j \overset{i.i.d.}{\sim} \{\pm 1\}$ for null j. Then
   $$(W_1,\dots,W_p) \overset{d}{=} (W_1\varepsilon_1,\dots,W_p\varepsilon_p).$$
   You may prove and use the antisymmetry property which states that swapping X_j and $\widetilde{X}_j$ has the effect of switching the sign of W_j.

5 Recasting Knockoff filter as a sequential testing procedure

Let $m = \#\{j\ :\ W_j \neq 0\}$. The knockoff filter never selects variable j when $W_j = 0$, hence we can ignore such variables. Assume without loss of generality that $|W_1| \geq |W_2| \geq \cdots \geq |W_m| > 0$, and set

$$\forall j \in [m], \quad p_j = \frac{1}{2}\mathbb{1}_{W_j>0} + \mathbb{1}_{W_j<0},$$

which can be thought of as 1-bit p-values. It then follows from Lemma 1 that the null p-values are i.i.d. with $\mathbb{P}\big(p_j = 1/2\big) = 1/2 = \mathbb{P}\big(p_j = 1\big)$ and are independent from the others. Setting K to be the indices of the strict inequalities,

$$K = \{k \in [m]\ :\ |W_k| > |W_{k+1}|\} \cup \{m\},$$

one sees that the correlation-preserving proxy method is now equivalent to a sequential testing procedure on these p-values. Indeed for any $k \in K$,

$$\frac{1 + \#\{j \leq k\ :\ p_j > 1/2\}}{\#\{j \leq k\ :\ p_j \leq 1/2\} \vee 1} = \frac{1 + \#\{j \leq k\ :\ W_j < 0\}}{\#\{j \leq k\ :\ W_j > 0\} \vee 1} = \frac{1 + \#\{j\ :\ W_j \leq -|W_k|\}}{\#\{j\ :\ W_j \geq |W_k|\} \vee 1}.$$

The first equality follows from the definition of p_j. The second equality holds because the absolute values of W are arranged in nonincreasing order; by definition of K, the inequality $|W_j| \geq |W_k|$ is only possible if it holds that $j \leq k$. Therefore $W_j \geq |W_k|$ (respectively, $W_j \leq -|W_k|$) is true if and only if $j \leq k$ and $W_j > 0$ (respectively, $j \leq k$ and $W_j < 0$). Hence, finding the largest k such that the ratio in the left-hand side is below q is the same as finding the smallest $|W_k|$ such that the right-hand side is below q. This is equivalent to finding the minimum $t \in \mathcal{W}$ such that

$$\frac{1 + \#\{j\ :\ W_j \leq -t\}}{\#\{j\ :\ W_j \geq t\} \vee 1} \leq q,$$

which are the knockoff and knockoff+ thresholds. Finally, rejecting the p-values obeying $p_j \leq 1/2$ is the same as rejecting the positive W_j 's.

6 Proof

9. For $k = m, m-1, \dots, 1, 0$, put

$$V^+(k) = \#\{null\ j\ :\ 1 \le j \le k,\ p_j \le \frac{1}{2}\}, \quad \text{and} \quad V^-(k) = \#\{null\ j\ :\ 1 \le j \le k,\ p_j > \frac{1}{2}\}$$

with the convention that $V^{\pm}(0) = 0$. Let $\mathscr{F}_k$ be the filtration defined by knowing all the non-null p-values, as well as $V^{\pm}(k_0)$ for all $k_0 \ge k$.
Prove then that the process

$$M(k) = \frac{V^+(k)}{1 + V^-(k)}$$

is a super-martingale running backward in time with respect to $\mathscr{F}_k$.

10. Deduce that

$$\mathbb{E}\left[M(\hat{k})\right] \le \mathbb{E}\left[\frac{\#\{null\ j\ :\ 1 \le j \le k,\ p_j \le \frac{1}{2}\}}{1 + \#\{null\ j\ :\ 1 \le j \le k,\ p_j > \frac{1}{2}\}}\right].$$

11. Set $X = \#\{null\ j\ :\ 1 \le j \le k,\ p_j \le \frac{1}{2}\}$. The independence of the nulls together with the stochastic dominance $p_j \overset{d}{\ge} \text{Unif}[0,1]$[1] valid for all nulls imply that $X \overset{d}{\le} Y$, where $Y \sim \text{Binomial}(N, \frac{1}{2})$, where N is the total number of nulls.
Using this property, show that

$$\mathbb{E}\left[\frac{X}{1 + N - X}\right] \le 1.$$

12. Recall that $V = \#\{null\ j \le \hat{k}\ :\ p_j \le \frac{1}{2}\}$ and $R = \#\{j \le \hat{k}\ :\ p_j \le \frac{1}{2}\}$. Conclude by showing that

$$\mathbb{E}\left[\frac{V}{R \vee 1}\right] \le q.$$

7 Discussion

Multiple testing is an active field of research and a large span of methods have recently emerged. Until recently, most methods were assuming independent p-values. The knockoff filter has the advantage to bypass this limitation. Nevertheless, the above version of Knockoffs can only be use when $n > p$ which is a strong assumption and makes the method inapplicable in most industrial problems that are typically high dimensional. Candès and al. proposed an extension of the Knockoff filter in [2] that can be used in high dimensional settings. This approach consider that the design is randomly sampled from a known distribution. The method is proved to control the FDR. The generation of knockoffs is no more deterministic but random and can be tricky to implement in practice. That's why Candès and al. introduced in [3] a deep learning approach to learn knockoffs to overcome the previous computational issue (but we lose the nice theoretical guarantees regarding FDR control).

References

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

[2] E. Candes, Y. Fan, L. Janson, and J. Lv. Panning for gold: Model-x knockoffs for high-dimensional controlled variable selection. *arXiv preprint arXiv:1610.02351*, 2016.

[3] Y. Romano, M. Sesia, and E. Candès. Deep knockoffs. *Journal of the American Statistical Association*, 115(532):1861–1872, Oct 2019.

[1] that is, for all null j and all $u \in [0,1]$, $\mathbb{P}\left(p_j \le u\right) \le u$.