Suppose we wish to perform controlled variable selection in a linear model of the form
\[ y = X\beta + \epsilon, \]
where \( y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times p}, \beta \in \mathbb{R}^p \) and \( \epsilon \) is a vector of independent standard normal errors. We wish to test \( H_j : \beta_j = 0 \) (vs. \( \beta_j \neq 0 \)) while controlling the FDR. Here and below, we assume \( X \) is fixed and \( n \geq p \) (in fact, we shall assume \( n \geq 2p \) for convenience). We shall study fixed-X knockoffs.

1. Suppose we apply Benjamini-Hochberg to classical ordinary least-squares p-values. Would this control the FDR?

2. Assume we have generated another matrix \( \tilde{X} \in \mathbb{R}^{n \times p} \) with the following properties:
   \[ \tilde{X}^\top \tilde{X} = X^\top X, \quad \text{offdiag}(\tilde{X}^\top X) = \text{offdiag}(X^\top X); \]
the second equality means that all off-diagonal entries match. We will call such a matrix a knockoff matrix. As in model-X knockoffs, suppose we have feature importance statistics \( Z = z(X,Y) = z(X^\top X, X^\top y) \) where the equality means that \( z \) only depends on the problem data through \( X^\top X \) and \( X^\top y \). We run the importance statistic on \( [X, \tilde{X}] \) and \( y \) and obtain scores \( Z \) and \( \tilde{Z} \) which are combined as usual:
\[ W_j = w(Z_j, \tilde{Z}_j), \]
with \( w \) antisymmetric. Show that the signs of the null \( W_j \)'s are i.i.d Ber(1/2), and are independent of \( |W| \).

3. If we apply SeqStep to the \( W_j \)'s, would we control the FDR?

4. Run a simulation to examine the power of this technique. You probably want to use the R knockoff package available [here]. This should involve minimum coding since all the heavy lifting is done by the package.
   - Set \( n = 3000, p = 1000 \) and \( \alpha = 0.2 \).
   - Sample a matrix \( X \) with i.i.d. \( \mathcal{N}(0, 1/n) \) entries and normalize the columns so they have unit norm.
   - Sample a vector \( \beta \) with 35 entries equal to 3.5 and all others equal to zero.
   - Create a knockoff matrix using either the SDP strategy, the equi-correlated strategy, or (ideally) both.
   - Sample 200 realizations of \( y \) and report the FDR and power of the procedure when \( \text{offset} \) is either 0 or 1: here, \( \text{offset} \) is a parameter in the SeqStep procedure
   \[ \tau = \min \left\{ t : \frac{\text{offset} + |\{j : W_j \leq -t\}|}{1 \vee |\{j : W_j \geq t\}|} \leq \alpha \right\}. \]
   We shall use
   \[ W_j = Z_j - \tilde{Z}_j \]
   where \( Z_j = |\beta_j(\lambda = 1.6)| \) is the magnitude of the lasso coefficient at \( \lambda = 1.6 \). That is, \( \beta(\lambda), \tilde{\beta}(\lambda) \) are solutions to
   \[ \text{minimize } \frac{1}{2} ||y - Xb - \tilde{X}\tilde{b}||_2^2 + \lambda||b||_1 + \lambda||\tilde{b}||_1. \]
   - Compare FDR and power with those achieved via the BH procedure.
   - Comment on your results.

\[ \text{1If you work in Python, let me know as I also have some code.} \]