STAT 689: Large-scale and high-dimensional statistical inference **Fall 2024**

Lecture 10: Part I

1 Variable selection problem

Consider a response variable Y and a set of covariates $X_1, \ldots, X_p$ in the number of thousands or millions. We want to select a subset of interesting covariates that can affect the response Y. For example, in a genome-wide association study (GWAS), we may be interested in selecting genes X_j that a phenotype Y of interest truly depends on. Mathematically, we define the null variable to be the one that satisfies

$$Y \perp\!\!\!\perp X_j | X_{-j}$$

that is X_j is conditionally independent of Y given $X_{-j} = \{X_i : 1 \leq i \leq p, i \neq j\}$.

Consider a linear model

$$Y = \sum_{j=1}^{p} X_j \beta_j + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$ which is independent of the covariates $(X_1, \ldots, X_p)$. In this case, testing $Y \perp\!\!\!\perp X_j | X_{-j}$ is equivalent to testing $\beta_j = 0$. Many variable selection methods compute an important statistic for each covariate. These statistics serve as a basis for deciding whether to include the variable in our model. For example, the feature importance statistic may be the magnitude of a coefficient computed with the Lasso, the point at which a variable enters the Lasso path or even a more complicated feature importance statistic computed with random forests or neural nets.

2 Conditional randomization tests

The conditional randomization tests work as follows. Suppose we know the conditional distribution of X_j given X_{-j}. Then, we can sample a synthetic null $\tilde{X}_j$ from this conditional distribution. Under the null $H_{0,j} : Y \perp\!\!\!\perp X_j | X_{-j}$, we have

$$\begin{aligned}
\mathbb{P}(Y, X_j, X_{-j}) =& \mathbb{P}(Y, X_j | X_{-j}) \mathbb{P}(X_{-j}) \\
=& \mathbb{P}(X_j | X_{-j}) \mathbb{P}(Y | X_{-j}) \mathbb{P}(X_{-j}) \\
=& \mathbb{P}(\tilde{X}_j | X_{-j}) \mathbb{P}(Y | X_{-j}) \mathbb{P}(X_{-j}) \\
=& \mathbb{P}(Y, \tilde{X}_j, X_{-j}).
\end{aligned}$$

In words, (X_j, X_{-j}, Y) and $(\tilde{X}_j, X_{-j}, Y)$ have the same joint distribution when $Y \perp\!\!\!\perp X_j | X_{-j}$. Thus, we can test if

$$(Y, X_j, X_{-j}) \stackrel{d}{=} (Y, \tilde{X}_j, X_{-j})$$

to decide if X_j is under the null.

By generating multiple samples from the conditional distribution of X_j given X_{-j}, we can compute the p-values for testing $H_{0,j}$.

1. Compute an important statistic $T_j = T(Y, X_j, X_{-j})$ (assume that a larger value provides stronger evidence against the null).
2. For $b = 1, 2, \ldots, B$, sample $\tilde{X}_j^{(b)}$ from the the conditional distribution of X_j given X_{-j} and compute $T_j^{(b)} = T(Y, \tilde{X}_j^{(b)}, X_{-j})$.

3. Compute

$$p_j := \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{T_j^{(b)} \geq T_j\}}{B+1}.$$

Remark. In practice, we have multiple samples. So you can think of $X_j = (X_{j,1}, \ldots, X_{j,n})$, where $X_{j,i}$ is the jth covariate from the ith sample.

Note that if X_j is indeed null, then the resulting p_j is indeed a p-value. This follows since, under the null,

$$(Y, X_j, X_{-j}) \stackrel{d}{=} (Y, \tilde{X}_j^{(b)}, X_{-j})$$

would hold for all b and hence $T_j, T_j^{(1)}, \ldots, T_j^{(B)}$ are identically distributed.

Exercise 10.1: Show that when X_j is null, for any $\alpha \in (0,1)$, we have

$$\mathbb{P}(p_j \leq \alpha) \leq \alpha.$$

Notice that in the above procedure, it is crucial to sample from the conditional, and not the marginal of X_j, in order to preserve the dependence structure between X_j and the other covariates. For example, suppose we have X_1, X_2, and Y with $\text{cov}(X_1, X_2) = 0.5$ and

$$Y = X_2 + \epsilon, \quad \epsilon \sim N(0, \sigma^2).$$

Here $\mathbb{E}[X_1] = \mathbb{E}[X_2] = 0$ and ϵ is independent of (X_1, X_2). Then, we have

$$\mathbb{E}[YX_1] = \mathbb{E}[(X_2 + \epsilon)X_1] = \mathbb{E}[X_2 X_1] = 0.5.$$

If $\tilde{X}_1$ is generated according to the marginal distribution of X_1, then $\tilde{X}_1$ is independent of Y and hence

$$\mathbb{E}[Y\tilde{X}_1] = 0 \neq \mathbb{E}[YX_1] = 0.5,$$

even though X_1 is null in the above model. Therefore, sampling from the marginal does not provide good control: X_1 would likely appear to be significant since we would expect the resulting p-value from the above procedure to be small.

3 Knockoffs: the fixed design case

Consider the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{Y} \in \mathbb{R}^{n \times 1}$ is a response variable, $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ is a vector of coefficients, and $\boldsymbol{\epsilon} \in \mathbb{R}^{n \times 1} \sim N(0, \sigma^2 \mathbf{I}_n)$. Suppose we have a variable selection method that returns a set of indices $\hat{S} \subseteq \{1, 2, \ldots, p\}$. In the context of variable selection, we define the FDR to be

$$\text{FDR} = \mathbb{E}\left[\frac{\#\{j : \beta_j = 0, j \in \hat{S}\}}{1 \vee |\hat{S}|}\right]$$

where $|\hat{S}|$ denotes the size of the set $\hat{S}$.

The goal of the knockoff filter is to provide a variable selection method that controls the FDR at level α for any finite sample of data whenever the number of observations n is larger than the number of variables p. Below, we describe the knockoff filter for the case of fixed design matrix $\mathbf{X}$.

3.1 Constructing the knockoffs

Let $\Sigma = \mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{p\times p}$ be the Gram matrix and $\mathbf{X}_j$ be the jth column of the design matrix $\mathbf{X}$. In practice, we can normalize the columns of $\mathbf{X}$ so that $\mathbf{X}_j^\top \mathbf{X}_j = 1$ for all j. We construct a knockoff copy $\tilde{\mathbf{X}}_j$ such that

A. $\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} = \Sigma$

B. $\mathbf{X}^\top \tilde{\mathbf{X}} = \Sigma - \text{diag}(\mathbf{s})$

where $\tilde{\mathbf{X}} = [\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_p]$ and $\mathbf{s} = (s_1, \ldots, s_p)$ with $s_i \geq 0$. The above requirements imply that

1. $\tilde{\mathbf{X}}_j^\top \tilde{\mathbf{X}}_k = \mathbf{X}_j^\top \mathbf{X}_k$ for all $1 \leq j, k \leq p$;
2. $\mathbf{X}_j^\top \tilde{\mathbf{X}}_k = \mathbf{X}_j^\top \mathbf{X}_k$ for $1 \leq j \neq k \leq p$;
3. $\mathbf{X}_j^\top \tilde{\mathbf{X}}_j = \mathbf{X}_j^\top \mathbf{X}_j - s_j = 1 - s_j$ for $1 \leq j \leq p$.

Exercise 10.2: A strategy to construct the knockoff copies is through the following method. Choose the non-negative vector $\mathbf{s}$ such that $2\Sigma - \text{diag}(\mathbf{s})$ is positive semi-definite.

- Show that $\mathbf{G} := 2\text{diag}(\mathbf{s}) - \text{diag}(\mathbf{s})\Sigma^{-1}\text{diag}(\mathbf{s})$ is positive semi-definite. As a result, there exists a Cholesky decomposition $\mathbf{C}$ such that $\mathbf{C}^\top \mathbf{C} = \mathbf{G}$.
- Set $\tilde{\mathbf{X}} = \mathbf{X}(\mathbf{I}_p - \Sigma^{-1}\text{diag}(\mathbf{s})) + \tilde{\mathbf{U}}\mathbf{C}$, where $\mathbf{U} \in \mathbb{R}^{n\times p}$ is an orthonormal matrix that is orthogonal to the span of the features of $\mathbf{X}$. Show that $\tilde{\mathbf{X}}$ satisfies the above requirements.

3.2 Calculate statistics

We define statistics T_j for each variable such that large positive values give evidence against the null hypothesis (or support that the $j$the variable is a signal). One way to construct such a statistic is through Lasso. The Lasso estimator for the regression coefficients can be defined as

$$\hat{\boldsymbol{\beta}}(\lambda) = \text{argmin}_{\mathbf{b}} \left\{ \|\mathbf{Y} - \mathbf{X}\mathbf{b}\|_2^2 + \lambda \|\mathbf{b}\|_1 \right\}$$

where $\|\cdot\|_p$ denotes the l_p norm of a vector, and λ is a tuning parameter. As λ gets larger, the components of $\hat{\boldsymbol{\beta}}(\lambda)$ gets smaller in absolute value. For each j, we let

$$Z_j = \sup\{\lambda \geq 0 : \hat{\beta}_j(\lambda) \neq 0\},$$

which is the largest λ such that the jth variable enters the model.

To use the knockoff method, we first run Lasso on an augmented matrix, which consists of concatenating the original design and knockoff design matrices, which can be written as $[\mathbf{X}, \tilde{\mathbf{X}}]$. The Lasso method gives $2p$ statistics, namely $Z_1, \ldots, Z_p$ and $\tilde{Z}_1, \ldots, \tilde{Z}_p$. Next, we define

$$W_j = \begin{cases} Z_j & \text{if } Z_j > \tilde{Z}_j, \\ -\tilde{Z}_j & \text{if } Z_j < \tilde{Z}_j. \end{cases}$$

3.3 Calculate the cutoff to control FDR

For a desired FDR level α, we define the threshold

$$T = \min \left\{ t \in \mathcal{W} : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq \alpha \right\},$$

where $\mathcal{W} = \{|W_j| : j = 1, 2, \ldots, p\} \setminus \{0\}$. This is essentially the BC procedure we have learned before. The constant one in the numerator is essential for achieving the FDR control in theory. The selected model $\hat{S}$ is given by

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

Theorem. The above procedure controls the FDR at level α, i.e.,

$$\text{FDR} = \mathbb{E}\left[\frac{\#\{j : \beta_j = 0, j \in \hat{S}\}}{1 \vee |\hat{S}|}\right] \leq \alpha.$$

Proof. Check Barber and Candès (2015, AOS). This proof relies on showing that W_j for $\beta_j = 0$ is symmetric about zero.

3.4 Rationale behind knockoffs

We first try to answer the question of why $\tilde{\mathbf{X}}$ has to be constructed in the above way. The knockoff dataset $\tilde{\mathbf{X}}$ can be thought of as the solution to the desired correlation structure:

$$[\mathbf{X}, \tilde{\mathbf{X}}]^\top [\mathbf{X}, \tilde{\mathbf{X}}] = \begin{pmatrix} \Sigma & \Sigma - \text{diag}(\mathbf{s}) \\ \Sigma - \text{diag}(\mathbf{s}) & \Sigma \end{pmatrix} = \Gamma,$$

where the first equality follows from Conditions (A)-(B). The condition that $\mathbf{G}$ is positive semi-definite ensures that Γ is positive semi-definite.

Let $S = \{j : \beta_j \neq 0\}$ be the active set. For $j \in S$, we expect Z_j to be greater than $\tilde{Z}_j$ and hence $W_j = Z_j > 0$. On the other hand, if $j \notin S$, intuitively $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ play the same role. So we expect W_j to be Z_j (or $-\tilde{Z}_j$) with the same probability, i.e., $\text{sign}(W_j) \sim \text{Bern}(0.5)$. Therefore, we have

$$\#\{j : W_j \geq t, j \notin S\} \approx \#\{j : W_j \leq -t, j \notin S\}$$

which implies that

$$\text{FDP}(t) \leq \frac{1 + \#\{j : W_j \geq t, j \notin S\}}{\#\{j : W_j \geq t\} \vee 1} \approx \frac{1 + \#\{j : W_j \leq -t, j \notin S\}}{\#\{j : W_j \geq t\} \vee 1} \leq \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1}.$$

So, the LHS can be viewed as a conservative estimate of the FDP based on the rejection rule $W_j \geq t$.

What properties do W_j's have to satisfy in general? Let $\mathbf{W} = (W_1, \ldots, W_p)$. Write $[\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(\tilde{S})}$ to mean that the columns $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ have been swapped in the matrix $[\mathbf{X}, \tilde{\mathbf{X}}]$ for each $j \in \tilde{S}$, where $\tilde{S} \subseteq \{1, 2, \ldots, p\}$. Suppose

$$\mathbf{W} = \mathbf{W}([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{Y}) = f([\mathbf{X}, \tilde{\mathbf{X}}]^\top [\mathbf{X}, \tilde{\mathbf{X}}], [\mathbf{X}, \tilde{\mathbf{X}}]\mathbf{Y}),$$

for some function f. We require $\mathbf{W}$ to obey the antisymmetry property

$$\mathbf{W}_j([\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(\tilde{S})}, \mathbf{Y}) = \mathbf{W}_j([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{Y}) \cdot \begin{cases} +1 & \text{if } j \notin \tilde{S}, \\ -1 & \text{if } j \in \tilde{S}. \end{cases}$$

Moreover, we require

$$(W_1, \ldots, W_p) \overset{d}{=} (W_1 e_1, \ldots, W_p e_p),$$

where $e_j = 1$ for $j \in S$ and $P(e_j = \pm 1) = 1/2$ for $j \notin S$.

3.5 How to choose s?

Intuitively, we want $\tilde{\mathbf{X}}_j$ to be different from $\mathbf{X}_j$. Note that by construction, $\mathbf{X}_j^\top \tilde{\mathbf{X}}_j = 1 - s_j$. Thus, we want to make $|1 - s_j|$ as small as possible. One way to choose s_j's is by solving the following optimization problem:

$$\min_{\mathbf{s}} \sum_j |1 - s_j|$$

subject to $s_j \geq 0$ and $2\Sigma - \text{diag}(\mathbf{s})$ is positive semi-definite.