universität wien

MAGISTERARBEIT / MASTER'S THESIS

Titel der Magisterarbeit / Title of the Master's Thesis

„Variable selection via fixed-X and model-X Knockoff procedures – detailed proofs“

verfasst von / submitted by

Mathias Wörndl, BSc BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Magister der Sozial- und Wirtschaftswissenschaften (Mag. rer. soc. oec.)

Wien, 2020 / Vienna 2020

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 951
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Magisterstudium Statistik
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Hannes Leeb

Contents

1 General introduction **6**

2 Fixed-X Knockoff methods **6**
2.1 Introduction . . . 6
2.2 FDR-Control . . . 9
2.3 Power . . . 49
2.4 Examples . . . 50
2.4.1 Exact construction of FX Knockoffs . . . 50
2.4.2 Approximate construction of FX Knockoffs (if $n < 2 \cdot p$) . . . 56
2.4.3 Construction of FX-Knockoff feature statistics . . . 56

3 Model-X Knockoff methods **61**
3.1 Introduction . . . 61
3.2 FDR Control . . . 70
3.3 Power . . . 102
3.4 Examples . . . 103
3.4.1 Exact construction of MX Knockoffs . . . 103
3.4.1.1 SCIP Algorithm . . . 106
3.4.1.2 Special case - Gaussian rows . . . 113
3.4.2 Approximate construction of MX Knockoffs . . . 117
3.4.2.1 Second-order MX-Knockoffs . . . 117
3.4.3 Construction of MX-Knockoff feature statistics . . . 118

A Appendix **122**

B Appendix **125**

C Appendix - English abstract **131**

D Appendix - German abstract **132**

Notation and General Conventions

i.i.d.	independent and identically distributed (random variables)
cdf	cumulative distribution function
LHS, RHS	left hand side, right hand side
Prep	Preparation (used in proofs to denote supportive parts)
wlog	without loss of generality
w.r.t.	with respect to
wts	want to show (used in proofs)
$a \wedge b$	minimum of $\{a, b\}$
$a \vee b$	maximum of $\{a, b\}$
$A := B$ or $B =: A$	A is defined as B
$\mathbb{1}_A(\cdot)$	indicator on A
$A \subseteq B$	A is a subset of B
$A \subsetneq B$ or $A \subset B$	A is a proper/strict subset of B
$A \preceq B$	matrix A is smaller than matrix B in Löwner partial order
X^T	transpose of X
$\text{diag}(a)$	diagonal matrix with entries of vector a on its diagonal
I_n	$n \times n$ identity matrix
$(W_1, \ldots, W_p)$	row vector containing entries W_1 up to W_p
$\begin{pmatrix} a \\ b \end{pmatrix}$	column vector or binomial coefficient (depends on context)
$e^{\cdot}$ or $\exp(\cdot)$	exponential function of " $\cdot$ "
$\ln(\cdot)$	natural logarithm (logarithm to base e)
$\lvert A \rvert$	cardinality if A is a set, absolute value if A is a number
$\lvert\lvert x \rvert\rvert_2$	euclidean norm of the vector x
$k!$	factorial of k
$X \perp\!\!\!\perp Y$	X and Y are independent
$\sigma(\cdot)$	sigma-algebra
$\mathbb{P}(\cdot)$	probability
$\mathbb{E}(\cdot)$	expectation
$\mathbb{V}ar(\cdot)$	variance
$\mathbb{C}ov(\cdot)$	covariance, cross-covariance or covariance matrix (depends on context)
$\mathcal{L}(\cdot)$	law/distribution
$\mathcal{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
$\sim$	distributed as
$X \stackrel{d}{=} Y$	X and Y are equal in distribution
$\mathbb{R}$	real numbers
$\mathbb{R}_{\geq 0}$	non-negative real numbers
$[0, 1]$	interval of real numbers between 0 and 1
$\emptyset$	empty set
Ω	sample space (set of all possible outcomes)

Comments for Readers

This master thesis is written for people who are interested in the theoretical background of Knockoff procedures. It can be used as lecture notes to teach fixed-X and model-X Knockoff procedures. The requirements for understanding the proofs are basic knowledge of probability theory (including conditional expectation and martingales) and linear model theory. Experience with other variable selection procedures is helpful. No asymptotic theory is used.

In contrast to the two main papers, on which this thesis is based on, results are presented in a structured definition-lemma-theorem style with detailed proofs. Instead of a short summary of important steps, proofs are presented by showing all the small steps. Longer equations are aligned in a consistent way. Comments inside such equations are placed on the right hand side. A dot is used as multiplication symbol to avoid misinterpretation. Vectors and matrices are boldfaced throughout the thesis whereas scalars are not (exceptions are highlighted separately).

Regarding citation/plagiarism, it should be noted that none of the procedures presented in this thesis are my own work. Most of the parts are taken from two papers, which are presented at the beginning of the chapters. Other parts are cited separately. My work was to present findings of the papers in a mathematically structured way, to prove claims where no proof was given and to expand existing proofs and proof sketches. The amount of "copied parts" can be checked by simply comparing this thesis with the two main papers.

1 General introduction

In many fields of science, a response variable (called response throughout the thesis) together with a large number of potential explanatory variables (called variables/features/covariates) is observed and researchers would like to know which variables are truly associated with the response. There exists a wide variety of variable selection procedures which can be used to find important variables. In terms of null hypothesis testing, such procedures can often be pooled in different type I error control groups. One such group consists of methods which have theoretical control of the false discovery rate (FDR), the expected fraction of false discoveries among all discoveries (Benjamini and Hochberg, 1995). This type I error control group is used in settings, where high power is preferred at the cost of an increased number of false discoveries. Genetics would be an application, where researchers would like to reduce the number of markers/genes/etc which are associated with a response (e.g. a specific cancer) from potentially thousands or millions to around a dozen first, so that each one of those candidates can be investigated carefully in a second step.
In this thesis, essentially two (actually, it will be four) FDR-controlling variable selection procedures are discussed. The first one, called fixed-X Knockoff method, is designed for a common linear model setting where the number of observations has to be higher than the number of variables.
The second procedure, called model-X Knockoff method, is using ideas of the first procedure but is designed for a totally different setting. Instead of assuming knowledge of the (conditional) distribution of the response of a linear model and requiring a high number of observations, it assumes exact knowledge of the joint distribution of the variables and can work with an arbitrary number of covariates.
Both, the fixed-X and model-X methods, use the same ideas to tease apart important and unimportant variables. First, "Knockoffs" are constructed for every variable which mimic the structure of the original variables. Then, instead of using p-values to select variables, both methods use "Knockoff feature statistics" as variable importance measure. In the construction of feature statistics, Knockoffs are used to decide which variables are important. Finally, variables whose feature statistics are higher than a specific threshold are selected.

The fixed-X and model-X methods are presented in a way, such that comparison will be easy. Because both methods are similar to a large extent, the ideas are mainly explained in the fixed-X parts.

2 Fixed-X Knockoff methods

Fixed-X Knockoff procedures were introduced first by Barber and Candes (2015) and the whole chapter is based on that paper.

2.1 Introduction

We assume that the observations obey the linear regression model

$$\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{z}, \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^n$ is a vector of responses, $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a known and fixed design matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown vector of coefficients β_j and $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \cdot \mathbf{I}_n)$ is Gaussian noise. We assume that $n \geq 2 \cdot p$. The Gram matrix $\mathbf{X}^T \cdot \mathbf{X}$ is assumed to be invertible as the model would otherwise not be identifiable.

Definition 2.1. An index j or variable $\mathbf{X}_j$ (column of $\mathbf{X}$) is said to be *null* if $\beta_j = 0$ and *non-null* if $\beta_j \neq 0$ in the true model.

Comments:
i) Of course, we cannot know which variables are null and which are non-null in practice. If we would know, there would be no reason to apply variable selection techniques in the first place. The definition of nulls and non-nulls is needed to study how well our procedures work in theory.
ii) In this chapter, we will present two variable selection procedures, which attempt to discover as many non-nulls as possible while keeping the FDR (or a modified version of it) under control.
iii) The first step of both procedures is to create one fake variable for each original variable $\mathbf{X}_j$. These fake variables will be used to decide which of the original variables are important. In a certain way, these fake versions have to mimic the structure of the original variables as the next definition shows.

Definition 2.2. Let $\mathbf{X}$ be a fixed design matrix with normalized columns $\mathbf{X}_j$; that is, $||\mathbf{X}_j||_2^2 = 1$ for all $j \in \{1, \dots, p\}$. A new matrix $\tilde{\mathbf{X}}$ is called *FX-Knockoff matrix* (with *FX-Knockoff vectors* $\tilde{\mathbf{X}}_j$ as columns), if the following assumptions are satisfied:
i) $\tilde{\mathbf{X}}^T \cdot \tilde{\mathbf{X}} = \mathbf{X}^T \cdot \mathbf{X}$
ii) $\mathbf{X}^T \cdot \tilde{\mathbf{X}} = \mathbf{X}^T \cdot \mathbf{X} - \text{diag}(\boldsymbol{s})$, where $\boldsymbol{s}$ is a p-dimensional vector satisfying $\text{diag}(\boldsymbol{s}) \succeq \mathbf{0}$ and $2 \cdot \mathbf{X}^T \cdot \mathbf{X} - \text{diag}(\boldsymbol{s}) \succeq \mathbf{0}$.

Comments:
a) Note that $\succeq$ denotes the Löwner partial order.
b) The reason why $\boldsymbol{s}$ has to satisfy $\text{diag}(\boldsymbol{s}) \succeq \mathbf{0}$ and $2 \cdot \mathbf{X}^T \cdot \mathbf{X} - \text{diag}(\boldsymbol{s}) \succeq \mathbf{0}$ is given in Lemma 2.14.
c) We denote by $\left[\mathbf{X}, \tilde{\mathbf{X}}\right] \in \mathbb{R}^{n \times 2 \cdot p}$ the columnwise concatenation of the original design matrix and the FX-Knockoff matrix.
d) This definition implies that all derived properties and procedures are designed for $\mathbf{X}$ with normalized columns. If fixed-X Knockoff procedures are applied in practice, $\mathbf{X}$ has to be normalized first.
e) Since $\mathbf{X}^T \cdot \mathbf{X}$ is assumed to be invertible, the normalized version is again invertible.
f) We will treat $\tilde{\mathbf{X}}$ and therefore $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]$ as non-random (fixed).
g) Construction of FX-Knockoff matrices in practice is presented in subsections 2.4.1 and 2.4.2.

Definition 2.3. Let $\text{swap}(S) : \mathbb{R}^{n \times 2 \cdot p} \to \mathbb{R}^{n \times 2 \cdot p}$ be a function where the matrix $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}$ is obtained from $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]$ by swapping columns $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ for each $j \in S$.

Comments:
Illustration of the swap-function for $p = 3$ and $S = \{2, 3\}$:

$$\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)} = \left[\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, \tilde{\mathbf{X}}_3\right]_{\text{swap}(\{2,3\})} = \left[\mathbf{X}_1, \tilde{\mathbf{X}}_2, \tilde{\mathbf{X}}_3, \tilde{\mathbf{X}}_1, \mathbf{X}_2, \mathbf{X}_3\right].$$

Next, we introduce statistics which will help to tease apart important (non-null) and unimportant (null) variables in the variable selection process.

Definition 2.4. For each $j \in \{1, \ldots, p\}$, statistics $W_j = w_j\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right], \mathbf{Y}\right)$ are said to be *FX-Knockoff feature statistics* for some real valued measurable function w_j which satisfies:
i) *("flip-sign property")*: For any $S \subseteq \{1, \ldots, p\}$,

$$w_j\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}, \mathbf{Y}\right) = \begin{cases} w_j\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right], \mathbf{Y}\right), & j \notin S, \\ -w_j\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right], \mathbf{Y}\right), & j \in S. \end{cases} \tag{2}$$

ii) *("sufficiency property")*: W_j depends upon the design matrix $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]$ only through $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right]$, and on the response $\mathbf{Y}$ only through $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \mathbf{Y}$, i.e.,

$$W_j = w_j\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right], \left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \mathbf{Y}\right). \tag{3}$$

Comments:
a) The flip-sign property says that swapping $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ has the effect of switching the sign of W_j.
b) The sufficiency property implies we can write $\mathbf{W} = w\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right], \left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \mathbf{Y}\right)$ for the vector valued function $w(x) = (w_1(x), \ldots, w_p(x))$.
c) FX-Knockoff feature statistics play a similar role as p-values in p-value-based variable selection procedures. Our methods (see Definitions 2.5 and 2.6) will be defined for a setting where the FX-Knockoff feature statistics are designed such that high positive values of W_j give evidence that j is non-null. This requirement is not part of the definition of FX-Knockoff feature statistics, because violation would just lead to poor power. The analogous requirement in p-value-based methods would be "small p-values p_j give evidence that j is non-null".
d) Examples of FX-Knockoff feature statistics are given in subsection 2.4.3.

We now present two variable selection procedures for the fixed-X setting.

Definition 2.5. (Definition 1 in Barber and Candes (2015)). Let $\tilde{\mathbf{X}}$ be a FX-Knockoff matrix (see Definition 2.2), let $W_1, \ldots, W_p$ be FX-Knockoff feature statistics (see Definition 2.4). Let

$$\mathcal{W} = \{|W_j| : \ W_j \neq 0, \ j \in \{1, \ldots, p\}\}$$

be the set of unique nonzero values attained by the $|W_j|$'s. Let $q \in [0,1]$ be the target FDR level and let

$$T = \min\left\{ t \in \mathcal{W} : \ \frac{|\{j : \ W_j \leq -t\}|}{|\{j : \ W_j \geq t\}| \vee 1} \leq q \right\}$$

(or $T = +\infty$ if this set is empty) be the data-dependent threshold. Then the *FX-Knockoff method* selects variables with indices

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

Comments:
i) Note that T and therefore $\hat{S}$ both implicitly depend on the choice of the target FDR level q.
ii) A hat is used to emphasize that $\hat{S}$ is random.
iii) We use the notation $a \vee b = \max\{a, b\}$. The maximum is used in the denominator to avoid dividing by zero, which would be the case when zero features get selected by the selection procedure.
iv) As we will see later on, the FX-Knockoff method only controls a modified FDR. FDR-control can be achieved by simply adding $+1$ in the numerator of the threshold-inequality:

Definition 2.6. Let $\tilde{\mathbf{X}}$ be a FX-Knockoff matrix (see Definition 2.2), let $W_1, \ldots, W_p$ be FX-Knockoff feature statistics (see Definition 2.4). Let

$$\mathcal{W} = \{|W_j| : \ W_j \neq 0, \ j \in \{1, \ldots, p\}\}$$

be the set of unique nonzero values attained by the $|W_j|$'s. Let $q \in [0,1]$ be the target FDR level and let

$$T_+ = \min\left\{ t \in \mathcal{W} : \ \frac{1 + |\{j : \ W_j \leq -t\}|}{|\{j : \ W_j \geq t\}| \vee 1} \leq q \right\}$$

(or $T_+ = +\infty$ if this set is empty) be the data-dependent threshold. Then the *FX-Knockoff+ method* selects variables with indices

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

2.2 FDR-Control

Definition 2.7. In the FX-Knockoff setting where we have a selection procedure returning a subset $\hat{S} \subseteq \{1, \ldots, p\}$ of indices of variables,
i) the *false discovery proportion* (FDP) is given by

$$\text{FDP} = \frac{|\left\{ j : \ \beta_j = 0 \text{ and } j \in \hat{S} \right\}|}{|\hat{S}| \vee 1}.$$

ii) the *false discovery rate* (FDR) is given by

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

Comments:
The false discovery proportion is defined as the number of falsely selected indices divided by the number of selected indices. The definition of nulls (Definition 2.1) explicitly tells us what "falsely" means in our setting, namely indices j with corresponding coefficient $\beta_j = 0$.

To prove that the FX-Knockoff method controls a modified version of FDR and that the FX-Knockoff+ method controls FDR, we first present Lemmas which will be useful. The first Lemma provides simple properties of FX-Knockoff matrices and vectors:

Lemma 2.8. *Let $\tilde{\mathbf{X}}$ be a FX-Knockoff matrix (containing FX-Knockoff vectors $\tilde{\mathbf{X}}_j$ as columns). Then:*
i) $\mathbf{X}_j^T \cdot \mathbf{X}_k = \tilde{\mathbf{X}}_j^T \cdot \tilde{\mathbf{X}}_k$ *for all* j, k,
ii) $\mathbf{X}_j^T \cdot \tilde{\mathbf{X}}_k = \mathbf{X}_j^T \cdot \mathbf{X}_k$ *for all* $j \neq k$,
iii) $\tilde{\mathbf{X}}_j^T \cdot \mathbf{X}_k = \mathbf{X}_j^T \cdot \mathbf{X}_k$ *for all* $j \neq k$,
iv) $\mathbf{X}_j^T \cdot \tilde{\mathbf{X}}_k = \tilde{\mathbf{X}}_j^T \cdot \tilde{\mathbf{X}}_k$ *for all* $j \neq k$,
v) $\tilde{\mathbf{X}}_j^T \cdot \mathbf{X}_k = \tilde{\mathbf{X}}_j^T \cdot \tilde{\mathbf{X}}_k$ *for all* $j \neq k$,
vi) $\tilde{\mathbf{X}}_j^T \cdot \mathbf{X}_k = \mathbf{X}_j^T \cdot \tilde{\mathbf{X}}_k$ *for all* $j \neq k$,
vii) $\mathbf{X}_j^T \cdot \mathbf{X}_j = \tilde{\mathbf{X}}_j^T \cdot \tilde{\mathbf{X}}_j = 1$,
viii) $\tilde{\mathbf{X}}^T \cdot \mathbf{X} = \tilde{\mathbf{X}}^T \cdot \tilde{\mathbf{X}} - \text{diag}(\boldsymbol{s})$,
ix) $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right] = \begin{bmatrix} \mathbf{X}^T \cdot \mathbf{X} & \mathbf{X}^T \cdot \tilde{\mathbf{X}} \\ \tilde{\mathbf{X}}^T \cdot \mathbf{X} & \tilde{\mathbf{X}}^T \cdot \tilde{\mathbf{X}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s}) \\ \boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s}) & \boldsymbol{\Sigma} \end{bmatrix}$, *where* $\boldsymbol{\Sigma} = \mathbf{X}^T \cdot \mathbf{X}$.

Proof:

ad i) wts $\mathbf{X}_j^T \cdot \mathbf{X}_k = \tilde{\mathbf{X}}_j^T \cdot \tilde{\mathbf{X}}_k$ for all j, k:

By Definition 2.2 i), we know that $\mathbf{X}^T \cdot \mathbf{X} = \tilde{\mathbf{X}}^T \cdot \tilde{\mathbf{X}}$.
In other words,

$$\begin{bmatrix} \mathbf{X}_1 & \cdots & \mathbf{X}_p \end{bmatrix}^T \cdot \begin{bmatrix} \mathbf{X}_1 & \cdots & \mathbf{X}_p \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{X}}_1 & \cdots & \tilde{\mathbf{X}}_p \end{bmatrix}^T \cdot \begin{bmatrix} \tilde{\mathbf{X}}_1 & \cdots & \tilde{\mathbf{X}}_p \end{bmatrix}.$$

Hence,

$$\begin{bmatrix} \mathbf{X}_1^T \\ \vdots \\ \mathbf{X}_p^T \end{bmatrix} \cdot \begin{bmatrix} \mathbf{X}_1 & \cdots & \mathbf{X}_p \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{X}}_1^T \\ \vdots \\ \tilde{\mathbf{X}}_p^T \end{bmatrix} \cdot \begin{bmatrix} \tilde{\mathbf{X}}_1 & \cdots & \tilde{\mathbf{X}}_p \end{bmatrix}.$$

So,

$$\begin{bmatrix} \mathbf{X}_1^T \cdot \mathbf{X}_1 & \cdots & \mathbf{X}_1^T \cdot \mathbf{X}_p \\ \vdots & \ddots & \vdots \\ \mathbf{X}_p^T \cdot \mathbf{X}_1 & \cdots & \mathbf{X}_p^T \cdot \mathbf{X}_p \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{X}}_1^T \cdot \tilde{\mathbf{X}}_1 & \cdots & \tilde{\mathbf{X}}_1^T \cdot \tilde{\mathbf{X}}_p \\ \vdots & \ddots & \vdots \\ \tilde{\mathbf{X}}_p^T \cdot \tilde{\mathbf{X}}_1 & \cdots & \tilde{\mathbf{X}}_p^T \cdot \tilde{\mathbf{X}}_p \end{bmatrix}.$$

And we see that $\mathbf{X}_j^T \cdot \mathbf{X}_k = \tilde{\mathbf{X}}_j^T \cdot \tilde{\mathbf{X}}_k$ for all j, k.

ad ii) wts $\mathbf{X}_j^T \cdot \tilde{\mathbf{X}}_k = \mathbf{X}_j^T \cdot \mathbf{X}_k$ for all $j \neq k$:

Prep2.1:

$$\begin{aligned}
&\mathbf{X}^T \cdot \tilde{\mathbf{X}} \\
&= \begin{bmatrix} \mathbf{X}_1 & \cdots & \mathbf{X}_p \end{bmatrix}^T \cdot \begin{bmatrix} \tilde{\mathbf{X}}_1 & \cdots & \tilde{\mathbf{X}}_p \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{X}_1^T \\ \vdots \\ \mathbf{X}_p^T \end{bmatrix} \cdot \begin{bmatrix} \tilde{\mathbf{X}}_1 & \cdots & \tilde{\mathbf{X}}_p \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{X}_1^T \cdot \tilde{\mathbf{X}}_1 & \cdots & \mathbf{X}_1^T \cdot \tilde{\mathbf{X}}_p \\ \vdots & \ddots & \vdots \\ \mathbf{X}_p^T \cdot \tilde{\mathbf{X}}_1 & \cdots & \mathbf{X}_p^T \cdot \tilde{\mathbf{X}}_p \end{bmatrix}.
\end{aligned}$$

Prep2.2:

$$\begin{aligned}
&\mathbf{X}^T \cdot \mathbf{X} - \operatorname{diag}(\boldsymbol{s}) \\
&= \begin{bmatrix} \mathbf{X}_1^T \cdot \mathbf{X}_1 & \cdots & \mathbf{X}_1^T \cdot \mathbf{X}_p \\ \vdots & \ddots & \vdots \\ \mathbf{X}_p^T \cdot \mathbf{X}_1 & \cdots & \mathbf{X}_p^T \cdot \mathbf{X}_p \end{bmatrix} - \begin{bmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & s_p \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{X}_1^T \cdot \mathbf{X}_1 - s_1 & \mathbf{X}_1^T \cdot \mathbf{X}_2 & \cdots & \mathbf{X}_1^T \cdot \mathbf{X}_p \\ \mathbf{X}_2^T \cdot \mathbf{X}_1 & \mathbf{X}_2^T \cdot \mathbf{X}_2 - s_2 & & \vdots \\ \vdots & & \ddots & \mathbf{X}_{p-1}^T \cdot \mathbf{X}_p \\ \mathbf{X}_p^T \cdot \mathbf{X}_1 & \cdots & \mathbf{X}_p^T \cdot \mathbf{X}_{p-1} & \mathbf{X}_p^T \cdot \mathbf{X}_p - s_p \end{bmatrix}.
\end{aligned}$$

By Definition 2.2 ii) it holds that $\mathbf{X}^T \cdot \tilde{\mathbf{X}} = \mathbf{X}^T \cdot \mathbf{X} - \operatorname{diag}(\boldsymbol{s})$, so Prep2.1 and Prep2.2 give

$$\begin{bmatrix} \mathbf{X}_1^T \cdot \tilde{\mathbf{X}}_1 & \cdots & \mathbf{X}_1^T \cdot \tilde{\mathbf{X}}_p \\ \vdots & \ddots & \vdots \\ \mathbf{X}_p^T \cdot \tilde{\mathbf{X}}_1 & \cdots & \mathbf{X}_p^T \cdot \tilde{\mathbf{X}}_p \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^T \cdot \mathbf{X}_1 - s_1 & \mathbf{X}_1^T \cdot \mathbf{X}_2 & \cdots & \mathbf{X}_1^T \cdot \mathbf{X}_p \\ \mathbf{X}_2^T \cdot \mathbf{X}_1 & \mathbf{X}_2^T \cdot \mathbf{X}_2 - s_2 & & \vdots \\ \vdots & & \ddots & \mathbf{X}_{p-1}^T \cdot \mathbf{X}_p \\ \mathbf{X}_p^T \cdot \mathbf{X}_1 & \cdots & \mathbf{X}_p^T \cdot \mathbf{X}_{p-1} & \mathbf{X}_p^T \cdot \mathbf{X}_p - s_p \end{bmatrix}.$$

Therefore, we can see that $\mathbf{X}_j^T \cdot \tilde{\mathbf{X}}_k = \mathbf{X}_j^T \cdot \mathbf{X}_k$ for all $j \neq k$.

ad iii) wts $\tilde{\mathbf{X}}_j^T \cdot \mathbf{X}_k = \mathbf{X}_j^T \cdot \mathbf{X}_k$ for all $j \neq k$:

Transposing both sides of ii) proves the claim.

ad iv) wts $\mathbf{X}_j^T \cdot \tilde{\mathbf{X}}_k = \tilde{\mathbf{X}}_j^T \cdot \tilde{\mathbf{X}}_k$ for all $j \neq k$:

Prep4.1:

$$\mathbf{X}^T \cdot \tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X}_1^T \cdot \tilde{\mathbf{X}}_1 & \cdots & \mathbf{X}_1^T \cdot \tilde{\mathbf{X}}_p \\ \vdots & \ddots & \vdots \\ \mathbf{X}_p^T \cdot \tilde{\mathbf{X}}_1 & \cdots & \mathbf{X}_p^T \cdot \tilde{\mathbf{X}}_p \end{bmatrix}.$$

Prep4.2:

$$\begin{aligned} &\mathbf{X}^T \cdot \mathbf{X} - \operatorname{diag}(\boldsymbol{s}) \\ &\qquad\qquad \text{By Definition 2.2 i)} \\ &= \tilde{\mathbf{X}}^T \cdot \tilde{\mathbf{X}} - \operatorname{diag}(\boldsymbol{s}) \\ &= \begin{bmatrix} \tilde{\mathbf{X}}_1^T \cdot \tilde{\mathbf{X}}_1 - s_1 & \tilde{\mathbf{X}}_1^T \cdot \tilde{\mathbf{X}}_2 & \cdots & \tilde{\mathbf{X}}_1^T \cdot \tilde{\mathbf{X}}_p \\ \tilde{\mathbf{X}}_2^T \cdot \tilde{\mathbf{X}}_1 & \tilde{\mathbf{X}}_2^T \cdot \tilde{\mathbf{X}}_2 - s_2 & & \vdots \\ \vdots & & \ddots & \tilde{\mathbf{X}}_{p-1}^T \cdot \tilde{\mathbf{X}}_p \\ \tilde{\mathbf{X}}_p^T \cdot \tilde{\mathbf{X}}_1 & \cdots & \tilde{\mathbf{X}}_p^T \cdot \tilde{\mathbf{X}}_{p-1} & \tilde{\mathbf{X}}_p^T \cdot \tilde{\mathbf{X}}_p - s_p \end{bmatrix}. \end{aligned}$$

By Definition 2.2 ii) it holds that $\mathbf{X}^T \cdot \tilde{\mathbf{X}} = \mathbf{X}^T \cdot \mathbf{X} - \operatorname{diag}(\boldsymbol{s})$, so Prep4.1 and Prep4.2 give

$$\begin{bmatrix} \mathbf{X}_1^T \cdot \tilde{\mathbf{X}}_1 & \cdots & \mathbf{X}_1^T \cdot \tilde{\mathbf{X}}_p \\ \vdots & \ddots & \vdots \\ \mathbf{X}_p^T \cdot \tilde{\mathbf{X}}_1 & \cdots & \mathbf{X}_p^T \cdot \tilde{\mathbf{X}}_p \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{X}}_1^T \cdot \tilde{\mathbf{X}}_1 - s_1 & \tilde{\mathbf{X}}_1^T \cdot \tilde{\mathbf{X}}_2 & \cdots & \tilde{\mathbf{X}}_1^T \cdot \tilde{\mathbf{X}}_p \\ \tilde{\mathbf{X}}_2^T \cdot \tilde{\mathbf{X}}_1 & \tilde{\mathbf{X}}_2^T \cdot \tilde{\mathbf{X}}_2 - s_2 & & \vdots \\ \vdots & & \ddots & \tilde{\mathbf{X}}_{p-1}^T \cdot \tilde{\mathbf{X}}_p \\ \tilde{\mathbf{X}}_p^T \cdot \tilde{\mathbf{X}}_1 & \cdots & \tilde{\mathbf{X}}_p^T \cdot \tilde{\mathbf{X}}_{p-1} & \tilde{\mathbf{X}}_p^T \cdot \tilde{\mathbf{X}}_p - s_p \end{bmatrix}.$$

Therefore, we have $\mathbf{X}_j^T \cdot \tilde{\mathbf{X}}_k = \tilde{\mathbf{X}}_j^T \cdot \tilde{\mathbf{X}}_k$ for all $j \neq k$.

ad v) wts $\tilde{\mathbf{X}}_j^T \cdot \mathbf{X}_k = \tilde{\mathbf{X}}_j^T \cdot \tilde{\mathbf{X}}_k$ for all $j \neq k$:

Transposing both sides of iv) proves the claim.

ad vi) wts $\tilde{\mathbf{X}}_j^T \cdot \mathbf{X}_k = \mathbf{X}_j^T \cdot \tilde{\mathbf{X}}_k$ for all $j \neq k$:

Let $j \neq k$.

$$\begin{aligned} &\tilde{\mathbf{X}}_j^T \cdot \mathbf{X}_k \\ &\qquad\qquad \text{By iii)} \\ &= \mathbf{X}_j^T \cdot \mathbf{X}_k \\ &\qquad\qquad \text{By ii)} \\ &= \mathbf{X}_j^T \cdot \tilde{\mathbf{X}}_k. \end{aligned}$$

ad vii) wts $\mathbf{X}_j^T \cdot \mathbf{X}_j = \tilde{\mathbf{X}}_j^T \cdot \tilde{\mathbf{X}}_j = 1$:

Note that

$$
\begin{aligned}
&1 \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{By Definition 2.2} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{it holds that } ||\mathbf{X}_j||_2^2 = 1 \text{ for all } j \in \{1, \ldots, p\}. \\
&= ||\mathbf{X}_j||_2^2 \\
&= \mathbf{X}_j^T \cdot \mathbf{X}_j.
\end{aligned}
$$

And by i), we know that $\mathbf{X}_j^T \cdot \mathbf{X}_j = \tilde{\mathbf{X}}_j^T \cdot \tilde{\mathbf{X}}_j$ for all j.
Hence, we proved that $\mathbf{X}_j^T \cdot \mathbf{X}_j = \tilde{\mathbf{X}}_j^T \cdot \tilde{\mathbf{X}}_j = 1$.

ad viii) wts $\tilde{\mathbf{X}}^T \cdot \mathbf{X} = \tilde{\mathbf{X}}^T \cdot \tilde{\mathbf{X}} - \text{diag}(\boldsymbol{s})$:

Definition 2.2 i) and ii) together gives $\mathbf{X}^T \cdot \tilde{\mathbf{X}} = \tilde{\mathbf{X}}^T \cdot \tilde{\mathbf{X}} - \text{diag}(\boldsymbol{s})$.
Both sides transposed gives $\tilde{\mathbf{X}}^T \cdot \mathbf{X} = \tilde{\mathbf{X}}^T \cdot \tilde{\mathbf{X}} - [\text{diag}(\boldsymbol{s})]^T = \tilde{\mathbf{X}}^T \cdot \tilde{\mathbf{X}} - \text{diag}(\boldsymbol{s})$.

ad ix) wts $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right] = \begin{bmatrix} \mathbf{X}^T \cdot \mathbf{X} & \mathbf{X}^T \cdot \tilde{\mathbf{X}} \\ \tilde{\mathbf{X}}^T \cdot \mathbf{X} & \tilde{\mathbf{X}}^T \cdot \tilde{\mathbf{X}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s}) \\ \boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s}) & \boldsymbol{\Sigma} \end{bmatrix}$ where $\boldsymbol{\Sigma} = \mathbf{X}^T \cdot \mathbf{X}$:

$$
\begin{aligned}
&\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right] \\
&= \begin{bmatrix} \mathbf{X}^T \\ \tilde{\mathbf{X}}^T \end{bmatrix} \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right] \\
&= \begin{bmatrix} \mathbf{X}^T \cdot \mathbf{X} & \mathbf{X}^T \cdot \tilde{\mathbf{X}} \\ \tilde{\mathbf{X}}^T \cdot \mathbf{X} & \tilde{\mathbf{X}}^T \cdot \tilde{\mathbf{X}} \end{bmatrix} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{By Definition 2.2 i), } \tilde{\mathbf{X}}^T \cdot \tilde{\mathbf{X}} = \boldsymbol{\Sigma}. \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{By Definition 2.2 ii), } \mathbf{X}^T \cdot \tilde{\mathbf{X}} = \boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s}). \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{By viii) and Definition 2.2 i), } \tilde{\mathbf{X}}^T \cdot \mathbf{X} = \boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s}). \\
&= \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s}) \\ \boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s}) & \boldsymbol{\Sigma} \end{bmatrix}.
\end{aligned}
$$

□

The next two Lemmas are needed in the proof of Lemma 2.11.

Lemma 2.9. *(Lemma 2 in Barber and Candes (2015)). For any subset $S \subseteq \{1, \ldots, p\}$,*

$$\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}\right)^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)} = \left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right].$$

That is, the Gram matrix of $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]$ is unchanged when we swap $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ for each $j \in S$.

Proof:

$$
\begin{aligned}
&\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\mathrm{swap}(S)}\right)^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\mathrm{swap}(S)} \\
&= \left(\left[\begin{array}{cccccccc} \mathbf{X}_1 & \cdots & \mathbf{X}_p & \tilde{\mathbf{X}}_1 & \cdots & \tilde{\mathbf{X}}_p \end{array}\right]_{\mathrm{swap}(S)}\right)^T \cdot \left[\begin{array}{cccccccc} \mathbf{X}_1 & \cdots & \mathbf{X}_p & \tilde{\mathbf{X}}_1 & \cdots & \tilde{\mathbf{X}}_p \end{array}\right]_{\mathrm{swap}(S)} \\
&\qquad\qquad \text{Define } \mathbf{X}'_i = \begin{cases} \tilde{\mathbf{X}}_i, & i \in S, \\ \mathbf{X}_i, & i \notin S, \end{cases} \\
&\qquad\qquad \text{and } \tilde{\mathbf{X}}'_i = \begin{cases} \mathbf{X}_i, & i \in S, \\ \tilde{\mathbf{X}}_i, & i \notin S. \end{cases} \\
&= \left[\begin{array}{cccccccc} \mathbf{X}'_1 & \mathbf{X}'_2 & \cdots & \mathbf{X}'_p & \tilde{\mathbf{X}}'_1 & \tilde{\mathbf{X}}'_2 & \cdots & \tilde{\mathbf{X}}'_p \end{array}\right]^T \cdot \left[\begin{array}{cccccccc} \mathbf{X}'_1 & \mathbf{X}'_2 & \cdots & \mathbf{X}'_p & \tilde{\mathbf{X}}'_1 & \tilde{\mathbf{X}}'_2 & \cdots & \tilde{\mathbf{X}}'_p \end{array}\right] \\
&= \begin{bmatrix} (\mathbf{X}'_1)^T \\ (\mathbf{X}'_2)^T \\ \vdots \\ (\mathbf{X}'_p)^T \\ \left(\tilde{\mathbf{X}}'_1\right)^T \\ \left(\tilde{\mathbf{X}}'_2\right)^T \\ \vdots \\ \left(\tilde{\mathbf{X}}'_p\right)^T \end{bmatrix} \cdot \left[\begin{array}{cccccccc} \mathbf{X}'_1 & \mathbf{X}'_2 & \cdots & \mathbf{X}'_p & \tilde{\mathbf{X}}'_1 & \tilde{\mathbf{X}}'_2 & \cdots & \tilde{\mathbf{X}}'_p \end{array}\right] \\
&= \left[\begin{array}{cccc|cccc} (\mathbf{X}'_1)^T \cdot \mathbf{X}'_1 & (\mathbf{X}'_1)^T \cdot \mathbf{X}'_2 & \cdots & (\mathbf{X}'_1)^T \cdot \mathbf{X}'_p & (\mathbf{X}'_1)^T \cdot \tilde{\mathbf{X}}'_1 & (\mathbf{X}'_1)^T \cdot \tilde{\mathbf{X}}'_2 & \cdots & (\mathbf{X}'_1)^T \cdot \tilde{\mathbf{X}}'_p \\ (\mathbf{X}'_2)^T \cdot \mathbf{X}'_1 & (\mathbf{X}'_2)^T \cdot \mathbf{X}'_2 & \cdots & (\mathbf{X}'_2)^T \cdot \mathbf{X}'_p & (\mathbf{X}'_2)^T \cdot \tilde{\mathbf{X}}'_1 & (\mathbf{X}'_2)^T \cdot \tilde{\mathbf{X}}'_2 & \cdots & (\mathbf{X}'_2)^T \cdot \tilde{\mathbf{X}}'_p \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ (\mathbf{X}'_p)^T \cdot \mathbf{X}'_1 & (\mathbf{X}'_p)^T \cdot \mathbf{X}'_2 & \cdots & (\mathbf{X}'_p)^T \cdot \mathbf{X}'_p & (\mathbf{X}'_p)^T \cdot \tilde{\mathbf{X}}'_1 & (\mathbf{X}'_p)^T \cdot \tilde{\mathbf{X}}'_2 & \cdots & (\mathbf{X}'_p)^T \cdot \tilde{\mathbf{X}}'_p \\ \hline \left(\tilde{\mathbf{X}}'_1\right)^T \cdot \mathbf{X}'_1 & \left(\tilde{\mathbf{X}}'_1\right)^T \cdot \mathbf{X}'_2 & \cdots & \left(\tilde{\mathbf{X}}'_1\right)^T \cdot \mathbf{X}'_p & \left(\tilde{\mathbf{X}}'_1\right)^T \cdot \tilde{\mathbf{X}}'_1 & \left(\tilde{\mathbf{X}}'_1\right)^T \cdot \tilde{\mathbf{X}}'_2 & \cdots & \left(\tilde{\mathbf{X}}'_1\right)^T \cdot \tilde{\mathbf{X}}'_p \\ \left(\tilde{\mathbf{X}}'_2\right)^T \cdot \mathbf{X}'_1 & \left(\tilde{\mathbf{X}}'_2\right)^T \cdot \mathbf{X}'_2 & \cdots & \left(\tilde{\mathbf{X}}'_2\right)^T \cdot \mathbf{X}'_p & \left(\tilde{\mathbf{X}}'_2\right)^T \cdot \tilde{\mathbf{X}}'_1 & \left(\tilde{\mathbf{X}}'_2\right)^T \cdot \tilde{\mathbf{X}}'_2 & \cdots & \left(\tilde{\mathbf{X}}'_2\right)^T \cdot \tilde{\mathbf{X}}'_p \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \left(\tilde{\mathbf{X}}'_p\right)^T \cdot \mathbf{X}'_1 & \left(\tilde{\mathbf{X}}'_p\right)^T \cdot \mathbf{X}'_2 & \cdots & \left(\tilde{\mathbf{X}}'_p\right)^T \cdot \mathbf{X}'_p & \left(\tilde{\mathbf{X}}'_p\right)^T \cdot \tilde{\mathbf{X}}'_1 & \left(\tilde{\mathbf{X}}'_p\right)^T \cdot \tilde{\mathbf{X}}'_2 & \cdots & \left(\tilde{\mathbf{X}}'_p\right)^T \cdot \tilde{\mathbf{X}}'_p \end{array}\right] =: (*)
\end{aligned}
$$

By Lemma 2.8 ix), we know that $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right] = \begin{bmatrix} \mathbf{X}^T \cdot \mathbf{X} & \mathbf{X}^T \cdot \tilde{\mathbf{X}} \\ \tilde{\mathbf{X}}^T \cdot \mathbf{X} & \tilde{\mathbf{X}}^T \cdot \tilde{\mathbf{X}} \end{bmatrix}$.

Therefore, it remains to show that $(*) = \begin{bmatrix} \mathbf{X}^T \cdot \mathbf{X} & \mathbf{X}^T \cdot \tilde{\mathbf{X}} \\ \tilde{\mathbf{X}}^T \cdot \mathbf{X} & \tilde{\mathbf{X}}^T \cdot \tilde{\mathbf{X}} \end{bmatrix}$.

ad diagonal of upper left block and lower right block:

> No matter what indices are in S, we will always end up with pairs $\mathbf{X}_i^T \cdot \mathbf{X}_i$ or $\tilde{\mathbf{X}}_i^T \cdot \tilde{\mathbf{X}}_i$.
> And by Lemma 2.8 i), $\tilde{\mathbf{X}}_i^T \cdot \tilde{\mathbf{X}}_i = \mathbf{X}_i^T \cdot \mathbf{X}_i$ for all i.

ad diagonal of lower left and upper right block:

> In the lower left block, we need $\tilde{\mathbf{X}}_i^T \cdot \mathbf{X}_i$. For $i \notin S$, we have that already. For $i \in S$, we have $\mathbf{X}_i^T \cdot \tilde{\mathbf{X}}_i$, but as it is 1×1, we can transpose it and get $\tilde{\mathbf{X}}_i^T \cdot \mathbf{X}_i$.
> In the upper right block, we need $\mathbf{X}_i^T \cdot \tilde{\mathbf{X}}_i$. It is already the case for $i \notin S$. For $i \in S$, we can change $\tilde{\mathbf{X}}_i^T \cdot \mathbf{X}_i$ to $\mathbf{X}_i^T \cdot \tilde{\mathbf{X}}_i$ by transposing.

ad off-diagonal of the four blocks:

In the upper left block, we need $\mathbf{X}_i^T \cdot \mathbf{X}_j$. For $i, j \notin S$, we have that already.
Case1 $i, j \in S$: we can use Lemma 2.8 i) to get $\tilde{\mathbf{X}}_i^T \cdot \tilde{\mathbf{X}}_j = \mathbf{X}_i^T \cdot \mathbf{X}_j$.
Case2 $i \in S$, $j \notin S$: we have $\tilde{\mathbf{X}}_i^T \cdot \mathbf{X}_j$. Using Lemma 2.8 iii), we get $\tilde{\mathbf{X}}_i^T \cdot \mathbf{X}_j = \mathbf{X}_i^T \cdot \mathbf{X}_j$.
Case3 $i \notin S$, $j \in S$: we have $\mathbf{X}_i^T \cdot \tilde{\mathbf{X}}_j$ and by using Lemma 2.8 ii), we get $\mathbf{X}_i^T \cdot \tilde{\mathbf{X}}_j = \mathbf{X}_i^T \cdot \mathbf{X}_j$.

In the lower right block, we need $\tilde{\mathbf{X}}_i^T \cdot \tilde{\mathbf{X}}_j$. For $i, j \notin S$, we have that already. And for the other cases, we can use similar arguments as for the upper left block:
Case1 $i, j \in S$: we can use Lemma 2.8 i) to get $\mathbf{X}_i^T \cdot \mathbf{X}_j = \tilde{\mathbf{X}}_i^T \cdot \tilde{\mathbf{X}}_j$.
Case2 $i \in S$, $j \notin S$: we have $\mathbf{X}_i^T \cdot \tilde{\mathbf{X}}_j$ and by using Lemma 2.8 iv), we get $\mathbf{X}_i^T \cdot \tilde{\mathbf{X}}_j = \tilde{\mathbf{X}}_i^T \cdot \tilde{\mathbf{X}}_j$.
Case3 $i \notin S$, $j \in S$: we have $\tilde{\mathbf{X}}_i^T \cdot \mathbf{X}_j$. By Lemma 2.8 v), we get $\tilde{\mathbf{X}}_i^T \cdot \mathbf{X}_j = \tilde{\mathbf{X}}_i^T \cdot \tilde{\mathbf{X}}_j$.

In the lower left block, we need $\tilde{\mathbf{X}}_i^T \cdot \mathbf{X}_j$. For $i, j \notin S$, we have that already.
Case1 $i, j \in S$: we have $\mathbf{X}_i^T \cdot \tilde{\mathbf{X}}_j$ and can use Lemma 2.8 vi) to get $\tilde{\mathbf{X}}_i^{\ T} \cdot \mathbf{X}_j$.
Case2 $i \in S$, $j \notin S$: we have $\mathbf{X}_i^T \cdot \mathbf{X}_j$ and can use Lemma 2.8 iii) to get $\tilde{\mathbf{X}}_i^{\ T} \cdot \mathbf{X}_j$.
Case3 $i \notin S$, $j \in S$: we have $\tilde{\mathbf{X}}_i^T \cdot \tilde{\mathbf{X}}_j$ and can use Lemma 2.8 v) to get $\tilde{\mathbf{X}}_i^{\ T} \cdot \mathbf{X}_j$.

In the upper right block, we need $\mathbf{X}_i^T \cdot \tilde{\mathbf{X}}_j$. For $i, j \notin S$, we have that already.
Case1 $i, j \in S$: we have $\tilde{\mathbf{X}}_i^T \cdot \mathbf{X}_j$ and can use Lemma 2.8 vi) to get $\mathbf{X}_i^T \cdot \tilde{\mathbf{X}}_j$.
Case2 $i \in S$, $j \notin S$: we have $\tilde{\mathbf{X}}_i^T \cdot \tilde{\mathbf{X}}_j$ and can use Lemma 2.8 iv) to get $\mathbf{X}_i^T \cdot \tilde{\mathbf{X}}_j$.
Case3 $i \notin S$, $j \in S$: we have $\mathbf{X}_i^T \cdot \mathbf{X}_j$ and can use Lemma 2.8 ii) to get $\mathbf{X}_i^T \cdot \tilde{\mathbf{X}}_j$.

Therefore, we get

$$(*) = \begin{bmatrix} \mathbf{X}_1^T \cdot \mathbf{X}_1 & \mathbf{X}_1^T \cdot \mathbf{X}_2 & \cdots & \mathbf{X}_1^T \cdot \mathbf{X}_p & \mathbf{X}_1^T \cdot \tilde{\mathbf{X}}_1 & \mathbf{X}_1^T \cdot \tilde{\mathbf{X}}_2 & \cdots & \mathbf{X}_1^T \cdot \tilde{\mathbf{X}}_p \\ \mathbf{X}_2^T \cdot \mathbf{X}_1 & \mathbf{X}_2^T \cdot \mathbf{X}_2 & \cdots & \mathbf{X}_2^T \cdot \mathbf{X}_p & \mathbf{X}_2^T \cdot \tilde{\mathbf{X}}_1 & \mathbf{X}_2^T \cdot \tilde{\mathbf{X}}_2 & \cdots & \mathbf{X}_2^T \cdot \tilde{\mathbf{X}}_p \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}_p^T \cdot \mathbf{X}_1 & \mathbf{X}_p^T \cdot \mathbf{X}_2 & \cdots & \mathbf{X}_p^T \cdot \mathbf{X}_p & \mathbf{X}_p^T \cdot \tilde{\mathbf{X}}_1 & \mathbf{X}_p^T \cdot \tilde{\mathbf{X}}_2 & \cdots & \mathbf{X}_p^T \cdot \tilde{\mathbf{X}}_p \\ \tilde{\mathbf{X}}_1^T \cdot \mathbf{X}_1 & \tilde{\mathbf{X}}_1^T \cdot \mathbf{X}_2 & \cdots & \tilde{\mathbf{X}}_1^T \cdot \mathbf{X}_p & \tilde{\mathbf{X}}_1^T \cdot \tilde{\mathbf{X}}_1 & \tilde{\mathbf{X}}_1^T \cdot \tilde{\mathbf{X}}_2 & \cdots & \tilde{\mathbf{X}}_1^T \cdot \tilde{\mathbf{X}}_p \\ \tilde{\mathbf{X}}_2^T \cdot \mathbf{X}_1 & \tilde{\mathbf{X}}_2^T \cdot \mathbf{X}_2 & \cdots & \tilde{\mathbf{X}}_2^T \cdot \mathbf{X}_p & \tilde{\mathbf{X}}_2^T \cdot \tilde{\mathbf{X}}_1 & \tilde{\mathbf{X}}_2^T \cdot \tilde{\mathbf{X}}_2 & \cdots & \tilde{\mathbf{X}}_2^T \cdot \tilde{\mathbf{X}}_p \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{X}}_p^T \cdot \mathbf{X}_1 & \tilde{\mathbf{X}}_p^T \cdot \mathbf{X}_2 & \cdots & \tilde{\mathbf{X}}_p^T \cdot \mathbf{X}_p & \tilde{\mathbf{X}}_p^T \cdot \tilde{\mathbf{X}}_1 & \tilde{\mathbf{X}}_p^T \cdot \tilde{\mathbf{X}}_2 & \cdots & \tilde{\mathbf{X}}_p^T \cdot \tilde{\mathbf{X}}_p \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T \cdot \mathbf{X} & \mathbf{X}^T \cdot \tilde{\mathbf{X}} \\ \tilde{\mathbf{X}}^T \cdot \mathbf{X} & \tilde{\mathbf{X}}^T \cdot \tilde{\mathbf{X}} \end{bmatrix}.$$

□

Lemma 2.10. *(Lemma 3 in Barber and Candes (2015)). For any subset S of nulls,*

$$\left(\left[\mathbf{X}, \tilde{\mathbf{X}} \right]_{\text{swap}(S)} \right)^T \cdot \mathbf{Y} \overset{d}{=} \left[\mathbf{X}, \tilde{\mathbf{X}} \right]^T \cdot \mathbf{Y}.$$

That is, the distribution of the product $\left[\mathbf{X}, \tilde{\mathbf{X}} \right]^T \cdot \mathbf{Y}$ is unchanged when we swap $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ for each $j \in S$, as long as none of the swapped features appear in the true model.

Proof:

We know that $\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{z}$, where $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \cdot \mathbf{I}_n)$.

By using Lemma A.2, we get

$$(\mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{z}) \sim \mathcal{N}\left(\mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{0},\, \sigma^2 \cdot \mathbf{I}_n\right) = \mathcal{N}\left(\mathbf{X} \cdot \boldsymbol{\beta},\, \sigma^2 \cdot \mathbf{I}_n\right).$$

Hence,

$$\mathbf{Y} \sim \mathcal{N}\left(\mathbf{X} \cdot \boldsymbol{\beta},\, \sigma^2 \cdot \mathbf{I}_n\right). \tag{4}$$

Let $S' \subseteq \{1, \ldots, p\}$ be arbitrary.

Note that $\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S')}\right)^T$ is a non-random $2 \cdot p \times n$ matrix. By using (4) and Lemma A.2, we get

$$\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S')}\right)^T \cdot \mathbf{Y} \;\sim\; \mathcal{N}\left(\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S')}\right)^T \cdot \mathbf{X} \cdot \boldsymbol{\beta},\, \sigma^2 \cdot \left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S')}\right)^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S')}\right). \tag{5}$$

Analogously, by again using (4) and Lemma A.2, we get

$$\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \mathbf{Y} \sim \mathcal{N}\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \mathbf{X} \cdot \boldsymbol{\beta},\, \sigma^2 \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right]\right). \tag{6}$$

Because S' was arbitrary, it holds for all $S' \subseteq \{1, \ldots, p\}$ that $\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S')}\right)^T \cdot \mathbf{Y}$ and $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \mathbf{Y}$ are both normally distributed.

To show that these two random vectors have the same distribution for any subset S of nulls, it remains to show that they both have the same expected value and the same covariance matrix for all S.

Case1 $S = \emptyset$:

$$\mathbb{E}\left[\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}\right)^T \cdot \mathbf{Y}\right] = \mathbb{E}\left[\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(\emptyset)}\right)^T \cdot \mathbf{Y}\right] = \mathbb{E}\left[\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \mathbf{Y}\right].$$

$$\mathbb{C}ov\left[\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}\right)^T \cdot \mathbf{Y}\right] = \mathbb{C}ov\left[\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(\emptyset)}\right)^T \cdot \mathbf{Y}\right] = \mathbb{C}ov\left[\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \mathbf{Y}\right].$$

Case2 $S \neq \emptyset$ is a subset of nulls:

Part1 wts $\mathbb{C}ov\left[\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}\right)^T \cdot \mathbf{Y}\right] = \mathbb{C}ov\left[\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \mathbf{Y}\right]$:

$$\mathbb{C}ov\left[\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}\right)^T \cdot \mathbf{Y}\right]$$

By (5)

$$= \sigma^2 \cdot \left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}\right)^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}$$

By Lemma 2.9

$$= \sigma^2 \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right]$$

By (6)

$$= \mathbb{C}ov\left[\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \mathbf{Y}\right].$$

Prep1 wts $\mathbf{X}_j^T \cdot \mathbf{X} \cdot \boldsymbol{\beta} = \tilde{\mathbf{X}}_j^T \cdot \mathbf{X} \cdot \boldsymbol{\beta} \quad \forall j \in S$:

We can write $\mathbf{X} \cdot \boldsymbol{\beta} = \sum_{i=1}^{p} \beta_i \cdot \mathbf{X}_i$.
We know that for nulls, the corresponding true β_j is zero.

So,

$$\mathbf{X} \cdot \boldsymbol{\beta} = \sum_{\substack{i=1 \\ i \notin S}}^{p} \beta_i \cdot \mathbf{X}_i. \tag{7}$$

Let $j \in S$ be arbitrary but fixed.

Equation (7) multiplied by $\mathbf{X}_j^T$ from the left side (on both sides) yields

$$\mathbf{X}_j^T \cdot \mathbf{X} \cdot \boldsymbol{\beta} = \mathbf{X}_j^T \cdot \sum_{\substack{i=1 \\ i \notin S}}^{p} \beta_i \cdot \mathbf{X}_i.$$

Multiplied out, the RHS is a sum of terms of the form $\beta_i \cdot \mathbf{X}_j^T \cdot \mathbf{X}_i$.
Note that (because of $j \in S$ and $i \notin S$) $i \neq j$.
By Lemma 2.8 iii) we can write it as a sum of terms of the form $\beta_i \cdot \tilde{\mathbf{X}}_j^T \cdot \mathbf{X}_i$.
Hence,

$$\mathbf{X}_j^T \cdot \mathbf{X} \cdot \boldsymbol{\beta} = \mathbf{X}_j^T \cdot \sum_{\substack{i=1 \\ i \notin S}}^{p} \beta_i \cdot \mathbf{X}_i = \tilde{\mathbf{X}}_j^T \cdot \sum_{\substack{i=1 \\ i \notin S}}^{p} \beta_i \cdot \mathbf{X}_i = \tilde{\mathbf{X}}_j^T \cdot \mathbf{X} \cdot \boldsymbol{\beta}.$$

Since $j \in S$ was arbitrary, we proved $\mathbf{X}_j^T \cdot \mathbf{X} \cdot \boldsymbol{\beta} = \tilde{\mathbf{X}}_j^T \cdot \mathbf{X} \cdot \boldsymbol{\beta}$ for all $j \in S$.

Part2 wts $\mathbb{E}\left[\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}\right)^T \cdot \mathbf{Y}\right] = \mathbb{E}\left[\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \mathbf{Y}\right]$:

$$\mathbb{E}\left[\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}\right)^T \cdot \mathbf{Y}\right]$$

By (5)

$$= \left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}\right)^T \cdot \mathbf{X} \cdot \boldsymbol{\beta}$$

$$= \left(\left[\mathbf{X}_1, \ldots, \mathbf{X}_p, \tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_p\right]_{\text{swap}(S)}\right)^T \cdot \mathbf{X} \cdot \boldsymbol{\beta}$$

Define $\mathbf{X}'_j = \begin{cases} \tilde{\mathbf{X}}_j, & j \in S \\ \mathbf{X}_j, & j \notin S \end{cases}$ and $\tilde{\mathbf{X}}'_j = \begin{cases} \mathbf{X}_j, & j \in S \\ \tilde{\mathbf{X}}_j, & j \notin S \end{cases}$.

$$= \left[\mathbf{X}'_1, \mathbf{X}'_2, \ldots, \mathbf{X}'_p, \tilde{\mathbf{X}}'_1, \tilde{\mathbf{X}}'_2, \ldots, \tilde{\mathbf{X}}'_p\right]^T \cdot \mathbf{X} \cdot \boldsymbol{\beta}$$

$$= \begin{bmatrix} (\mathbf{X}'_1)^T \\ (\mathbf{X}'_2)^T \\ \vdots \\ (\mathbf{X}'_p)^T \\ \left(\tilde{\mathbf{X}}'_1\right)^T \\ \left(\tilde{\mathbf{X}}'_2\right)^T \\ \vdots \\ \left(\tilde{\mathbf{X}}'_p\right)^T \end{bmatrix} \cdot \mathbf{X} \cdot \boldsymbol{\beta}$$

$$= \begin{bmatrix} (\mathbf{X}'_1)^T \cdot \mathbf{X} \cdot \boldsymbol{\beta} \\ (\mathbf{X}'_2)^T \cdot \mathbf{X} \cdot \boldsymbol{\beta} \\ \vdots \\ (\mathbf{X}'_p)^T \cdot \mathbf{X} \cdot \boldsymbol{\beta} \\ \left(\tilde{\mathbf{X}}'_1\right)^T \cdot \mathbf{X} \cdot \boldsymbol{\beta} \\ \left(\tilde{\mathbf{X}}'_2\right)^T \cdot \mathbf{X} \cdot \boldsymbol{\beta} \\ \vdots \\ \left(\tilde{\mathbf{X}}'_p\right)^T \cdot \mathbf{X} \cdot \boldsymbol{\beta} \end{bmatrix}$$

Top block (first p rows):

Case1 $j \in S:$ We have $\tilde{\mathbf{X}}_j^T \cdot \mathbf{X} \cdot \boldsymbol{\beta}$ and can use Prep1 to get $\mathbf{X}_j^T \cdot \mathbf{X} \cdot \boldsymbol{\beta}$.

Case2 $j \notin S:$ we already have $\mathbf{X}_j^T \cdot \mathbf{X} \cdot \boldsymbol{\beta}$.

Bottom block (last p rows):

Case1 $j \in S:$ We have $\mathbf{X}_j^T \cdot \mathbf{X} \cdot \boldsymbol{\beta}$ and can use Prep1 to get $\tilde{\mathbf{X}}_j^T \cdot \mathbf{X} \cdot \boldsymbol{\beta}$.

Case2 $j \notin S:$ we already have $\tilde{\mathbf{X}}_j^T \cdot \mathbf{X} \cdot \boldsymbol{\beta}$.

$$= \begin{bmatrix} \mathbf{X}_1^T \cdot \mathbf{X} \cdot \boldsymbol{\beta} \\ \mathbf{X}_2^T \cdot \mathbf{X} \cdot \boldsymbol{\beta} \\ \vdots \\ \mathbf{X}_p^T \cdot \mathbf{X} \cdot \boldsymbol{\beta} \\ \tilde{\mathbf{X}}_1^T \cdot \mathbf{X} \cdot \boldsymbol{\beta} \\ \tilde{\mathbf{X}}_2^T \cdot \mathbf{X} \cdot \boldsymbol{\beta} \\ \vdots \\ \tilde{\mathbf{X}}_p^T \cdot \mathbf{X} \cdot \boldsymbol{\beta} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{X}_1^T \\ \mathbf{X}_2^T \\ \vdots \\ \mathbf{X}_p^T \\ \tilde{\mathbf{X}}_1^T \\ \tilde{\mathbf{X}}_2^T \\ \vdots \\ \tilde{\mathbf{X}}_p^T \end{bmatrix} \cdot \mathbf{X} \cdot \boldsymbol{\beta}$$

$$= \left[\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p, \tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, \ldots, \tilde{\mathbf{X}}_p\right]^T \cdot \mathbf{X} \cdot \boldsymbol{\beta}$$

$$= \left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \mathbf{X} \cdot \boldsymbol{\beta}$$

$$= \mathbb{E}\left[\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \mathbf{Y}\right]. \qquad \text{By (6)}$$

□

The next Lemma is the key to prove FDR-control.

Lemma 2.11. *(Lemma 1 in Barber and Candes (2015)). Let* $\mathbf{W} = (W_1, \ldots, W_p)$ *be a vector of FX-Knockoff feature statistics (see Definition 2.4). Let* $W_j \neq 0$ *for all* $j \in \{1, \ldots, p\}$*. Then:*
i) For null j*:* $\text{sign}(W_j)$ *are mutually (jointly) independent and* $\mathbb{P}(\text{sign}(W_j) = -1) = \mathbb{P}(\text{sign}(W_j) = +1) = \frac{1}{2}$*,*
ii) $\text{sign}(W_j)$*,* j *null, are mutually (jointly) independent of* $\text{sign}(W_i)$ *for* i *non-null.*

Proof:

For any set $S \subseteq \{1, \ldots, p\}$, let $\mathbf{W}_{\text{swap}(S)}$ be the statistic we would get if we had replaced $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]$ with $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}$ when calculating $\mathbf{W}$.

By Definition 2.4 i),

$$w_j\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}, \mathbf{Y}\right) = \begin{cases} w_j\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right], \mathbf{Y}\right), & j \notin S \\ -w_j\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right], \mathbf{Y}\right), & j \in S. \end{cases}$$

We can write that as

$$w_j\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}, \mathbf{Y}\right) = w_j\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right], \mathbf{Y}\right) \cdot \delta_j, \quad \text{where } \delta_j = \begin{cases} +1, & j \notin S \\ -1, & j \in S \end{cases}. \tag{8}$$

Prep1 wts $\mathbf{W}_{\text{swap}(S)} = (W_1 \cdot \delta_1, W_2 \cdot \delta_2, \ldots, W_p \cdot \delta_p)$:

$$\begin{aligned} &\mathbf{W}_{\text{swap}(S)} \\ &= \left(w_1\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}, \mathbf{Y}\right), w_2\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}, \mathbf{Y}\right), \ldots, w_p\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}, \mathbf{Y}\right)\right) \\ &\qquad \text{By (8)} \\ &= \left(w_1\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right], \mathbf{Y}\right) \cdot \delta_1, w_2\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right], \mathbf{Y}\right) \cdot \delta_2, \ldots, w_p\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right], \mathbf{Y}\right) \cdot \delta_p\right) \\ &= (W_1 \cdot \delta_1, W_2 \cdot \delta_2, \ldots, W_p \cdot \delta_p). \end{aligned}$$

Prep2 wts $w\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right], \left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}\right)^T \cdot \mathbf{Y}\right) \overset{d}{=} w\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right], \left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \mathbf{Y}\right)$ for any subset of nulls S and for w as in Definition 2.4:

Because S is a subset of nulls, we can apply Lemma 2.10 and get

$$\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}\right)^T \cdot \mathbf{Y} \overset{d}{=} \left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \mathbf{Y}.$$

Note that we treat $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]$ and therefore $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right]$ as fixed (non-random).
Thus we can apply Lemma A.13 ii) with $\mathbf{A} = \left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right]$ and get

$$w\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right], \left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}\right)^T \cdot \mathbf{Y}\right) \overset{d}{=} w\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right], \left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \mathbf{Y}\right).$$

Prep3 wts $(W_1, \ldots, W_p) \overset{d}{=} (W_1 \cdot \delta_1, \ldots, W_p \cdot \delta_p)$ for an arbitrary set of nulls S:

Let $S \subseteq \{1, \ldots, p\}$ be an arbitrary set of nulls.
Since S only contains nulls, we can use Lemma 2.10 for S.

And note that the definition of S is matching with the definition of δ_j above (we just add the restriction that all indices have to be nulls).

Therefore, we can use Prep1 here and have $\mathbf{W}_{\text{swap}(S)} = (W_1 \cdot \delta_1, W_2 \cdot \delta_2, \ldots, W_p \cdot \delta_p)$.

So, it remains to show that $(W_1, \ldots, W_p) \overset{d}{=} \mathbf{W}_{\text{swap}(S)}$.

$\mathbf{W}_{\text{swap}(S)}$

Definition 2.4 ii) used for $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}$ instead of $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]$.

$$= w\left(\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}\right)^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}, \left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}\right)^T \cdot \mathbf{Y}\right)$$

Because $S \subseteq \{1, \ldots, p\}$, we can use Lemma 2.9, which states that $\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}\right)^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)} = \left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right]$.

$$= w\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right], \left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}\right)^T \cdot \mathbf{Y}\right)$$

By Prep2

$$\overset{d}{=} w\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right], \left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \mathbf{Y}\right)$$

By Definition 2.4 ii).

$$= \mathbf{W}$$
$$= (W_1, \ldots, W_p).$$

For the rest of the proof, let wlog $\{1, \ldots, m\}$ be the set of null indices and $\{m+1, \ldots, p\}$ be the set of non-null indices.

ad i) wts $(\text{sign}(W_1), \ldots, \text{sign}(W_m))$ are independent and $\mathbb{P}(\text{sign}(W_j) = -1) = \mathbb{P}(\text{sign}(W_j) = 1) = \frac{1}{2}$ for all $j \in \{1, \ldots, m\}$:

Part1.1 wts $(\text{sign}(W_1), \ldots, \text{sign}(W_m))$ are independent:

We have to show that $\mathbb{P}(\text{sign}(W_1) = k_1, \ldots, \text{sign}(W_m) = k_m) = \prod_{j=1}^{m} \mathbb{P}(\text{sign}(W_j) = k_j)$ for $k_1, \ldots, k_m \in \{-1, 0, +1\}$.

Since $W_j \neq 0$ for all $j \in \{1, \ldots, p\}$ by assumption, it suffices to show this for all $k_1, \ldots, k_m \in \{-1, +1\}$.

Note that $\{\text{sign}(W_j) = k_j\} = \{W_j \cdot k_j > 0\}$ for $j \in \{1, \ldots, p\}$, because:

Case1 $k_j = 1$:

$$\begin{aligned}
&\{\text{sign}(W_j) = k_j\} \\
&\qquad\qquad \text{In this case, } k_j = 1. \\
&= \{\text{sign}(W_j) = 1\} \\
&\qquad\qquad \text{By Definition A.1 (sign-function)} \\
&= \{W_j > 0\} \\
&= \{W_j \cdot 1 > 0\} \\
&\qquad\qquad \text{In this case, } k_j = 1. \\
&= \{W_j \cdot k_j > 0\}.
\end{aligned}$$

Case2 $k_j = -1$:

$$\begin{aligned}
&\{\text{sign}(W_j) = k_j\} \\
&\qquad\qquad \text{In this case, } k_j = -1. \\
&= \{\text{sign}(W_j) = -1\} \\
&\qquad\qquad \text{By Definition A.1 (sign-function)} \\
&= \{W_j < 0\} \\
&= \{W_j \cdot (-1) > 0\} \\
&\qquad\qquad \text{In this case, } k_j = -1. \\
&= \{W_j \cdot k_j > 0\}.
\end{aligned}$$

The same holds for the whole random vector.
Hence, it suffices to show

$$\mathbb{P}(W_1 \cdot k_1 > 0, \ldots, W_m \cdot k_m > 0) = \prod_{j=1}^{m} \mathbb{P}(W_j \cdot k_j > 0) \quad \text{for all } k_1, \ldots, k_m \in \{-1, +1\}. \tag{9}$$

By Prep3 we know that $(W_1, \ldots, W_p) \overset{d}{=} (W_1 \cdot \delta_1, \ldots, W_p \cdot \delta_p)$ for $\delta_j = \begin{cases} +1, & j \notin S \\ -1, & j \in S \end{cases}$ and all $S \subseteq \{1, \ldots, m\}$.
In particular, we get

$$\mathbb{P}(W_1 \cdot k_1 > 0, \ldots, W_m \cdot k_m > 0) = \mathbb{P}(W_1 > 0, \ldots, W_m > 0)$$

for all $k_1, \ldots, k_m \in \{-1, +1\}$ (take $S = \{j \in \{1, \ldots, m\} : k_j = -1\}$).
And of course, $\mathbb{P}(W_j \cdot k_j > 0) = \mathbb{P}(W_j > 0)$ for all $k_j \in \{-1, +1\}$, $j \in \{1, \ldots, m\}$.

We now prove (9) by showing that both sides of (9) are equal to $\frac{1}{2^m}$.
First, consider the LHS of (9):

1

The sum of all possible probabilities has to be 1.

$$= \sum_{k_1,\ldots,k_m \in \{-1,1\}} \mathbb{P}\left(W_1 \cdot k_1 > 0, \ldots, W_m \cdot k_m > 0\right)$$

We know that $\mathbb{P}\left(W_1 \cdot k_1 > 0, \ldots, W_m \cdot k_m > 0\right) = \mathbb{P}\left(W_1 > 0, \ldots, W_m > 0\right)$.

$$= \sum_{k_1,\ldots,k_m \in \{-1,1\}} \mathbb{P}\left(W_1 > 0, \ldots, W_m > 0\right)$$

The probability term does not depend on $k_1, \ldots, k_m$ anymore.

$$= \mathbb{P}\left(W_1 > 0, \ldots, W_m > 0\right) \cdot \sum_{k_1,\ldots,k_m \in \{-1,1\}} 1$$

There are 2^m different combinations of $k_1, \ldots, k_m$.

$$= \mathbb{P}\left(W_1 > 0, \ldots, W_m > 0\right) \cdot 2^m.$$

Rearranged gives $\mathbb{P}\left(W_1 > 0, \ldots, W_m > 0\right) = \frac{1}{2^m}$.
And since we know $\mathbb{P}\left(W_1 \cdot k_1 > 0, \ldots, W_m \cdot k_m > 0\right) = \mathbb{P}\left(W_1 > 0, \ldots, W_m > 0\right)$ for all $k_1, \ldots, k_m \in \{-1, +1\}$, we get $\mathbb{P}\left(W_1 \cdot k_1 > 0, \ldots, W_m \cdot k_m > 0\right) = \frac{1}{2^m}$ for all $k_1, \ldots, k_m \in \{-1, +1\}$.

Second, consider the RHS of (9):

Let $j \in \{1, \ldots, m\}$ be arbitrary.
By assumption, $W_j \neq 0$.
So, we know that $\mathbb{P}\left(W_j = 0\right) = 0$ and hence it has to hold $\mathbb{P}\left(W_j > 0\right) + \mathbb{P}\left(W_j < 0\right) = 1$.
In other words, $\mathbb{P}\left(W_j > 0\right) + \mathbb{P}\left(W_j \cdot (-1) > 0\right) = 1$.
Since we know $\mathbb{P}\left(W_j \cdot k_j > 0\right) = \mathbb{P}\left(W_j > 0\right)$ for all $k_j \in \{-1, +1\}$ from above, we get $\mathbb{P}\left(W_j > 0\right) = \mathbb{P}\left(W_j \cdot (-1) > 0\right)$.
Hence, $\mathbb{P}\left(W_j > 0\right) = \mathbb{P}\left(W_j < 0\right) = \frac{1}{2}$.
Since $j \in \{1, \ldots, m\}$ was arbitrary, we have shown $\mathbb{P}\left(W_j > 0\right) = \mathbb{P}\left(W_j < 0\right) = \frac{1}{2}$ for all $j \in \{1, \ldots, m\}$.

Let $k_1, \ldots, k_m \in \{-1, +1\}$ be arbitrary.

$$\prod_{j=1}^{m} \mathbb{P}\left(W_j \cdot k_j > 0\right)$$

We know that $\mathbb{P}\left(W_j \cdot k_j > 0\right) = \mathbb{P}\left(W_j > 0\right)$ for all $k_j \in \{-1, +1\}$.

$$= \prod_{j=1}^{m} \mathbb{P}\left(W_j > 0\right)$$

We know that $\mathbb{P}\left(W_j > 0\right) = \frac{1}{2}$ for all $j \in \{1, \ldots, m\}$.

$$= \prod_{j=1}^{m} \frac{1}{2}$$
$$= \left(\frac{1}{2}\right)^m$$
$$= \frac{1^m}{2^m}$$
$$= \frac{1}{2^m}.$$

Part1.2 wts $\mathbb{P}(\text{sign}(W_j) = -1) = \mathbb{P}(\text{sign}(W_j) = 1) = \frac{1}{2}$ for all null j:

We know from Part1.1 that $\mathbb{P}(W_j < 0) = \mathbb{P}(W_j > 0) = \frac{1}{2}$ for all null j.
This is equivalent to $\mathbb{P}(\text{sign}(W_j) = -1) = \mathbb{P}(\text{sign}(W_j) = 1) = \frac{1}{2}$ for all null j.

ad ii) wts $(\text{sign}(W_1), \ldots, \text{sign}(W_m))$ is independent of $(\text{sign}(W_{m+1}), \ldots, \text{sign}(W_p))$:

Let $k_1, \ldots, k_m \in \{-1, +1\}$ and $l_{m+1}, \ldots, l_p \in \{-1, +1\}$ be arbitrary (we can ignore 0 since $W_j \neq 0$ for all j by assumption).
We have to show that

$$\begin{aligned}&\mathbb{P}(\text{sign}(W_1) = k_1, \ldots, \text{sign}(W_m) = k_m, \text{sign}(W_{m+1}) = l_{m+1}, \ldots, \text{sign}(W_p) = l_p)\\ &= \mathbb{P}(\text{sign}(W_1) = k_1, \ldots, \text{sign}(W_m) = k_m) \cdot \mathbb{P}(\text{sign}(W_{m+1}) = l_{m+1}, \ldots, \text{sign}(W_p) = l_p).\end{aligned}$$

This is equivalent (by the same argument as in Part1.1) to show that

$$\begin{aligned}&\mathbb{P}(W_1 \cdot k_1 > 0, \ldots, W_m \cdot k_m > 0, \text{sign}(W_{m+1}) = l_{m+1}, \ldots, \text{sign}(W_p) = l_p)\\ &= \mathbb{P}(W_1 \cdot k_1 > 0, \ldots, W_m \cdot k_m > 0) \cdot \mathbb{P}(\text{sign}(W_{m+1}) = l_{m+1}, \ldots, \text{sign}(W_p) = l_p).\end{aligned}$$

By Prep3 we know that $(W_1, \ldots, W_p) \stackrel{d}{=} (W_1 \cdot \delta_1, \ldots, W_p \cdot \delta_p)$ for $\delta_j = \begin{cases} +1, & j \notin S \\ -1, & j \in S \end{cases}$ and all $S \subseteq \{1, \ldots, m\}$.
In particular,

$$\begin{aligned}&\mathbb{P}(W_1 \cdot k_1 > 0, \ldots, W_m \cdot k_m > 0, \text{sign}(W_{m+1}) = l_{m+1}, \ldots, \text{sign}(W_p) = l_p)\\ &= \mathbb{P}(W_1 > 0, \ldots, W_m > 0, \text{sign}(W_{m+1}) = l_{m+1}, \ldots, \text{sign}(W_p) = l_p)\end{aligned} \tag{10}$$

for all $k_1, \ldots, k_m \in \{-1, +1\}$ and all $l_{m+1}, \ldots, l_p \in \{-1, +1\}$.
And of course, $\mathbb{P}(W_1 \cdot k_1 > 0, \ldots, W_m \cdot k_m > 0) = \mathbb{P}(W_1 > 0, \ldots, W_m > 0)$ for all $k_1, \ldots, k_m \in \{-1, +1\}$.
(To see this, take $S = \{j \in \{1, \ldots, m\} : \ k_j = -1\}$ in both cases.)

Therefore, it suffices to show that

$$\begin{aligned}&\mathbb{P}(W_1 > 0, \ldots, W_m > 0, \text{sign}(W_{m+1}) = l_{m+1}, \ldots, \text{sign}(W_p) = l_p)\\ &= \mathbb{P}(W_1 > 0, \ldots, W_m > 0) \cdot \mathbb{P}(\text{sign}(W_{m+1}) = l_{m+1}, \ldots, \text{sign}(W_p) = l_p).\end{aligned}$$

Note that we can marginalize out the first m Knockoff statistics:

$$\begin{aligned}&\sum_{k_1, \ldots, k_m \in \{-1,1\}} \mathbb{P}(W_1 \cdot k_1 > 0, \ldots, W_m \cdot k_m > 0, \text{sign}(W_{m+1}) = l_{m+1}, \ldots, \text{sign}(W_p) = l_p)\\ &= \mathbb{P}(\text{sign}(W_{m+1}) = l_{m+1}, \ldots, \text{sign}(W_p) = l_p).\end{aligned}$$

By using (10), we get

$$\sum_{k_1,\dots,k_m\in\{-1,1\}} \mathbb{P}\left(W_1>0,\dots,W_m>0,\operatorname{sign}(W_{m+1})=l_{m+1},\dots,\operatorname{sign}(W_p)=l_p\right)$$
$$=\mathbb{P}\left(\operatorname{sign}(W_{m+1})=l_{m+1},\dots,\operatorname{sign}(W_p)=l_p\right).$$

And since the probability term does not depend on $k_1,\dots,k_m$ anymore,

$$\mathbb{P}\left(W_1>0,\dots,W_m>0,\operatorname{sign}(W_{m+1})=l_{m+1},\dots,\operatorname{sign}(W_p)=l_p\right)\cdot \sum_{k_1,\dots,k_m\in\{-1,1\}} 1$$
$$=\mathbb{P}\left(\operatorname{sign}(W_{m+1})=l_{m+1},\dots,\operatorname{sign}(W_p)=l_p\right).$$

There are 2^m different combinations of $k_1,\dots,k_m$. Hence we get

$$\mathbb{P}\left(W_1>0,\dots,W_m>0,\operatorname{sign}(W_{m+1})=l_{m+1},\dots,\operatorname{sign}(W_p)=l_p\right)\cdot 2^m$$
$$=\mathbb{P}\left(\operatorname{sign}(W_{m+1})=l_{m+1},\dots,\operatorname{sign}(W_p)=l_p\right).$$

Rearranged,

$$\mathbb{P}\left(W_1>0,\dots,W_m>0,\operatorname{sign}(W_{m+1})=l_{m+1},\dots,\operatorname{sign}(W_p)=l_p\right)$$
$$=\frac{1}{2^m}\cdot\mathbb{P}\left(\operatorname{sign}(W_{m+1})=l_{m+1},\dots,\operatorname{sign}(W_p)=l_p\right).$$

And we know from i) that $\mathbb{P}\left(W_1>0,\dots,W_m>0\right)=\frac{1}{2^m}$.
Therefore, we get

$$\mathbb{P}\left(W_1>0,\dots,W_m>0,\operatorname{sign}(W_{m+1})=l_{m+1},\dots,\operatorname{sign}(W_p)=l_p\right)$$
$$=\mathbb{P}\left(W_1>0,\dots,W_m>0\right)\cdot\mathbb{P}\left(\operatorname{sign}(W_{m+1})=l_{m+1},\dots,\operatorname{sign}(W_p)=l_p\right).$$

□

Comments:
In the comments after Definition 2.4, we listed similarities between FX-Knockoff feature statistics and p-values. Lemma 2.11 shows that our feature statistics (more precicely, the signs of our feature statistics) have much stronger properties than p-values usually have. Usually, the distribution of p-values is known under the null hypothesis. Signs of feature statistics share this property. But for p-values, we usually do not know anything about their dependence structure, even among the nulls. However, we know that signs of null feature statistics are independent.

Theorem 2.12. *(Theorem 1 in Barber and Candes (2015)). For any $q\in(0,1)$, the FX-Knockoff method (see Definition 2.5) which is selecting variables with indices*

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

satisfies

$$\mathrm{mFDR}:=\mathbb{E}\left[\frac{|\left\{j:\ \beta_j=0\ \textit{and}\ j\in\hat{S}\right\}|}{|\hat{S}|+1/q}\right]\leq q,$$

where the expectation is taken over the Gaussian noise $\boldsymbol{z}$ in the model (1) while treating $\mathbf{X}$ and $\tilde{\mathbf{X}}$ as fixed.

Proof:

We only prove the case where:
- $W_j \neq 0$ for all j (hence, the assumptions of Lemma 2.11 are satisfied).
- $|W_j|$ are distinct, i.e., $|W_i| \neq |W_j|$ for $i \neq j$.
- $T < \infty$.

The proof is split into five parts. Part2 and Part3 are martingale results (needed to use optional stopping theorem). Part1 and Part4 are inequalities we are going to use in Part5, where we finally prove FDR-control.

Let $q \in (0, 1)$ be arbitrary but fixed.

We would like to assume that $|W_j|$ are sorted in increasing order to make the proof easier. Such sorting implies that null indices change too.
To address this problem, we define random indices $[1], \ldots, [p]$ through $|W_{[1]}| < |W_{[2]}| < \ldots < |W_{[p]}|$.
The concept is similar to order statistics, except that we use it for ordered absolute values of terms (e.g. for $W_1 = 3$, $W_2 = -1$ and $W_3 = -4$, we have $|W_{[1]}| < |W_{[2]}| < |W_{[3]}|$ where $W_{[1]} = W_2 = -1$, $W_{[2]} = W_1 = 3$ and $W_{[3]} = W_3 = -4$).

For each $k \in \{1, \ldots, p\}$, define

$$A(k) := |\left\{j : \ j \in \{k, \ldots, p\}\,,\ W_{[j]} < 0\right\}|,$$

$$B(k) := |\left\{j : \ j \in \{k, \ldots, p\}\,,\ W_{[j]} > 0\right\}|.$$

In other words, $A(k)$ is the number of negative W_j among the $p - (k-1)$ biggest $|W_j|$.
And $B(k)$ is the number of positive W_j among the $p - (k-1)$ biggest $|W_j|$.

Hence, we can write

$$A(k) = |\left\{j : \ |W_j| \geq |W_{[k]}|,\ W_j < 0\right\}|,$$

$$B(k) = |\left\{j : \ |W_j| \geq |W_{[k]}|,\ W_j > 0\right\}|.$$

In Definition 2.5, we have $\mathcal{W} = \{|W_j| : \ W_j \neq 0,\ j \in \{1, \ldots, p\}\}$.
Since we assume $W_j \neq 0$ in this proof, we can write that as $\mathcal{W} = \left\{|W_{[j]}| : \ j \in \{1, \ldots, p\}\right\}$.

Prep0.1 wts $\left\{j : \ W_j \leq -|W_{[k]}|\right\} = \left\{j : \ |W_j| \geq |W_{[k]}|,\ W_j < 0\right\}$ for $k \in \{1, \ldots, p\}$:

Let $k \in \{1, \ldots, p\}$ be arbitrary.
wts $\left\{j : \ W_j \leq -|W_{[k]}|\right\} \subseteq \left\{j : \ |W_j| \geq |W_{[k]}|,\ W_j < 0\right\}$:

Let $j \in \left\{j : \ W_j \leq -|W_{[k]}|\right\}$ be arbitrary.
This means that $W_j \leq -|W_{[k]}|$.
Since $W_j \neq 0$ for all j, this implies that $W_j < 0$.
Multiplying (-1) on both sides of $W_j \leq -|W_{[k]}|$ gives $-W_j \geq |W_{[k]}|$.
And since $W_j < 0$, we have $-W_j > 0$.
Therefore, $|-W_j| = -W_j$.
And of course (property of absolute value), $|-W_j| = |W_j|$.
Hence, $|W_j| \geq |W_{[k]}|$.
So, we have $|W_j| \geq |W_{[k]}|$ and $W_j < 0$.
Hence, $j \in \left\{j : \ |W_j| \geq |W_{[k]}|,\ W_j < 0\right\}$.

wts $\left\{j : \ W_j \leq -|W_{[k]}|\right\} \supseteq \left\{j : \ |W_j| \geq |W_{[k]}|,\ W_j < 0\right\}$:

Let $j \in \{j : |W_j| \geq |W_{[k]}|,\ W_j < 0\}$ be arbitrary.
This means that $|W_j| \geq |W_{[k]}|$ and $W_j < 0$.
Multiplying (-1) on both sides of $|W_j| \geq |W_{[k]}|$ gives $-|W_j| \leq -|W_{[k]}|$.
And since $W_j < 0$, we have $|W_j| = -W_j$ (by definition of the absolute value).
Hence, $-|W_j| = -(-W_j) = W_j$.
Therefore, $W_j \leq -|W_{[k]}|$.
Hence, $j \in \{j : W_j \leq -|W_{[k]}|\}$.

Both together give $\{j : W_j \leq -|W_{[k]}|\} = \{j : |W_j| \geq |W_{[k]}|,\ W_j < 0\}$.
And since $k \in \{1, \ldots, p\}$ was arbitrary, we have shown this result for all $k \in \{1, \ldots, p\}$.

Prep0.2 wts $\{j : W_j \geq |W_{[k]}|\} = \{j : |W_j| \geq |W_{[k]}|,\ W_j > 0\}$ for $k \in \{1, \ldots, p\}$:

Let $k \in \{1, \ldots, p\}$ be arbitrary.
wts $\{j : W_j \geq |W_{[k]}|\} \subseteq \{j : |W_j| \geq |W_{[k]}|,\ W_j > 0\}$:

Let $j \in \{j : W_j \geq |W_{[k]}|\}$ be arbitrary.
This means that $W_j \geq |W_{[k]}|$.
Since $W_j \neq 0$ for all j, this implies that $W_j > 0$.
Therefore, $|W_j| = W_j$.
So, we have $|W_j| \geq |W_{[k]}|$.
Hence, $j \in \{j : |W_j| \geq |W_{[k]}|,\ W_j > 0\}$.

wts $\{j : W_j \geq |W_{[k]}|\} \supseteq \{j : |W_j| \geq |W_{[k]}|,\ W_j > 0\}$:

Let $j \in \{j : |W_j| \geq |W_{[k]}|,\ W_j > 0\}$ be arbitrary.
This means that $|W_j| \geq |W_{[k]}|$ and $W_j > 0$.
Since $W_j > 0$, we have $|W_j| = W_j$.
Therefore, $W_j \geq |W_{[k]}|$.
Hence, $j \in \{j : W_j \geq |W_{[k]}|\}$.

Both together give $\{j : W_j \geq |W_{[k]}|\} = \{j : |W_j| \geq |W_{[k]}|,\ W_j > 0\}$.
And since $k \in \{1, \ldots, p\}$ was arbitrary, we have shown this result for all $k \in \{1, \ldots, p\}$.

Then

$$
\begin{aligned}
& T \\
& \qquad \text{By Definition 2.5} \\
&= \min\left\{t \in \mathcal{W} : \ \frac{|\{j : W_j \leq -t\}|}{|\{j : W_j \geq t\}| \vee 1} \leq q\right\} \\
& \qquad \text{We know that } \mathcal{W} = \left\{|W_{[j]}| : \ j \in \{1, \ldots, p\}\right\}. \\
&= \min\left\{|W_{[k]}| : \ k \in \{1, \ldots, p\}, \ \frac{|\{j : W_j \leq -|W_{[k]}|\}|}{|\{j : W_j \geq |W_{[k]}|\}| \vee 1} \leq q\right\} \\
& \qquad \text{By Prep0.1 and Prep0.2, we know} \\
& \qquad \text{that } \{j : W_j \leq -|W_{[k]}|\} = \{j : |W_j| \geq |W_{[k]}|,\ W_j < 0\} \\
& \qquad \text{and } \{j : W_j \geq |W_{[k]}|\} = \{j : |W_j| \geq |W_{[k]}|,\ W_j > 0\}. \\
&= \min\left\{|W_{[k]}| : \ k \in \{1, \ldots, p\}, \ \frac{|\{j : |W_j| \geq |W_{[k]}|,\ W_j < 0\}|}{|\{j : |W_j| \geq |W_{[k]}|,\ W_j > 0\}| \vee 1} \leq q\right\} \\
& \qquad \text{By definition of } A(k) \text{ and } B(k) \\
&= \min\left\{|W_{[k]}| : \ k \in \{1, \ldots, p\}, \ \frac{A(k)}{B(k) \vee 1} \leq q\right\}.
\end{aligned}
$$

Set

$$
K := \min\left\{k : \ k \in \{1, \ldots, p\}, \ \frac{A(k)}{B(k) \vee 1} \leq q\right\}.
$$

Since $T < \infty$, we know that at least one $k \in \{1, \ldots, p\}$ satisfies the inequality in the definition of K.
We know from above that T is equal to the smallest $|W_{[k]}|$ such that this inequality is satisfied.
And we know that $|W_{[1]}| < |W_{[2]}| < \ldots < |W_{[p]}|$.
Therefore,

$$T = |W_{[K]}|.$$

Side note: If we would relax the assumption $|W_{[1]}| < |W_{[2]}| < \ldots < |W_{[p]}|$ to $|W_{[1]}| \leq |W_{[2]}| \leq \ldots \leq |W_{[p]}|$, then $A(k)$, $B(k)$ and therefore K would not be unique in general, as a small example shows:

Take $p = 3$ with $W_1 = 3$, $W_2 = -1$ and $W_3 = -3$. Then we can choose if we write $|W_2| \leq |W_1| \leq |W_3|$ or $|W_2| \leq |W_3| \leq |W_1|$.

Case1 $W_{[1]} = W_2$, $W_{[2]} = W_1$, $W_{[3]} = W_3$:

Here, we have
$A(3) = |\{j : \ j \in \{3\}, \ W_{[j]} < 0\}| = |\{3\}| = 1$ and
$B(3) = |\{j : \ j \in \{3\}, \ W_{[j]} > 0\}| = |\emptyset| = 0$.
Hence, in particular,

$$\frac{A(3)}{B(3) \vee 1} = \frac{1}{0 \vee 1} = 1.$$

Case2 $W_{[1]} = W_2$, $W_{[2]} = W_3$, $W_{[3]} = W_1$:

Here, we have
$A(3) = |\{j : \ j \in \{3\}, \ W_{[j]} < 0\}| = |\emptyset| = 0$ and
$B(3) = |\{j : \ j \in \{3\}, \ W_{[j]} > 0\}| = |\{3\}| = 1$.
Hence, in particular,

$$\frac{A(3)}{B(3) \vee 1} = \frac{0}{1 \vee 1} = 0.$$

To continue the proof, note that

$$\begin{aligned}
&\hat{S} \\
&= \{j : \ W_j \geq T\} \\
&\qquad\qquad \text{We know that } T = |W_{[K]}|. \\
&= \{j : \ W_j \geq |W_{[K]}|\} \\
&\qquad\qquad \text{By the same argument as in Prep0.2} \\
&= \{j : \ |W_j| \geq |W_{[K]}|, \ W_j > 0\} \\
&\qquad\qquad \text{By the same argument as above, where} \\
&\qquad\qquad \text{we introduced a different formula for } B(k). \\
&= \{j : \ j \in \{K, \ldots, p\}, \ W_{[j]} > 0\}.
\end{aligned}$$

Prep1.1 wts $1 \leq \dfrac{1 + |\{j : \ j \in \{K, \ldots, p\}, \ W_{[j]} < 0\}|}{1 + |\{j : \ j \in \{K, \ldots, p\}, \ W_{[j]} < 0, \ \beta_{[j]} = 0\}|}$:

It is clear, that $\{j : \ j \in \{K, \ldots, p\}, \ W_{[j]} < 0\} \supseteq \{j : \ j \in \{K, \ldots, p\}, \ W_{[j]} < 0, \ \beta_{[j]} = 0\}$.
Hence, $|\{j : \ j \in \{K, \ldots, p\}, \ W_{[j]} < 0\}| \geq |\{j : \ j \in \{K, \ldots, p\}, \ W_{[j]} < 0, \ \beta_{[j]} = 0\}|$.
So, $1 + |\{j : \ j \in \{K, \ldots, p\}, \ W_{[j]} < 0\}| \geq 1 + |\{j : \ j \in \{K, \ldots, p\}, \ W_{[j]} < 0, \ \beta_{[j]} = 0\}|$.
Rearranged gives

$$1 \leq \frac{1 + |\{j : \ j \in \{K, \ldots, p\}, \ W_{[j]} < 0\}|}{1 + |\{j : \ j \in \{K, \ldots, p\}, \ W_{[j]} < 0, \ \beta_{[j]} = 0\}|}.$$

Prep1.2 wts $\frac{1+A(K)}{B(K)+1/q} \leq q$:

We know that

$$K = \min\left\{k: \ k \in \{1,\dots,p\}, \ \frac{A(k)}{B(k)\vee 1} \leq q\right\}.$$

Therefore, K is one such $k \in \{1,\dots,p\}$ satisfying this inequality. So, it holds

$$\frac{A(K)}{B(K)\vee 1} \leq q.$$

Note that $B(K) > 0$, because:

Assume otherwise, $B(K) = 0$.
Note that $A(k) + B(k) = p - (k-1)$ for all $k \in \{1,\dots,p\}$.
So, $A(K) = p - (K-1)$.
And $B(K) = 0$ also implies $\frac{A(K)}{B(K)\vee 1} = \frac{A(K)}{1}$. Hence by above inequality, $A(K) \leq q$.
Since $q < 1$, this implies $A(K) = 0$.
But that is impossible, because $A(K) = p - (K-1)$.
Even vor $K = p$, we would have $A(K) = p - (p-1) = 1$.
Contradiction!
Hence, the assumption $B(K) = 0$ is wrong.
Therefore (since $B(\cdot)$ is nonnegative by definition), $B(K) > 0$.

Hence we can write

$$\frac{A(K)}{B(K)} \leq q.$$

Rearranged gives (note that $B(K) > 0$)

$$A(K) \leq q \cdot B(K). \tag{11}$$

Therefore,

$$\begin{aligned}
&\frac{1+A(K)}{B(K)+1/q} \\
&\qquad\qquad \text{Multiplied by } q \text{ in numerator and denominator} \\
&\qquad\qquad \text{(possible because } q > 0) \\
&= q \cdot \frac{1+A(K)}{q\cdot B(K)+1} \\
&\qquad\qquad \text{Note that all terms are nonnegative. So, we can use (11).} \\
&\leq q \cdot \frac{1+A(K)}{A(K)+1} \\
&= q \cdot 1 \\
&= q.
\end{aligned}$$

Define

$$V^+(K) := |\left\{j: \ j \in \{K,\dots,p\}, \ W_{[j]} > 0, \ \beta_{[j]} = 0\right\}|,$$

$$V^-(K) := |\left\{j: \ j \in \{K,\dots,p\}, \ W_{[j]} < 0, \ \beta_{[j]} = 0\right\}|.$$

Part1 wts $\dfrac{|\left\{j:\ \beta_{[j]}=0 \text{ and } j\in\hat{S}\right\}|}{|\hat{S}|+1/q}\le q\cdot\dfrac{V^+(K)}{1+V^-(K)}$:

$$\frac{|\left\{j:\ \beta_{[j]}=0 \text{ and } j\in\hat{S}\right\}|}{|\hat{S}|+1/q}$$

We know that $\hat{S}=\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}>0\right\}$.

$$=\frac{|\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}>0,\ \beta_{[j]}=0\right\}|}{|\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}>0\right\}|+1/q}$$
$$=\frac{|\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}>0,\ \beta_{[j]}=0\right\}|}{|\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}>0\right\}|+1/q}\cdot 1$$

The first term is nonnegative.
So, by Prep1.1

$$\le\frac{|\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}>0,\ \beta_{[j]}=0\right\}|}{|\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}>0\right\}|+1/q}\cdot\frac{1+|\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}<0\right\}|}{1+|\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}<0,\ \beta_{[j]}=0\right\}|}$$
$$=\frac{1+|\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}<0\right\}|}{|\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}>0\right\}|+1/q}\cdot\frac{|\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}>0,\ \beta_{[j]}=0\right\}|}{1+|\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}<0,\ \beta_{[j]}=0\right\}|}$$
$$=\frac{1+A(K)}{B(K)+1/q}\cdot\frac{V^+(K)}{1+V^-(K)}$$

The second term is nonnegative.
So, by Prep1.2

$$\le q\cdot\frac{V^+(K)}{1+V^-(K)}.$$

We now introduce martingale results.

Let $\mathfrak{F}=\{\mathcal{F}_k:\ k\in\{1,\dots,p\}\}$ be a filtration with sub-σ-algebras

$$\mathcal{F}_k=\sigma\left(A(l),B(l),V^+(l),V^-(l)\ \forall l\in\{1,\dots,k\};\ \mathbb{1}_{\left\{\beta_{[j]}=0\right\}}\ \forall j\in\{1,\dots,p\}\right).$$

Note that $\mathfrak{F}$ is a filtration, because $\mathcal{F}_1\subseteq\ldots\subseteq\mathcal{F}_p$.

Part2 wts K is a bounded stopping time w.r.t. $\mathfrak{F}$:

Note that $K\in\{1,\dots,p\}$. Therefore, K is bounded.
It remains to show that K is a stopping time.
This means we have to show $\{K=t\}\in\mathcal{F}_t$ for all $t\in\{1,\dots,p\}$.

Remember that

$$K=\min\left\{k:\ k\in\{1,\dots,p\},\ \frac{A(k)}{B(k)\vee 1}\le q\right\}.$$

We see that K is the minimal $k\in\{1,\dots,p\}$, such that a specific inequality is satisfied.

First, observe that $\{K=1\}\in\mathcal{F}_1$:

Note that by definition of $\mathcal{F}_1$, we know $A(1)$, $B(1)$.
Hence, we can tell if the inequality in the definition of K is satisfied.
In other words, we can tell if $K=1$.
Therefore, $\{K=1\}\in\mathcal{F}_1$.

Second, observe that $\{K = 1\}, \{K = 2\} \in \mathcal{F}_2$:

Note that by definition of $\mathcal{F}_2$, we know $A(l)$, $B(l)$ for $l \in \{1, \ldots, 2\}$.
So, we know $A(1)$, $B(1)$, $A(2)$ and $B(2)$.
Knowing $A(1)$ and $B(1)$ tells us if the inequality in the definition of K is satisfied for $k = 1$.
Therefore, $\{K = 1\} \in \mathcal{F}_2$.
Knowing $A(2)$ and $B(2)$ tells us if the inequality in the definition of K is satisfied for $k = 2$.
Therefore, $\{K = 2\} \in \mathcal{F}_2$.

By the same arguments for $3, \ldots, p$, we get overall: $\{K = t\} \in \mathcal{F}_t$ for all $t \in \{1, \ldots, p\}$.

Define $M(k) = \dfrac{V^+(k)}{1 + V^-(k)}$ for $k \in \{1, \ldots, p\}$.

Part3 wts $M(k)$ is a martingale w.r.t. $\mathfrak{F}$:

Part3.1 wts $M(k)$ is adapted to $\mathfrak{F}$:

Let $k \in \{1, \ldots, p\}$ be arbitrary.
We have to show that $M(k)$ is measurable w.r.t. $\mathcal{F}_k$.
By definition of $\mathcal{F}_k$, $V^+(k)$ and $V^-(k)$ are $\mathcal{F}_k$-measurable.
Note that constant functions are measureable.
We know that the sum of two measurable functions is measurable.
Hence, $1 + V^-(k)$ is $\mathcal{F}_k$-measurable.
And we know that the quotient of two measurable functions is measurable.
Therefore, $M(k)$ is $\mathcal{F}_k$-measurable.

Part3.2 wts $\mathbb{E}(|M(k)|) < \infty$ for all $k \in \{1, \ldots, p\}$:

Let $k \in \{1, \ldots, p\}$.

$$
\begin{aligned}
&\mathbb{E}(|M(k)|) && \\
&&& \text{By definition of } M(k) \\
&= \mathbb{E}\left(\left|\frac{V^+(k)}{1 + V^-(k)}\right|\right) && \\
&&& \text{We can omit the absolute value, because} \\
&&& \text{the term inside is always nonnegative} \\
&= \mathbb{E}\left(\frac{V^+(k)}{1 + V^-(k)}\right) && \\
&&& \text{We know that } j \in \{1, \ldots, p\}. \text{ Therefore,} \\
&&& V^+(k) \leq p \text{ and } V^-(k) \geq 0 \text{ for all } k. \\
&&& \text{So, } \frac{V^+(k)}{1 + V^-(k)} \leq \frac{p}{1 + 0} = p. \\
&\leq \mathbb{E}(p) && \\
&= p && \\
&< \infty. &&
\end{aligned}
$$

Part3.3 wts $\mathbb{E}[M(k+1) \mid \mathcal{F}_k] = M(k)$ for all $k \in \{1, \ldots, p-1\}$:

Let $k \in \{1, \ldots, p-1\}$.

Prep3.3.1 wts $M(k+1) = M(k)$ on the event $\left\{\beta_{[k]} \neq 0\right\}$:

$$
\begin{aligned}
& M(k+1) \\
& \qquad \text{By definition of } M(k+1) \\
&= \frac{V^+(k+1)}{1+V^-(k+1)} \\
& \qquad \text{By definition of } V^+(k+1) \text{ and } V^-(k+1) \\
&= \frac{\left|\left\{j:\ j \in \{k+1, \ldots, p\},\ W_{[j]} > 0,\ \beta_{[j]} = 0\right\}\right|}{1+\left|\left\{j:\ j \in \{k+1, \ldots, p\},\ W_{[j]} < 0,\ \beta_{[j]} = 0\right\}\right|} \\
& \qquad \text{The counts in numerator and denominator} \\
& \qquad \text{do not change if we start with } k \text{ instead of } k+1 \\
& \qquad \text{since we know that } \beta_{[k]} \neq 0. \\
&= \frac{\left|\left\{j:\ j \in \{k, \ldots, p\},\ W_{[j]} > 0,\ \beta_{[j]} = 0\right\}\right|}{1+\left|\left\{j:\ j \in \{k, \ldots, p\},\ W_{[j]} < 0,\ \beta_{[j]} = 0\right\}\right|} \\
& \qquad \text{By definition of } V^+(k) \text{ and } V^-(k) \\
&= \frac{V^+(k)}{1+V^-(k)} \\
& \qquad \text{By definition of } M(k) \\
&= M(k).
\end{aligned}
$$

Prep3.3.2 wts $M(k+1) = \dfrac{V^+(k)-1}{1+V^-(k)}$ on the event $\left\{\beta_{[k]} = 0,\ W_{[k]} > 0\right\}$:

$$
\begin{aligned}
& M(k+1) \\
& \qquad \text{By definition of } M(k+1) \\
&= \frac{V^+(k+1)}{1+V^-(k+1)} \\
& \qquad \text{By definition of } V^+(k+1) \text{ and } V^-(k+1) \\
&= \frac{\left|\left\{j:\ j \in \{k+1, \ldots, p\},\ W_{[j]} > 0,\ \beta_{[j]} = 0\right\}\right|}{1+\left|\left\{j:\ j \in \{k+1, \ldots, p\},\ W_{[j]} < 0,\ \beta_{[j]} = 0\right\}\right|} \\
& \qquad \text{Since we consider the event } \left\{\beta_{[k]} = 0,\ W_{[k]} > 0\right\}, \\
& \qquad \text{the numerator will increase by 1 if we start} \\
& \qquad \text{with } k \text{ instead of } k+1. \text{ The denominator is not affected.} \\
&= \frac{\left|\left\{j:\ j \in \{k, \ldots, p\},\ W_{[j]} > 0,\ \beta_{[j]} = 0\right\}\right| - 1}{1+\left|\left\{j:\ j \in \{k, \ldots, p\},\ W_{[j]} < 0,\ \beta_{[j]} = 0\right\}\right|} \\
& \qquad \text{By definition of } V^+(k) \text{ and } V^-(k) \\
&= \frac{V^+(k)-1}{1+V^-(k)}.
\end{aligned}
$$

Prep3.3.3 wts $M(k+1) = \dfrac{V^+(k)}{V^-(k)}$ on the event $\left\{\beta_{[k]} = 0,\ W_{[k]} < 0\right\}$:

$$
\begin{aligned}
& M(k+1) \\
& \qquad \text{By definition of } M(k+1) \\
&= \frac{V^+(k+1)}{1+V^-(k+1)}
\end{aligned}
$$

By definition of $V^+(k+1)$ and $V^-(k+1)$

$$= \frac{|\left\{j:\ j \in \{k+1,\dots,p\},\ W_{[j]} > 0,\ \beta_{[j]} = 0\right\}|}{1 + |\left\{j:\ j \in \{k+1,\dots,p\},\ W_{[j]} < 0,\ \beta_{[j]} = 0\right\}|}$$

Since we consider the event $\left\{\beta_{[k]} = 0,\ W_{[k]} < 0\right\}$, the denominator will increase by 1 if we start with k instead of $k+1$. The numerator is not affected.

$$= \frac{|\left\{j:\ j \in \{k,\dots,p\},\ W_{[j]} > 0,\ \beta_{[j]} = 0\right\}|}{1 + |\left\{j:\ j \in \{k,\dots,p\},\ W_{[j]} < 0,\ \beta_{[j]} = 0\right\}| - 1}$$

By definition of $V^+(k)$ and $V^-(k)$

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

Note that $V^-(k) > 0$ in this case, because we consider the event $\left\{\beta_{[k]} = 0,\ W_{[k]} < 0\right\}$ and hence, $V^-(k) \geq 1$.

Prep3.3.4 wts $\mathbb{1}_{\left\{\beta_{[k]}=0\right\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] = \mathbb{1}_{\left\{\beta_{[k]}=0\right\}} \cdot \frac{V^+(k)}{V^+(k)+V^-(k)}$:

Note that we can write $V^+(k) = \sum_{j=k}^{p} \mathbb{1}_{\left\{\beta_{[j]}=0\right\}} \cdot \mathbb{1}_{\left\{W_{[j]}>0\right\}}$.
Conditional expected value applied on both sides gives

$$\mathbb{E}\left[V^+(k) \mid \mathcal{F}_k\right] = \mathbb{E}\left[\sum_{j=k}^{p} \mathbb{1}_{\left\{\beta_{[j]}=0\right\}} \cdot \mathbb{1}_{\left\{W_{[j]}>0\right\}} \mid \mathcal{F}_k\right].$$

Since $V^+(k), \left\{\beta_{[j]} = 0\right\} \in \mathcal{F}_k$ for all $j \in \{1,\dots,p\}$, we can pull these terms out and get

$$V^+(k) \cdot \mathbb{E}\left[1 \mid \mathcal{F}_k\right] = \sum_{j=k}^{p} \mathbb{1}_{\left\{\beta_{[j]}=0\right\}} \cdot \mathbb{E}\left[\mathbb{1}_{\left\{W_{[j]}>0\right\}} \mid \mathcal{F}_k\right].$$

Hence,

$$V^+(k) = \sum_{j=k}^{p} \mathbb{1}_{\left\{\beta_{[j]}=0\right\}} \cdot \mathbb{P}\left[W_{[j]} > 0 \mid \mathcal{F}_k\right].$$

By Lemma 2.11 i) we know for null $[j]$: $\operatorname{sign}(W_{[j]})$ are independent and $\mathbb{P}\left(\operatorname{sign}(W_{[j]}) = -1\right) = \mathbb{P}\left(\operatorname{sign}(W_{[j]}) = +1\right) = \frac{1}{2}$.
The latter can be written as $\mathbb{P}\left(W_{[j]} < 0\right) = \mathbb{P}\left(W_{[j]} > 0\right) = \frac{1}{2}$.
In other words, the signs of the null FX-Knockoff feature statistics are i.i.d. and hence exchangeable.
By definition of $\mathcal{F}_k$, all nulls are known.
Therefore,

$$\mathbb{1}_{\left\{\beta_{[j]}=0\right\}} \cdot \mathbb{1}_{\left\{\beta_{[k]}=0\right\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] = \mathbb{1}_{\left\{\beta_{[j]}=0\right\}} \cdot \mathbb{1}_{\left\{\beta_{[k]}=0\right\}} \cdot \mathbb{P}\left[W_{[j]} > 0 \mid \mathcal{F}_k\right]$$

for all $j \in \{k,\dots,p\}$.

And note that

$\mathbb{1}_{\left\{\beta_{[k]}=0\right\}} \cdot V^+(k)$

We know from above that $V^+(k) = \sum_{j=k}^{p} \mathbb{1}_{\left\{\beta_{[j]}=0\right\}} \cdot \mathbb{P}\left[W_{[j]} > 0 \mid \mathcal{F}_k\right]$.

$$= \mathbb{1}_{\left\{\beta_{[k]}=0\right\}} \cdot \sum_{j=k}^{p} \mathbb{1}_{\left\{\beta_{[j]}=0\right\}} \cdot \mathbb{P}\left[W_{[j]} > 0 \mid \mathcal{F}_k\right]$$

$$= \sum_{j=k}^{p} \mathbb{1}_{\{\beta_{[j]}=0\}} \cdot \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{P}\left[W_{[j]} > 0 \mid \mathcal{F}_k\right]$$

We know that $\mathbb{1}_{\{\beta_{[j]}=0\}} \cdot \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] = \mathbb{1}_{\{\beta_{[j]}=0\}} \cdot \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{P}\left[W_{[j]} > 0 \mid \mathcal{F}_k\right]$.

$$= \sum_{j=k}^{p} \mathbb{1}_{\{\beta_{[j]}=0\}} \cdot \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right]$$

$$= \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] \cdot \sum_{j=k}^{p} \mathbb{1}_{\{\beta_{[j]}=0\}}$$

$$= \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] \cdot |\left\{j : \ j \in \{k, \ldots, p\}, \ \beta_{[j]} = 0\right\}|$$

Note that $V^+(k) + V^-(k) = |\left\{j : \ j \in \{k, \ldots, p\}, \ \beta_{[j]} = 0\right\}|$.

$$= \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] \cdot \left(V^+(k) + V^-(k)\right).$$

Rearranged gives

$$\mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] = \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \frac{V^+(k)}{V^+(k) + V^-(k)}.$$

Therefore,

$$\mathbb{E}\left[M(k+1) \mid \mathcal{F}_k\right]$$

$$= \mathbb{E}\left[M(k+1) \cdot \mathbb{1}_{\{\beta_{[k]}\neq 0\} \cup \{\beta_{[k]}=0\}} \mid \mathcal{F}_k\right]$$

Note that $\mathbb{1}_{C \cup D} = \mathbb{1}_C + \mathbb{1}_D$ for disjoint sets C, D.

$$= \mathbb{E}\left[M(k+1) \cdot \left(\mathbb{1}_{\{\beta_{[k]}\neq 0\}} + \mathbb{1}_{\{\beta_{[k]}=0\}}\right) \mid \mathcal{F}_k\right]$$

$$= \mathbb{E}\left[M(k+1) \cdot \left(\mathbb{1}_{\{\beta_{[k]}\neq 0\}} + \mathbb{1}_{\{\beta_{[k]}=0,\, W_{[k]}>0\} \cup \{\beta_{[k]}=0,\, W_{[k]}<0\}}\right) \mid \mathcal{F}_k\right]$$

Note that $\mathbb{1}_{C \cup D} = \mathbb{1}_C + \mathbb{1}_D$ for disjoint sets C, D.

$$= \mathbb{E}\left[M(k+1) \cdot \left(\mathbb{1}_{\{\beta_{[k]}\neq 0\}} + \mathbb{1}_{\{\beta_{[k]}=0,\, W_{[k]}>0\}} + \mathbb{1}_{\{\beta_{[k]}=0,\, W_{[k]}<0\}}\right) \mid \mathcal{F}_k\right]$$

$$= \mathbb{E}\left[M(k+1) \cdot \mathbb{1}_{\{\beta_{[k]}\neq 0\}} + M(k+1) \cdot \mathbb{1}_{\{\beta_{[k]}=0,\, W_{[k]}>0\}} + M(k+1) \cdot \mathbb{1}_{\{\beta_{[k]}=0,\, W_{[k]}<0\}} \mid \mathcal{F}_k\right]$$

$$= \mathbb{E}\left[M(k+1) \cdot \mathbb{1}_{\{\beta_{[k]}\neq 0\}} \mid \mathcal{F}_k\right]$$
$$+ \mathbb{E}\left[M(k+1) \cdot \mathbb{1}_{\{\beta_{[k]}=0,\, W_{[k]}>0\}} \mid \mathcal{F}_k\right]$$
$$+ \mathbb{E}\left[M(k+1) \cdot \mathbb{1}_{\{\beta_{[k]}=0,\, W_{[k]}<0\}} \mid \mathcal{F}_k\right]$$

By Prep3.3.1, Prep3.3.2 and Prep3.3.3

$$= \mathbb{E}\left[M(k) \cdot \mathbb{1}_{\{\beta_{[k]}\neq 0\}} \mid \mathcal{F}_k\right]$$
$$+ \mathbb{E}\left[\frac{V^+(k)-1}{1+V^-(k)} \cdot \mathbb{1}_{\{\beta_{[k]}=0,\, W_{[k]}>0\}} \mid \mathcal{F}_k\right]$$
$$+ \mathbb{E}\left[\frac{V^+(k)}{V^-(k)} \cdot \mathbb{1}_{\{\beta_{[k]}=0,\, W_{[k]}<0\}} \mid \mathcal{F}_k\right]$$

Note that $\mathbb{1}_{C \cap D} = \mathbb{1}_C \cdot \mathbb{1}_D$ for sets C, D.

$$= \mathbb{E}\left[M(k) \cdot \mathbb{1}_{\{\beta_{[k]}\neq 0\}} \mid \mathcal{F}_k\right]$$
$$+ \mathbb{E}\left[\frac{V^+(k)-1}{1+V^-(k)} \cdot \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{1}_{\{W_{[k]}>0\}} \mid \mathcal{F}_k\right]$$
$$+ \mathbb{E}\left[\frac{V^+(k)}{V^-(k)} \cdot \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{1}_{\{W_{[k]}<0\}} \mid \mathcal{F}_k\right]$$

By definition of $\mathcal{F}_k$, we know that $V^+(k), V^-(k), \{\beta_{[k]} = 0\} \in \mathcal{F}_k$.
Since complements of elements of a σ-algebra are
again in that σ-algebra, $\{\beta_{[k]} = 0\}^c = \{\beta_{[k]} \neq 0\} \in \mathcal{F}_k$.
Hence we can pull out these terms.

$$
\begin{aligned}
&= M(k) \cdot \mathbb{1}_{\{\beta_{[k]} \neq 0\}} \cdot \mathbb{E}\left[1 \mid \mathcal{F}_k\right] \\
&\quad + \frac{V^+(k) - 1}{1 + V^-(k)} \cdot \mathbb{1}_{\{\beta_{[k]} = 0\}} \cdot \mathbb{E}\left[\mathbb{1}_{\{W_{[k]} > 0\}} \mid \mathcal{F}_k\right] \\
&\quad + \frac{V^+(k)}{V^-(k)} \cdot \mathbb{1}_{\{\beta_{[k]} = 0\}} \cdot \mathbb{E}\left[\mathbb{1}_{\{W_{[k]} < 0\}} \mid \mathcal{F}_k\right] \\
&= M(k) \cdot \mathbb{1}_{\{\beta_{[k]} \neq 0\}} \\
&\quad + \frac{V^+(k) - 1}{1 + V^-(k)} \cdot \mathbb{1}_{\{\beta_{[k]} = 0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] \\
&\quad + \frac{V^+(k)}{V^-(k)} \cdot \mathbb{1}_{\{\beta_{[k]} = 0\}} \cdot \mathbb{P}\left[W_{[k]} < 0 \mid \mathcal{F}_k\right]
\end{aligned}
$$

By Prep3.3.4

$$
\begin{aligned}
&= M(k) \cdot \mathbb{1}_{\{\beta_{[k]} \neq 0\}} \\
&\quad + \frac{V^+(k) - 1}{1 + V^-(k)} \cdot \mathbb{1}_{\{\beta_{[k]} = 0\}} \cdot \frac{V^+(k)}{V^+(k) + V^-(k)} \\
&\quad + \frac{V^+(k)}{V^-(k)} \cdot \mathbb{1}_{\{\beta_{[k]} = 0\}} \cdot \left(1 - \frac{V^+(k)}{V^+(k) + V^-(k)}\right) \\
&= M(k) \cdot \mathbb{1}_{\{\beta_{[k]} \neq 0\}} \\
&\quad + \mathbb{1}_{\{\beta_{[k]} = 0\}} \cdot \frac{V^+(k) - 1}{1 + V^-(k)} \cdot \frac{V^+(k)}{V^+(k) + V^-(k)} \\
&\quad + \mathbb{1}_{\{\beta_{[k]} = 0\}} \cdot \frac{V^+(k)}{V^-(k)} \cdot \left(\frac{V^+(k) + V^-(k)}{V^+(k) + V^-(k)} - \frac{V^+(k)}{V^+(k) + V^-(k)}\right) \\
&= M(k) \cdot \mathbb{1}_{\{\beta_{[k]} \neq 0\}} \\
&\quad + \mathbb{1}_{\{\beta_{[k]} = 0\}} \cdot \frac{V^+(k) - 1}{1 + V^-(k)} \cdot \frac{V^+(k)}{V^+(k) + V^-(k)} \\
&\quad + \mathbb{1}_{\{\beta_{[k]} = 0\}} \cdot \frac{V^+(k)}{V^-(k)} \cdot \frac{V^-(k)}{V^+(k) + V^-(k)} \\
&= M(k) \cdot \mathbb{1}_{\{\beta_{[k]} \neq 0\}} + \mathbb{1}_{\{\beta_{[k]} = 0\}} \cdot \left(\frac{V^+(k) - 1}{1 + V^-(k)} \cdot \frac{V^+(k)}{V^+(k) + V^-(k)} + \frac{V^-(k)}{V^-(k)} \cdot \frac{V^+(k)}{V^+(k) + V^-(k)}\right) \\
&= M(k) \cdot \mathbb{1}_{\{\beta_{[k]} \neq 0\}} + \mathbb{1}_{\{\beta_{[k]} = 0\}} \cdot \left(\frac{V^+(k)}{V^+(k) + V^-(k)} \cdot \left(\frac{V^+(k) - 1}{1 + V^-(k)} + 1\right)\right) \\
&= M(k) \cdot \mathbb{1}_{\{\beta_{[k]} \neq 0\}} + \mathbb{1}_{\{\beta_{[k]} = 0\}} \cdot \left(\frac{V^+(k)}{V^+(k) + V^-(k)} \cdot \left(\frac{V^+(k) - 1}{1 + V^-(k)} + \frac{1 + V^-(k)}{1 + V^-(k)}\right)\right) \\
&= M(k) \cdot \mathbb{1}_{\{\beta_{[k]} \neq 0\}} + \mathbb{1}_{\{\beta_{[k]} = 0\}} \cdot \left(\frac{V^+(k)}{V^+(k) + V^-(k)} \cdot \frac{V^+(k) + V^-(k)}{1 + V^-(k)}\right) \\
&= M(k) \cdot \mathbb{1}_{\{\beta_{[k]} \neq 0\}} + \mathbb{1}_{\{\beta_{[k]} = 0\}} \cdot \frac{V^+(k)}{1 + V^-(k)}
\end{aligned}
$$

By definition of $M(k)$

$$
\begin{aligned}
&= M(k) \cdot \mathbb{1}_{\{\beta_{[k]} \neq 0\}} + \mathbb{1}_{\{\beta_{[k]} = 0\}} \cdot M(k) \\
&= M(k) \cdot \left(\mathbb{1}_{\{\beta_{[k]} \neq 0\}} + \mathbb{1}_{\{\beta_{[k]} = 0\}}\right)
\end{aligned}
$$

Note that $\mathbb{1}_{C \cup D} = \mathbb{1}_C + \mathbb{1}_D$ for disjoint sets C, D.

$$
\begin{aligned}
&= M(k) \cdot \mathbb{1}_{\{\beta_{[k]} \neq 0\} \cup \{\beta_{[k]} = 0\}} \\
&= M(k).
\end{aligned}
$$

Since $k \in \{1, \ldots, p-1\}$ was arbitrary, we have shown $\mathbb{E}\left[M(k+1) \mid \mathcal{F}_k\right] = M(k)$ for all $k \in \{1, \ldots, p-1\}$.

Define $N_0 = |\left\{j \in \{1,\ldots,p\} :\ \beta_{[j]} = 0\right\}|$ as the total number of nulls.

Prep4.1 wts $V^+(1) \sim \text{Binomial}\left(N_0, \frac{1}{2}\right)$:

Note that we can write

$$V^+(1) = |\left\{j :\ j \in \{1,\ldots,p\},\ W_{[j]} > 0,\ \beta_{[j]} = 0\right\}| = |\left\{j :\ j \in \{1,\ldots,p\},\ \text{sign}(W_{[j]}) = +1,\ \beta_{[j]} = 0\right\}|.$$

Define $U_j = \mathbb{1}_{\left\{\text{sign}(W_{[j]})=+1\right\}}$.
Note that $\mathbb{P}(U_j = 1) = \mathbb{P}\left(\text{sign}(W_{[j]}) = +1\right) = \frac{1}{2}$ for all j where $\beta_{[j]} = 0$ by Lemma 2.11 i).
By Lemma 2.11 i), $\text{sign}(W_{[j]})$ are independent for j where $\beta_{[j]} = 0$.
This means that U_j are independent random variables with $U_j \sim \text{Bernoulli}\left(\frac{1}{2}\right)$ for all j where $\beta_{[j]} = 0$.
Short: $U_j \overset{i.i.d.}{\sim} \text{Bernoulli}\left(\frac{1}{2}\right)$ for j where $\beta_{[j]} = 0$.
Note that $\sum_{\substack{j \in \{1,\ldots,p\},\\ \beta_{[j]}=0}} U_j = V^+(1)$.
By definition of the Binomial-distribution, $V^+(1) \sim \text{Binomial}\left(N_0, \frac{1}{2}\right)$.

Prep4.2 wts $V^-(1) = N_0 - V^+(1)$:

$$\begin{aligned}
&V^-(1)\\
&= |\left\{j :\ j \in \{1,\ldots,p\},\ W_{[j]} < 0,\ \beta_{[j]} = 0\right\}|\\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{Note that } N_0 = |\left\{j \in \{1,\ldots,p\} :\ \beta_{[j]} = 0\right\}|.\\
&= N_0 - |\left\{j :\ j \in \{1,\ldots,p\},\ W_{[j]} > 0,\ \beta_{[j]} = 0\right\}|\\
&= N_0 - V^+(1).
\end{aligned}$$

Part4 wts $\mathbb{E}\left[\frac{V^+(K)}{1+V^-(K)}\right] \leq 1$:

$$\begin{aligned}
&\mathbb{E}\left[\frac{V^+(K)}{1+V^-(K)}\right]\\
&\qquad\qquad \text{By Part2, we know that } K \text{ is a bounded stopping time.}\\
&\qquad\qquad \text{By Part3, we know that } \frac{V^+(k)}{1+V^-(k)} \text{ is a martingale.}\\
&\qquad\qquad \text{Therefore, we can use Lemma A.4 i)}\\
&\qquad\qquad \text{(optional stopping theorem for martingales).}\\
&= \mathbb{E}\left[\frac{V^+(1)}{1+V^-(1)}\right]
\end{aligned}$$

By Prep4.2

$$= \mathbb{E}\left[\frac{V^+(1)}{1+N_0-V^+(1)}\right]$$

In general, it holds for a discrete random variable Y and function g that $\mathbb{E}(g(Y)) = \sum_y g(y)\cdot\mathbb{P}(Y=y)$. In our case, $g(x) := \frac{x}{1+N_0-x}$ and $Y := V^+(1)$. Note that $N_0 \le p$ and $Y \in \{0,\dots,N_0\}$.

$$= \sum_{i=0}^{N_0} g(i)\cdot\mathbb{P}(Y=i)$$

By definition of g

$$= \sum_{i=0}^{N_0} \frac{i}{1+N_0-i}\cdot\mathbb{P}(Y=i)$$

By Prep4.1, $V^+(1) \sim \text{Binomial}\left(N_0, \frac{1}{2}\right)$.

$$= \sum_{i=0}^{N_0} \frac{i}{1+N_0-i}\cdot\binom{N_0}{i}\cdot\left(\frac{1}{2}\right)^i\cdot\left(1-\frac{1}{2}\right)^{N_0-i}$$

We can start the sum at $i=1$.

$$= \sum_{i=1}^{N_0} \frac{i}{1+N_0-i}\cdot\binom{N_0}{i}\cdot\left(\frac{1}{2}\right)^i\cdot\left(1-\frac{1}{2}\right)^{N_0-i}$$

$$= \sum_{i=1}^{N_0} \frac{i}{1+N_0-i}\cdot\frac{N_0!}{i!\cdot(N_0-i)!}\cdot\left(\frac{1}{2}\right)^i\cdot\left(\frac{1}{2}\right)^{N_0-i}$$

$$= \sum_{i=1}^{N_0} \frac{1}{1+N_0-i}\cdot\frac{N_0!}{(i-1)!\cdot(N_0-i)!}\cdot\left(\frac{1}{2}\right)^{i-1}\cdot\left(\frac{1}{2}\right)^1\cdot\left(\frac{1}{2}\right)^{N_0-(i-1)}\cdot\left(\frac{1}{2}\right)^{-1}$$

Note that $1+N_0-i = N_0-(i-1)$.

$$= \sum_{i=1}^{N_0} \frac{N_0!}{(i-1)!\cdot(N_0-(i-1))!}\cdot\left(\frac{1}{2}\right)^{i-1}\cdot\left(\frac{1}{2}\right)^{N_0-(i-1)}$$

$$= \sum_{i=1}^{N_0} \binom{N_0}{i-1}\cdot\left(\frac{1}{2}\right)^{i-1}\cdot\left(\frac{1}{2}\right)^{N_0-(i-1)}$$

Substitution: set $j = i-1$.

$$= \sum_{j=0}^{N_0-1} \underbrace{\binom{N_0}{j}\cdot\left(\frac{1}{2}\right)^j\cdot\left(\frac{1}{2}\right)^{N_0-j}}_{=\mathbb{P}(Y=j)}$$

The sum of probabilities the random variable Y can achieve has to be smaller than 1.

$$\le 1.$$

Part5 wts mFDR $\le q$:

mFDR

$$= \mathbb{E}\left[\frac{|\left\{j:\ \beta_{[j]}=0 \text{ and } j\in\hat{S}\right\}|}{|\hat{S}|+1/q}\right]$$

By Part1, we know that $\frac{|\left\{j:\ \beta_{[j]}=0 \text{ and } j\in\hat{S}\right\}|}{|\hat{S}|+1/q} \le q\cdot\frac{V^+(K)}{1+V^-(K)}$.

$$\le \mathbb{E}\left[q\cdot\frac{V^+(K)}{1+V^-(K)}\right]$$
$$= q\cdot\mathbb{E}\left[\frac{V^+(K)}{1+V^-(K)}\right]$$

By Part4, we know that $\mathbb{E}\left[\frac{V^+(K)}{1+V^-(K)}\right]\le 1$.

$$\le q\cdot 1$$
$$= q.$$

As $q\in(0,1)$ was arbitrary but fixed, we have shown the result for all $q\in(0,1)$.

□

Theorem 2.13. *(Theorem 2 of Barber and Candes (2015)) For any $q\in[0,1]$, the FX-Knockoff+ method (see Definition 2.6) which is selecting variables with indices*

$$\hat{S}=\{j:\ W_j\ge T_+\}$$

satisfies

$$\mathrm{FDR}=\mathbb{E}\left[\mathrm{FDP}\right]\le q,$$

where the expectation is taken over the Gaussian noise $\boldsymbol{z}$ in model (1) while treating $\mathbf{X}$ and $\tilde{\mathbf{X}}$ as fixed.

Proof:

Side note: This proof is almost identical to the proof of Theorem 2.12.

We only prove the case where:
- $W_j\neq 0$ for all j (hence, the assumptions of Lemma 2.11 are satisfied).
- $|W_j|$ are distinct, i.e., $|W_i|\neq|W_j|$ for $i\neq j$.
- $T_+<\infty$.

Let $q\in[0,1]$ be arbitrary but fixed.

We would like to assume that $|W_j|$ are sorted in increasing order to make the proof easier. Such sorting implies that null indices change too.
To address this problem, we define random indices $[1],\ldots,[p]$ through $|W_{[1]}|<|W_{[2]}|<\ldots<|W_{[p]}|$.
The concept is similar to order statistics, except that we use it for ordered absolute values of terms (e.g. for $W_1=3$, $W_2=-1$ and $W_3=-4$, we have $|W_{[1]}|<|W_{[2]}|<|W_{[3]}|$ where $W_{[1]}=W_2=-1$, $W_{[2]}=W_1=3$ and $W_{[3]}=W_3=-4$).

For each $k\in\{1,\ldots,p\}$, define

$$A(k):=|\left\{j:\ j\in\{k,\ldots,p\},\ W_{[j]}<0\right\}|,$$

$$B(k) := |\left\{j:\ j \in \{k,\dots,p\},\ W_{[j]} > 0\right\}|.$$

In other words, $A(k)$ is the number of negative W_j among the $p-(k-1)$ biggest $|W_j|$.
And $B(k)$ is the number of positive W_j among the $p-(k-1)$ biggest $|W_j|$.

Hence, we can write

$$A(k) = |\left\{j:\ |W_j| \geq |W_{[k]}|,\ W_j < 0\right\}|,$$

$$B(k) = |\left\{j:\ |W_j| \geq |W_{[k]}|,\ W_j > 0\right\}|.$$

In Definition 2.6, we have $\mathcal{W} = \{|W_j|:\ W_j \neq 0,\ j \in \{1,\dots,p\}\}$.
Since we assume $W_j \neq 0$ in this proof, we can write that as $\mathcal{W} = \left\{|W_{[j]}|:\ j \in \{1,\dots,p\}\right\}$.

Prep0.1 wts $\left\{j:\ W_j \leq -|W_{[k]}|\right\} = \left\{j:\ |W_j| \geq |W_{[k]}|,\ W_j < 0\right\}$ for $k \in \{1,\dots,p\}$:

Let $k \in \{1,\dots,p\}$ be arbitrary.
wts $\left\{j:\ W_j \leq -|W_{[k]}|\right\} \subseteq \left\{j:\ |W_j| \geq |W_{[k]}|,\ W_j < 0\right\}$:

Let $j \in \left\{j:\ W_j \leq -|W_{[k]}|\right\}$ be arbitrary.
This means that $W_j \leq -|W_{[k]}|$.
Since $W_j \neq 0$ for all j, this implies that $W_j < 0$.
Multiplying (-1) on both sides of $W_j \leq -|W_{[k]}|$ gives $-W_j \geq |W_{[k]}|$.
And since $W_j < 0$, we have $-W_j > 0$.
Therefore, $|-W_j| = -W_j$.
And of course (property of absolute value), $|-W_j| = |W_j|$.
Hence, $|W_j| \geq |W_{[k]}|$.
So, we have $|W_j| \geq |W_{[k]}|$ and $W_j < 0$.
Hence, $j \in \left\{j:\ |W_j| \geq |W_{[k]}|,\ W_j < 0\right\}$.

wts $\left\{j:\ W_j \leq -|W_{[k]}|\right\} \supseteq \left\{j:\ |W_j| \geq |W_{[k]}|,\ W_j < 0\right\}$:

Let $j \in \left\{j:\ |W_j| \geq |W_{[k]}|,\ W_j < 0\right\}$ be arbitrary.
This means that $|W_j| \geq |W_{[k]}|$ and $W_j < 0$.
Multiplying (-1) on both sides of $|W_j| \geq |W_{[k]}|$ gives $-|W_j| \leq -|W_{[k]}|$.
And since $W_j < 0$, we have $|W_j| = -W_j$ (by definition of the absolute value).
Hence, $-|W_j| = -(-W_j) = W_j$.
Therefore, $W_j \leq -|W_{[k]}|$.
Hence, $j \in \left\{j:\ W_j \leq -|W_{[k]}|\right\}$.

Both together give $\left\{j:\ W_j \leq -|W_{[k]}|\right\} = \left\{j:\ |W_j| \geq |W_{[k]}|,\ W_j < 0\right\}$.
And since $k \in \{1,\dots,p\}$ was arbitrary, we have shown this result for all $k \in \{1,\dots,p\}$.

Prep0.2 wts $\left\{j:\ W_j \geq |W_{[k]}|\right\} = \left\{j:\ |W_j| \geq |W_{[k]}|,\ W_j > 0\right\}$ for $k \in \{1,\dots,p\}$:

Let $k \in \{1,\dots,p\}$ be arbitrary.
wts $\left\{j:\ W_j \geq |W_{[k]}|\right\} \subseteq \left\{j:\ |W_j| \geq |W_{[k]}|,\ W_j > 0\right\}$:

Let $j \in \left\{j:\ W_j \geq |W_{[k]}|\right\}$ be arbitrary.
This means that $W_j \geq |W_{[k]}|$.
Since $W_j \neq 0$ for all j, this implies that $W_j > 0$.
Therefore, $|W_j| = W_j$.
So, we have $|W_j| \geq |W_{[k]}|$.
Hence, $j \in \left\{j:\ |W_j| \geq |W_{[k]}|,\ W_j > 0\right\}$.

wts $\left\{j:\ W_j \geq |W_{[k]}|\right\} \supseteq \left\{j:\ |W_j| \geq |W_{[k]}|,\ W_j > 0\right\}$:

Let $j \in \{j :\ |W_j| \geq |W_{[k]}|,\ W_j > 0\}$ be arbitrary.
This means that $|W_j| \geq |W_{[k]}|$ and $W_j > 0$.
Since $W_j > 0$, we have $|W_j| = W_j$.
Therefore, $W_j \geq |W_{[k]}|$.
Hence, $j \in \{j :\ W_j \geq |W_{[k]}|\}$.

Both together give $\{j :\ W_j \geq |W_{[k]}|\} = \{j :\ |W_j| \geq |W_{[k]}|,\ W_j > 0\}$.
And since $k \in \{1, \ldots, p\}$ was arbitrary, we have shown this result for all $k \in \{1, \ldots, p\}$.

Then

$$
\begin{aligned}
& T_+ \\
& \qquad \text{By Definition 2.6} \\
&= \min\left\{t \in \mathcal{W} :\ \frac{1 + |\{j :\ W_j \leq -t\}|}{|\{j :\ W_j \geq t\}| \vee 1} \leq q\right\} \\
& \qquad \text{We know that } \mathcal{W} = \{|W_{[j]}| :\ j \in \{1, \ldots, p\}\}. \\
&= \min\left\{|W_{[k]}| :\ k \in \{1, \ldots, p\},\ \frac{1 + |\{j :\ W_j \leq -|W_{[k]}|\}|}{|\{j :\ W_j \geq |W_{[k]}|\}| \vee 1} \leq q\right\} \\
& \qquad \text{By Prep0.1 and Prep0.2, we know} \\
& \qquad \text{that } \{j :\ W_j \leq -|W_{[k]}|\} = \{j :\ |W_j| \geq |W_{[k]}|,\ W_j < 0\} \\
& \qquad \text{and } \{j :\ W_j \geq |W_{[k]}|\} = \{j :\ |W_j| \geq |W_{[k]}|,\ W_j > 0\}. \\
&= \min\left\{|W_{[k]}| :\ k \in \{1, \ldots, p\},\ \frac{1 + |\{j :\ |W_j| \geq |W_{[k]}|,\ W_j < 0\}|}{|\{j :\ |W_j| \geq |W_{[k]}|,\ W_j > 0\}| \vee 1} \leq q\right\} \\
& \qquad \text{By definition of } A(k) \text{ and } B(k) \\
&= \min\left\{|W_{[k]}| :\ k \in \{1, \ldots, p\},\ \frac{1 + A(k)}{B(k) \vee 1} \leq q\right\}.
\end{aligned}
$$

Set

$$K := \min\left\{k :\ k \in \{1, \ldots, p\},\ \frac{1 + A(k)}{B(k) \vee 1} \leq q\right\}.$$

Since $T_+ < \infty$, we know that at least one $k \in \{1, \ldots, p\}$ satisfies the inequality in the definition of K.
We know from above that T_+ is equal to the smallest $|W_{[k]}|$ such that this inequality is satisfied.
And we know that $|W_{[1]}| < |W_{[2]}| < \ldots < |W_{[p]}|$.
Therefore,

$$T_+ = |W_{[K]}|.$$

Side note: If we would relax the assumption $|W_{[1]}| < |W_{[2]}| < \ldots < |W_{[p]}|$ to $|W_{[1]}| \leq |W_{[2]}| \leq \ldots \leq |W_{[p]}|$, then $A(k)$, $B(k)$ and therefore K would not be unique in general, as a small example shows:

Take $p = 3$ with $W_1 = 3$, $W_2 = -1$ and $W_3 = -3$. Then we can choose if we write $|W_2| \leq |W_1| \leq |W_3|$ or $|W_2| \leq |W_3| \leq |W_1|$.

Case1 $W_{[1]} = W_2$, $W_{[2]} = W_1$, $W_{[3]} = W_3$:

Here, we have
$A(3) = |\{j :\ j \in \{3\},\ W_{[j]} < 0\}| = |\{3\}| = 1$ and
$B(3) = |\{j :\ j \in \{3\},\ W_{[j]} > 0\}| = |\emptyset| = 0$.
Hence, in particular,

$$\frac{1 + A(3)}{B(3) \vee 1} = \frac{1 + 1}{0 \vee 1} = 2.$$

Case2 $W_{[1]} = W_2$, $W_{[2]} = W_3$, $W_{[3]} = W_1$:

Here, we have
$A(3) = |\left\{j : \ j \in \{3\}, \ W_{[j]} < 0\right\}| = |\emptyset| = 0$ and
$B(3) = |\left\{j : \ j \in \{3\}, \ W_{[j]} > 0\right\}| = |\{3\}| = 1.$
Hence, in particular,

$$\frac{1 + A(3)}{B(3) \vee 1} = \frac{1 + 0}{1 \vee 1} = 1.$$

To continue the proof, note that

$$\begin{aligned}
&\hat{S} \\
&= \{j : \ W_j \geq T_+\} \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{We know that } T_+ = |W_{[K]}|. \\
&= \left\{j : \ W_j \geq |W_{[K]}|\right\} \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{By the same argument as in Prep0.2} \\
&= \left\{j : \ |W_j| \geq |W_{[K]}|, \ W_j > 0\right\} \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{By the same argument as above, where} \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{we introduced a different formula for } B(k). \\
&= \left\{j : \ j \in \{K, \ldots, p\}, \ W_{[j]} > 0\right\}.
\end{aligned}$$

Prep1.1 wts $\dfrac{1 + A(K)}{B(K) \vee 1} \leq q$:

We know that

$$K = \min\left\{k : \ k \in \{1, \ldots, p\}, \ \frac{1 + A(k)}{B(k) \vee 1} \leq q\right\}.$$

Therefore, K is one such $k \in \{1, \ldots, p\}$ satisfying this inequality.

Define

$$V^+(K) := |\left\{j : \ j \in \{K, \ldots, p\}, \ W_{[j]} > 0, \ \beta_{[j]} = 0\right\}|,$$

$$V^-(K) := |\left\{j : \ j \in \{K, \ldots, p\}, \ W_{[j]} < 0, \ \beta_{[j]} = 0\right\}|.$$

Part1 wts FDP $\leq q \cdot \dfrac{V^+(K)}{1 + V^-(K)}$:

FDP

$$\begin{aligned}
&\qquad\qquad\qquad\qquad\qquad\qquad \text{By Definition 2.7 i)} \\
&= \frac{|\left\{j : \ \beta_{[j]} = 0 \text{ and } j \in \hat{S}\right\}|}{|\hat{S}| \vee 1} \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{We know that } \hat{S} = \left\{j : \ j \in \{K, \ldots, p\}, \ W_{[j]} > 0\right\}. \\
&= \frac{|\left\{j : \ j \in \{K, \ldots, p\}, \ W_{[j]} > 0, \ \beta_{[j]} = 0\right\}|}{|\left\{j : \ j \in \{K, \ldots, p\}, \ W_{[j]} > 0\right\}| \vee 1} \\
&= \frac{|\left\{j : \ j \in \{K, \ldots, p\}, \ W_{[j]} > 0, \ \beta_{[j]} = 0\right\}|}{|\left\{j : \ j \in \{K, \ldots, p\}, \ W_{[j]} > 0\right\}| \vee 1} \cdot 1
\end{aligned}$$

We multiply with $\frac{1+A(K)}{1+A(K)}$.

$$= \frac{|\{j:\ j\in\{K,\dots,p\},\ W_{[j]}>0,\ \beta_{[j]}=0\}|}{|\{j:\ j\in\{K,\dots,p\},\ W_{[j]}>0\}|\vee 1}\cdot\frac{1+|\{j:\ j\in\{K,\dots,p\},\ W_{[j]}<0\}|}{1+|\{j:\ j\in\{K,\dots,p\},\ W_{[j]}<0\}|}$$

$$= \frac{1+|\{j:\ j\in\{K,\dots,p\},\ W_{[j]}<0\}|}{|\{j:\ j\in\{K,\dots,p\},\ W_{[j]}>0\}|\vee 1}\cdot\frac{|\{j:\ j\in\{K,\dots,p\},\ W_{[j]}>0,\ \beta_{[j]}=0\}|}{1+|\{j:\ j\in\{K,\dots,p\},\ W_{[j]}<0\}|}$$

$$= \frac{1+A(K)}{B(K)\vee 1}\cdot\frac{V^+(K)}{1+|\{j:\ j\in\{K,\dots,p\},\ W_{[j]}<0\}|}$$

All terms are nonnegative and note that $|\{j:\ j\in\{K,\dots,p\},\ W_{[j]}<0\}|\geq|\{j:\ j\in\{K,\dots,p\},\ W_{[j]}<0,\ \beta_{[j]}=0\}|$.

$$\leq \frac{1+A(K)}{B(K)\vee 1}\cdot\frac{V^+(K)}{1+|\{j:\ j\in\{K,\dots,p\},\ W_{[j]}<0,\ \beta_{[j]}=0\}|}$$

$$= \frac{1+A(K)}{B(K)\vee 1}\cdot\frac{V^+(K)}{1+V^-(K)}$$

By Prep1.1

$$\leq q\cdot\frac{V^+(K)}{1+V^-(K)}.$$

We now introduce martingale results.

Let $\mathfrak{F}=\{\mathcal{F}_k:\ k\in\{1,\dots,p\}\}$ be a filtration with sub-σ-algebras

$$\mathcal{F}_k=\sigma\left(A(l),B(l),V^+(l),V^-(l)\ \forall l\in\{1,\dots,k\};\ \mathbb{1}_{\{\beta_{[j]}=0\}}\ \forall j\in\{1,\dots,p\}\right).$$

Note that $\mathfrak{F}$ is a filtration, because $\mathcal{F}_1\subseteq\ldots\subseteq\mathcal{F}_p$.

Part2 wts K is a bounded stopping time w.r.t. $\mathfrak{F}$:

Note that $K\in\{1,\dots,p\}$. Therefore, K is bounded.
It remains to show that K is a stopping time.
This means we have to show $\{K=t\}\in\mathcal{F}_t$ for all $t\in\{1,\dots,p\}$.

Remember that

$$K=\min\left\{k:\ k\in\{1,\dots,p\},\ \frac{1+A(k)}{B(k)\vee 1}\leq q\right\}.$$

We see that K is the minimal $k\in\{1,\dots,p\}$, such that a specific inequality is satisfied.

First, observe that $\{K=1\}\in\mathcal{F}_1$:

Note that by definition of $\mathcal{F}_1$, we know $A(1)$, $B(1)$.
Hence, we can tell if the inequality in the definition of K is satisfied.
In other words, we can tell if $K=1$.
Therefore, $\{K=1\}\in\mathcal{F}_1$.

Second, observe that $\{K=1\},\{K=2\}\in\mathcal{F}_2$:

Note that by definition of $\mathcal{F}_2$, we know $A(l)$, $B(l)$ for $l\in\{1,\dots,2\}$.
So, we know $A(1)$, $B(1)$, $A(2)$ and $B(2)$.
Knowing $A(1)$ and $B(1)$ tells us if the inequality in the definition of K is satisfied for $k=1$.
Therefore, $\{K=1\}\in\mathcal{F}_2$.
Knowing $A(2)$ and $B(2)$ tells us if the inequality in the definition of K is satisfied for $k=2$.
Therefore, $\{K=2\}\in\mathcal{F}_2$.

By the same arguments for $3, \ldots, p$, we get overall: $\{K = t\} \in \mathcal{F}_t$ for all $t \in \{1, \ldots, p\}$.

Define $M(k) = \dfrac{V^+(k)}{1 + V^-(k)}$ for $k \in \{1, \ldots, p\}$.

Part3 wts $M(k)$ is a martingale w.r.t. $\mathfrak{F}$:

Part3.1 wts $M(k)$ is adapted to $\mathfrak{F}$:

Let $k \in \{1, \ldots, p\}$ be arbitrary.
We have to show that $M(k)$ is measurable w.r.t. $\mathcal{F}_k$.
By definition of $\mathcal{F}_k$, $V^+(k)$ and $V^-(k)$ are $\mathcal{F}_k$-measurable.
Note that constant functions are measureable.
We know that the sum of two measurable functions is measurable.
Hence, $1 + V^-(k)$ is $\mathcal{F}_k$-measurable.
And we know that the quotient of two measurable functions is measurable.
Therefore, $M(k)$ is $\mathcal{F}_k$-measurable.

Part3.2 wts $\mathbb{E}(|M(k)|) < \infty$ for all $k \in \{1, \ldots, p\}$:

Let $k \in \{1, \ldots, p\}$.

$$\begin{aligned}
&\mathbb{E}(|M(k)|) && \\
&&& \text{By definition of } M(k) \\
&= \mathbb{E}\left(\left|\frac{V^+(k)}{1 + V^-(k)}\right|\right) && \\
&&& \text{We can omit the absolute value, because} \\
&&& \text{the term inside is always nonnegative} \\
&= \mathbb{E}\left(\frac{V^+(k)}{1 + V^-(k)}\right) && \\
&&& \text{We know that } j \in \{1, \ldots, p\}. \text{ Therefore,} \\
&&& V^+(k) \leq p \text{ and } V^-(k) \geq 0 \text{ for all } k. \\
&&& \text{So, } \frac{V^+(k)}{1 + V^-(k)} \leq \frac{p}{1 + 0} = p. \\
&\leq \mathbb{E}(p) && \\
&= p && \\
&< \infty. &&
\end{aligned}$$

Part3.3 wts $\mathbb{E}[M(k+1) \mid \mathcal{F}_k] = M(k)$ for all $k \in \{1, \ldots, p-1\}$:

Let $k \in \{1, \ldots, p-1\}$.

Prep3.3.1 wts $M(k+1) = M(k)$ on the event $\{\beta_{[k]} \neq 0\}$:

$$
\begin{aligned}
& M(k+1) \\
& \qquad \text{By definition of } M(k+1) \\
&= \frac{V^+(k+1)}{1+V^-(k+1)} \\
& \qquad \text{By definition of } V^+(k+1) \text{ and } V^-(k+1) \\
&= \frac{|\left\{j:\ j \in \{k+1,\ldots,p\},\ W_{[j]}>0,\ \beta_{[j]}=0\right\}|}{1+|\left\{j:\ j \in \{k+1,\ldots,p\},\ W_{[j]}<0,\ \beta_{[j]}=0\right\}|} \\
& \qquad \text{The counts in numerator and denominator} \\
& \qquad \text{do not change if we start with } k \text{ instead of } k+1 \\
& \qquad \text{since we know that } \beta_{[k]} \neq 0. \\
&= \frac{|\left\{j:\ j \in \{k,\ldots,p\},\ W_{[j]}>0,\ \beta_{[j]}=0\right\}|}{1+|\left\{j:\ j \in \{k,\ldots,p\},\ W_{[j]}<0,\ \beta_{[j]}=0\right\}|} \\
& \qquad \text{By definition of } V^+(k) \text{ and } V^-(k) \\
&= \frac{V^+(k)}{1+V^-(k)} \\
& \qquad \text{By definition of } M(k) \\
&= M(k).
\end{aligned}
$$

Prep3.3.2 wts $M(k+1) = \dfrac{V^+(k)-1}{1+V^-(k)}$ on the event $\left\{\beta_{[k]}=0,\ W_{[k]}>0\right\}$:

$$
\begin{aligned}
& M(k+1) \\
& \qquad \text{By definition of } M(k+1) \\
&= \frac{V^+(k+1)}{1+V^-(k+1)} \\
& \qquad \text{By definition of } V^+(k+1) \text{ and } V^-(k+1) \\
&= \frac{|\left\{j:\ j \in \{k+1,\ldots,p\},\ W_{[j]}>0,\ \beta_{[j]}=0\right\}|}{1+|\left\{j:\ j \in \{k+1,\ldots,p\},\ W_{[j]}<0,\ \beta_{[j]}=0\right\}|} \\
& \qquad \text{Since we consider the event } \left\{\beta_{[k]}=0,\ W_{[k]}>0\right\}, \\
& \qquad \text{the numerator will increase by 1 if we start} \\
& \qquad \text{with } k \text{ instead of } k+1. \text{ The denominator is not affected.} \\
&= \frac{|\left\{j:\ j \in \{k,\ldots,p\},\ W_{[j]}>0,\ \beta_{[j]}=0\right\}|-1}{1+|\left\{j:\ j \in \{k,\ldots,p\},\ W_{[j]}<0,\ \beta_{[j]}=0\right\}|} \\
& \qquad \text{By definition of } V^+(k) \text{ and } V^-(k) \\
&= \frac{V^+(k)-1}{1+V^-(k)}.
\end{aligned}
$$

Prep3.3.3 wts $M(k+1) = \dfrac{V^+(k)}{V^-(k)}$ on the event $\left\{\beta_{[k]}=0,\ W_{[k]}<0\right\}$:

$$
\begin{aligned}
& M(k+1) \\
& \qquad \text{By definition of } M(k+1) \\
&= \frac{V^+(k+1)}{1+V^-(k+1)} \\
& \qquad \text{By definition of } V^+(k+1) \text{ and } V^-(k+1) \\
&= \frac{|\left\{j:\ j \in \{k+1,\ldots,p\},\ W_{[j]}>0,\ \beta_{[j]}=0\right\}|}{1+|\left\{j:\ j \in \{k+1,\ldots,p\},\ W_{[j]}<0,\ \beta_{[j]}=0\right\}|}
\end{aligned}
$$

Since we consider the event $\{\beta_{[k]} = 0,\ W_{[k]} < 0\}$,
the denominator will increase by 1 if we start
with k instead of $k+1$. The numerator is not affected.

$$= \frac{|\{j :\ j \in \{k, \dots, p\},\ W_{[j]} > 0,\ \beta_{[j]} = 0\}|}{1 + |\{j :\ j \in \{k, \dots, p\},\ W_{[j]} < 0,\ \beta_{[j]} = 0\}| - 1}$$

By definition of $V^+(k)$ and $V^-(k)$

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

Note that $V^-(k) > 0$ in this case, because we consider the event $\{\beta_{[k]} = 0,\ W_{[k]} < 0\}$ and hence, $V^-(k) \geq 1$.

Prep3.3.4 wts $\mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] = \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \frac{V^+(k)}{V^+(k)+V^-(k)}$:

Note that we can write $V^+(k) = \sum_{j=k}^{p} \mathbb{1}_{\{\beta_{[j]}=0\}} \cdot \mathbb{1}_{\{W_{[j]}>0\}}$.
Conditional expected value applied on both sides gives

$$\mathbb{E}\left[V^+(k) \mid \mathcal{F}_k\right] = \mathbb{E}\left[\sum_{j=k}^{p} \mathbb{1}_{\{\beta_{[j]}=0\}} \cdot \mathbb{1}_{\{W_{[j]}>0\}} \mid \mathcal{F}_k\right].$$

Since $V^+(k), \{\beta_{[j]} = 0\} \in \mathcal{F}_k$ for all $j \in \{1, \dots, p\}$, we can pull these terms out and get

$$V^+(k) \cdot \mathbb{E}\left[1 \mid \mathcal{F}_k\right] = \sum_{j=k}^{p} \mathbb{1}_{\{\beta_{[j]}=0\}} \cdot \mathbb{E}\left[\mathbb{1}_{\{W_{[j]}>0\}} \mid \mathcal{F}_k\right].$$

Hence,

$$V^+(k) = \sum_{j=k}^{p} \mathbb{1}_{\{\beta_{[j]}=0\}} \cdot \mathbb{P}\left[W_{[j]} > 0 \mid \mathcal{F}_k\right].$$

By Lemma 2.11 i) we know for null $[j]$: $\text{sign}(W_{[j]})$ are independent and $\mathbb{P}\left(\text{sign}(W_{[j]}) = -1\right) = \mathbb{P}\left(\text{sign}(W_{[j]}) = +1\right) = \frac{1}{2}$.
The latter can be written as $\mathbb{P}\left(W_{[j]} < 0\right) = \mathbb{P}\left(W_{[j]} > 0\right) = \frac{1}{2}$.
In other words, the signs of the null FX-Knockoff feature statistics are i.i.d. and hence exchangeable.
By definition of $\mathcal{F}_k$, all nulls are known.
Therefore,

$$\mathbb{1}_{\{\beta_{[j]}=0\}} \cdot \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] = \mathbb{1}_{\{\beta_{[j]}=0\}} \cdot \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{P}\left[W_{[j]} > 0 \mid \mathcal{F}_k\right]$$

for all $j \in \{k, \dots, p\}$.

And note that

$\mathbb{1}_{\{\beta_{[k]}=0\}} \cdot V^+(k)$

We know from above

that $V^+(k) = \sum_{j=k}^{p} \mathbb{1}_{\{\beta_{[j]}=0\}} \cdot \mathbb{P}\left[W_{[j]} > 0 \mid \mathcal{F}_k\right]$.

$$= \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \sum_{j=k}^{p} \mathbb{1}_{\{\beta_{[j]}=0\}} \cdot \mathbb{P}\left[W_{[j]} > 0 \mid \mathcal{F}_k\right]$$

$$= \sum_{j=k}^{p} \mathbb{1}_{\{\beta_{[j]}=0\}} \cdot \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{P}\left[W_{[j]} > 0 \mid \mathcal{F}_k\right]$$

We know that $\mathbb{1}_{\{\beta_{[j]}=0\}} \cdot \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] = \mathbb{1}_{\{\beta_{[j]}=0\}} \cdot \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{P}\left[W_{[j]} > 0 \mid \mathcal{F}_k\right]$.

$$= \sum_{j=k}^{p} \mathbb{1}_{\{\beta_{[j]}=0\}} \cdot \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right]$$

$$= \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] \cdot \sum_{j=k}^{p} \mathbb{1}_{\{\beta_{[j]}=0\}}$$

$$= \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] \cdot |\left\{j : \ j \in \{k, \ldots, p\}, \ \beta_{[j]} = 0\right\}|$$

Note that $V^+(k) + V^-(k) = |\left\{j : \ j \in \{k, \ldots, p\}, \ \beta_{[j]} = 0\right\}|$.

$$= \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] \cdot \left(V^+(k) + V^-(k)\right).$$

Rearranged gives

$$\mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] = \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \frac{V^+(k)}{V^+(k) + V^-(k)}.$$

Therefore,

$$\mathbb{E}\left[M(k+1) \mid \mathcal{F}_k\right]$$

$$= \mathbb{E}\left[M(k+1) \cdot \mathbb{1}_{\{\beta_{[k]}\neq 0\} \cup \{\beta_{[k]}=0\}} \mid \mathcal{F}_k\right]$$

Note that $\mathbb{1}_{C \cup D} = \mathbb{1}_C + \mathbb{1}_D$ for disjoint sets C, D.

$$= \mathbb{E}\left[M(k+1) \cdot \left(\mathbb{1}_{\{\beta_{[k]}\neq 0\}} + \mathbb{1}_{\{\beta_{[k]}=0\}}\right) \mid \mathcal{F}_k\right]$$

$$= \mathbb{E}\left[M(k+1) \cdot \left(\mathbb{1}_{\{\beta_{[k]}\neq 0\}} + \mathbb{1}_{\{\beta_{[k]}=0,\, W_{[k]}>0\} \cup \{\beta_{[k]}=0,\, W_{[k]}<0\}}\right) \mid \mathcal{F}_k\right]$$

Note that $\mathbb{1}_{C \cup D} = \mathbb{1}_C + \mathbb{1}_D$ for disjoint sets C, D.

$$= \mathbb{E}\left[M(k+1) \cdot \left(\mathbb{1}_{\{\beta_{[k]}\neq 0\}} + \mathbb{1}_{\{\beta_{[k]}=0,\, W_{[k]}>0\}} + \mathbb{1}_{\{\beta_{[k]}=0,\, W_{[k]}<0\}}\right) \mid \mathcal{F}_k\right]$$

$$= \mathbb{E}\left[M(k+1) \cdot \mathbb{1}_{\{\beta_{[k]}\neq 0\}} + M(k+1) \cdot \mathbb{1}_{\{\beta_{[k]}=0,\, W_{[k]}>0\}} + M(k+1) \cdot \mathbb{1}_{\{\beta_{[k]}=0,\, W_{[k]}<0\}} \mid \mathcal{F}_k\right]$$

$$= \mathbb{E}\left[M(k+1) \cdot \mathbb{1}_{\{\beta_{[k]}\neq 0\}} \mid \mathcal{F}_k\right]$$
$$+ \mathbb{E}\left[M(k+1) \cdot \mathbb{1}_{\{\beta_{[k]}=0,\, W_{[k]}>0\}} \mid \mathcal{F}_k\right]$$
$$+ \mathbb{E}\left[M(k+1) \cdot \mathbb{1}_{\{\beta_{[k]}=0,\, W_{[k]}<0\}} \mid \mathcal{F}_k\right]$$

By Prep3.3.1, Prep3.3.2 and Prep3.3.3

$$= \mathbb{E}\left[M(k) \cdot \mathbb{1}_{\{\beta_{[k]}\neq 0\}} \mid \mathcal{F}_k\right]$$
$$+ \mathbb{E}\left[\frac{V^+(k)-1}{1+V^-(k)} \cdot \mathbb{1}_{\{\beta_{[k]}=0,\, W_{[k]}>0\}} \mid \mathcal{F}_k\right]$$
$$+ \mathbb{E}\left[\frac{V^+(k)}{V^-(k)} \cdot \mathbb{1}_{\{\beta_{[k]}=0,\, W_{[k]}<0\}} \mid \mathcal{F}_k\right]$$

Note that $\mathbb{1}_{C \cap D} = \mathbb{1}_C \cdot \mathbb{1}_D$ for sets C, D.

$$= \mathbb{E}\left[M(k) \cdot \mathbb{1}_{\{\beta_{[k]}\neq 0\}} \mid \mathcal{F}_k\right]$$
$$+ \mathbb{E}\left[\frac{V^+(k)-1}{1+V^-(k)} \cdot \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{1}_{\{W_{[k]}>0\}} \mid \mathcal{F}_k\right]$$
$$+ \mathbb{E}\left[\frac{V^+(k)}{V^-(k)} \cdot \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{1}_{\{W_{[k]}<0\}} \mid \mathcal{F}_k\right]$$

By definition of $\mathcal{F}_k$, we know that $V^+(k), V^-(k), \left\{\beta_{[k]} = 0\right\} \in \mathcal{F}_k$.
Since complements of elements of a σ-algebra are
again in that σ-algebra, $\left\{\beta_{[k]} = 0\right\}^c = \left\{\beta_{[k]} \neq 0\right\} \in \mathcal{F}_k$.
Hence we can pull out these terms.

$$
\begin{aligned}
&= M(k) \cdot \mathbb{1}_{\{\beta_{[k]} \neq 0\}} \cdot \mathbb{E}\left[1 \mid \mathcal{F}_k\right] \\
&\quad + \frac{V^+(k)-1}{1+V^-(k)} \cdot \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{E}\left[\mathbb{1}_{\{W_{[k]}>0\}} \mid \mathcal{F}_k\right] \\
&\quad + \frac{V^+(k)}{V^-(k)} \cdot \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{E}\left[\mathbb{1}_{\{W_{[k]}<0\}} \mid \mathcal{F}_k\right] \\
&= M(k) \cdot \mathbb{1}_{\{\beta_{[k]} \neq 0\}} \\
&\quad + \frac{V^+(k)-1}{1+V^-(k)} \cdot \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] \\
&\quad + \frac{V^+(k)}{V^-(k)} \cdot \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \mathbb{P}\left[W_{[k]} < 0 \mid \mathcal{F}_k\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{By Prep3.3.4} \\
&= M(k) \cdot \mathbb{1}_{\{\beta_{[k]} \neq 0\}} \\
&\quad + \frac{V^+(k)-1}{1+V^-(k)} \cdot \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \frac{V^+(k)}{V^+(k)+V^-(k)} \\
&\quad + \frac{V^+(k)}{V^-(k)} \cdot \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \left(1 - \frac{V^+(k)}{V^+(k)+V^-(k)}\right) \\
&= M(k) \cdot \mathbb{1}_{\{\beta_{[k]} \neq 0\}} \\
&\quad + \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \frac{V^+(k)-1}{1+V^-(k)} \cdot \frac{V^+(k)}{V^+(k)+V^-(k)} \\
&\quad + \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \frac{V^+(k)}{V^-(k)} \cdot \left(\frac{V^+(k)+V^-(k)}{V^+(k)+V^-(k)} - \frac{V^+(k)}{V^+(k)+V^-(k)}\right) \\
&= M(k) \cdot \mathbb{1}_{\{\beta_{[k]} \neq 0\}} \\
&\quad + \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \frac{V^+(k)-1}{1+V^-(k)} \cdot \frac{V^+(k)}{V^+(k)+V^-(k)} \\
&\quad + \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \frac{V^+(k)}{V^-(k)} \cdot \frac{V^-(k)}{V^+(k)+V^-(k)} \\
&= M(k) \cdot \mathbb{1}_{\{\beta_{[k]} \neq 0\}} + \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \left(\frac{V^+(k)-1}{1+V^-(k)} \cdot \frac{V^+(k)}{V^+(k)+V^-(k)} + \frac{V^-(k)}{V^-(k)} \cdot \frac{V^+(k)}{V^+(k)+V^-(k)}\right) \\
&= M(k) \cdot \mathbb{1}_{\{\beta_{[k]} \neq 0\}} + \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \left(\frac{V^+(k)}{V^+(k)+V^-(k)} \cdot \left(\frac{V^+(k)-1}{1+V^-(k)} + 1\right)\right) \\
&= M(k) \cdot \mathbb{1}_{\{\beta_{[k]} \neq 0\}} + \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \left(\frac{V^+(k)}{V^+(k)+V^-(k)} \cdot \left(\frac{V^+(k)-1}{1+V^-(k)} + \frac{1+V^-(k)}{1+V^-(k)}\right)\right) \\
&= M(k) \cdot \mathbb{1}_{\{\beta_{[k]} \neq 0\}} + \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \left(\frac{V^+(k)}{V^+(k)+V^-(k)} \cdot \frac{V^+(k)+V^-(k)}{1+V^-(k)}\right) \\
&= M(k) \cdot \mathbb{1}_{\{\beta_{[k]} \neq 0\}} + \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot \frac{V^+(k)}{1+V^-(k)} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{By definition of } M(k) \\
&= M(k) \cdot \mathbb{1}_{\{\beta_{[k]} \neq 0\}} + \mathbb{1}_{\{\beta_{[k]}=0\}} \cdot M(k) \\
&= M(k) \cdot \left(\mathbb{1}_{\{\beta_{[k]} \neq 0\}} + \mathbb{1}_{\{\beta_{[k]}=0\}}\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{Note that } \mathbb{1}_{C \cup D} = \mathbb{1}_C + \mathbb{1}_D \text{ for disjoint sets } C, D. \\
&= M(k) \cdot \mathbb{1}_{\{\beta_{[k]} \neq 0\} \cup \{\beta_{[k]}=0\}} \\
&= M(k).
\end{aligned}
$$

Since $k \in \{1, \ldots, p-1\}$ was arbitrary, we have shown $\mathbb{E}\left[M(k+1) \mid \mathcal{F}_k\right] = M(k)$ for all $k \in \{1, \ldots, p-1\}$.

Define $N_0 = |\left\{j \in \{1, \ldots, p\} : \ \beta_{[j]} = 0\right\}|$ as the total number of nulls.

Prep4.1 wts $V^+(1) \sim \text{Binomial}\left(N_0, \frac{1}{2}\right)$:

Note that we can write

$$V^+(1) = |\left\{j : \ j \in \{1,\ldots,p\}\,,\ W_{[j]} > 0,\ \beta_{[j]} = 0\right\}| = |\left\{j : \ j \in \{1,\ldots,p\}\,,\ \text{sign}(W_{[j]}) = +1,\ \beta_{[j]} = 0\right\}|.$$

Define $U_j = \mathbb{1}_{\left\{\text{sign}(W_{[j]})=+1\right\}}$.
Note that $\mathbb{P}\left(U_j = 1\right) = \mathbb{P}\left(\text{sign}(W_{[j]}) = +1\right) = \frac{1}{2}$ for all j where $\beta_{[j]} = 0$ by Lemma 2.11 i).
By Lemma 2.11 i), $\text{sign}(W_{[j]})$ are independent for j where $\beta_{[j]} = 0$.
This means that U_j are independent random variables with $U_j \sim \text{Bernoulli}\left(\frac{1}{2}\right)$ for all j where $\beta_{[j]} = 0$.
Short: $U_j \overset{i.i.d.}{\sim} \text{Bernoulli}\left(\frac{1}{2}\right)$ for j where $\beta_{[j]} = 0$.
Note that $\sum_{\substack{j \in \{1,\ldots,p\}\,,\\ \beta_{[j]}=0}} U_j = V^+(1)$.
By definition of the Binomial-distribution, $V^+(1) \sim \text{Binomial}\left(N_0, \frac{1}{2}\right)$.

Prep4.2 wts $V^-(1) = N_0 - V^+(1)$:

$$\begin{aligned}
&V^-(1)\\
&= |\left\{j : \ j \in \{1,\ldots,p\}\,,\ W_{[j]} < 0,\ \beta_{[j]} = 0\right\}| \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{Note that } N_0 = |\left\{j \in \{1,\ldots,p\} : \ \beta_{[j]} = 0\right\}|.\\
&= N_0 - |\left\{j : \ j \in \{1,\ldots,p\}\,,\ W_{[j]} > 0,\ \beta_{[j]} = 0\right\}| \\
&= N_0 - V^+(1).
\end{aligned}$$

Part4 wts $\mathbb{E}\left[\frac{V^+(K)}{1+V^-(K)}\right] \leq 1$:

$$\begin{aligned}
&\mathbb{E}\left[\frac{V^+(K)}{1+V^-(K)}\right]\\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{By Part2, we know that } K \text{ is a bounded stopping time.}\\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{By Part3, we know that } \frac{V^+(k)}{1+V^-(k)} \text{ is a martingale.}\\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{Therefore, we can use Lemma A.4 i)}\\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{(optional stopping theorem for martingales).}\\
&= \mathbb{E}\left[\frac{V^+(1)}{1+V^-(1)}\right]\\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{By Prep4.2}\\
&= \mathbb{E}\left[\frac{V^+(1)}{1+N_0-V^+(1)}\right]
\end{aligned}$$

In general, it holds for a discrete random variable Y and function g that $\mathbb{E}(g(Y)) = \sum_y g(y) \cdot \mathbb{P}(Y = y)$. In our case, $g(x) := \frac{x}{1 + N_0 - x}$ and $Y := V^+(1)$. Note that $N_0 \leq p$ and $Y \in \{0, \ldots, N_0\}$.

$$= \sum_{i=0}^{N_0} g(i) \cdot \mathbb{P}(Y = i)$$

By definition of g

$$= \sum_{i=0}^{N_0} \frac{i}{1 + N_0 - i} \cdot \mathbb{P}(Y = i)$$

By Prep4.1, $V^+(1) \sim \text{Binomial}\left(N_0, \frac{1}{2}\right)$.

$$= \sum_{i=0}^{N_0} \frac{i}{1 + N_0 - i} \cdot \binom{N_0}{i} \cdot \left(\frac{1}{2}\right)^i \cdot \left(1 - \frac{1}{2}\right)^{N_0 - i}$$

We can start the sum at $i = 1$.

$$= \sum_{i=1}^{N_0} \frac{i}{1 + N_0 - i} \cdot \binom{N_0}{i} \cdot \left(\frac{1}{2}\right)^i \cdot \left(1 - \frac{1}{2}\right)^{N_0 - i}$$

$$= \sum_{i=1}^{N_0} \frac{i}{1 + N_0 - i} \cdot \frac{N_0!}{i! \cdot (N_0 - i)!} \cdot \left(\frac{1}{2}\right)^i \cdot \left(\frac{1}{2}\right)^{N_0 - i}$$

$$= \sum_{i=1}^{N_0} \frac{1}{1 + N_0 - i} \cdot \frac{N_0!}{(i-1)! \cdot (N_0 - i)!} \cdot \left(\frac{1}{2}\right)^{i-1} \cdot \left(\frac{1}{2}\right)^1 \cdot \left(\frac{1}{2}\right)^{N_0 - (i-1)} \cdot \left(\frac{1}{2}\right)^{-1}$$

Note that $1 + N_0 - i = N_0 - (i - 1)$.

$$= \sum_{i=1}^{N_0} \frac{N_0!}{(i-1)! \cdot (N_0 - (i-1))!} \cdot \left(\frac{1}{2}\right)^{i-1} \cdot \left(\frac{1}{2}\right)^{N_0 - (i-1)}$$

$$= \sum_{i=1}^{N_0} \binom{N_0}{i-1} \cdot \left(\frac{1}{2}\right)^{i-1} \cdot \left(\frac{1}{2}\right)^{N_0 - (i-1)}$$

Substitution: set $j = i - 1$.

$$= \sum_{j=0}^{N_0 - 1} \underbrace{\binom{N_0}{j} \cdot \left(\frac{1}{2}\right)^j \cdot \left(\frac{1}{2}\right)^{N_0 - j}}_{= \mathbb{P}(Y = j)}$$

The sum of probabilities the random variable Y can achieve has to be smaller than 1.

$$\leq 1.$$

Part5 wts FDR $\leq q$:

FDR

By Definition 2.7 ii)

$$= \mathbb{E}[\text{FDP}]$$

$$\text{By Part1, we know that FDP} \leq q \cdot \frac{V^+(K)}{1+V^-(K)}$$

$$\leq \mathbb{E}\left[q \cdot \frac{V^+(K)}{1+V^-(K)}\right]$$

$$= q \cdot \mathbb{E}\left[\frac{V^+(K)}{1+V^-(K)}\right]$$

$$\text{By Part4, we know that } \mathbb{E}\left[\frac{V^+(K)}{1+V^-(K)}\right] \leq 1.$$

$$\leq q \cdot 1$$

$$= q.$$

As $q \in [0,1]$ was arbitrary but fixed, we have shown the result for all $q \in [0,1]$.

□

Comments:
i) Note that the assumptions of this theorem are the same as in Theorem 2.12, except for the fact that T_+ is used instead of T. This is why both proofs are almost the same.
ii) Also note that the threshold T_+ chosen by the FX-Knockoff+ method is always higher or equal than the threshold T chosen by the FX-Knockoff method. This means that FX-Knockoff+ is slightly more conservative (exact FDR-control at the cost of power).

Janson and Su (2016) presented a FX-Knockoff procedure for order statistics $W_{\pi(1)} \geq W_{\pi(2)} \cdots \geq W_{\pi(p)}$ (where $\pi(1), \ldots, \pi(p)$ is a permutation of $1, \ldots, p$) which controls the k-family wise error rate (k-FWER), a generalized version of the family wise error rate (Lehmann and Romano, 2005).
Katsevich and Sabatti (2019) introduced a multilayer knockoff filter which reduces the number of false discoveries at the cost of little power loss (compared to fixed-X Knockoff procedures).

2.3 Power

FDR control alone is not enough. If the power of a method is poor, FDR control is meaningless. To see this, consider a method which never selects any variables. This method would have FDR = 0, but the power would be zero as well.

For power analysis, Barber and Candes (2015) used simulations to compare FX-Knockoff and FX-Knockoff+ methods to other FDR-controlling procedures. First, permutation based methods were considered and authors noted, that such methods can have very poor control of FDR. E.g. if we construct $\tilde{\mathbf{X}}$, where

$$\tilde{\mathbf{X}}_{i,j} = \mathbf{X}_{\pi(i),j}$$

for some randomly chosen permutation π of the sample indices $\{1, \ldots, n\}$, the matrix $\tilde{\mathbf{X}}$ would satisfy $\tilde{\mathbf{X}}^T \cdot \tilde{\mathbf{X}} = \mathbf{X}^T \cdot \mathbf{X}$, but Lemmas 2.9 and 2.10 would not hold for $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]$ in a case where $\mathbf{X}$ displays non-vanishing correlations; and this can lead to very poor control of FDR.

For the rest of the power simulations, authors focussed on Benjamini-Hochberg procedures, mainly in a setting where $n = 3000,\ p = 1000$ and Gaussian noise. The conclusion was that FX-Knockoff and FX-Knockoff+ methods have better FDR and power performance than comparable Benjamini-Hochberg procedures. Nevertheless, power of both FX-Knockoff procedures declined as correlation of variables increased. The reason is that for highly correlated variables $\mathbf{X}_j$, the vector $\boldsymbol{s}$ has very small entries (close to zero). This means that FX-Knockoff vectors $\tilde{\mathbf{X}}_j$ and original variables $\mathbf{X}_j$ are highly correlated. This leads to poor power, because in such a setting, it is hard to distinguish between a non-null variable $\mathbf{X}_j$ and its fake version $\tilde{\mathbf{X}}_j$. If correlations of variables appear in groups, where the within-group correlations are high but between-group correlations are low, Dai and Barber (2016) present FX-group-Knockoffs for the fixed-X setting (where $n \geq p$).

Weinstein et al. (2017) provide theoretical power-results for methods similar to FX-Knockoff procedures when Lasso (Tibshirani, 1996) is used to create FX-Knockoff feature statistics and conclude that these methods asymptotically achieve close to optimal power with respect to an omniscient oracle.

2.4 Examples

2.4.1 Exact construction of FX Knockoffs

As mentioned in the comments after Definition 2.2, we first present answers to the question why $\boldsymbol{s}$ has to satisfy $\text{diag}(\boldsymbol{s}) \succeq \mathbf{0}$ and $2 \cdot \boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s}) \succeq \mathbf{0}$.

Lemma 2.14. *Let $\boldsymbol{s} \in \mathbb{R}^p$ be a vector. Then:*

$$\mathbf{G} := \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s}) \\ \boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s}) & \boldsymbol{\Sigma} \end{bmatrix}$$

is positive semidefinite if and only if $\text{diag}(\boldsymbol{s}) \succeq \mathbf{0}$ *and* $2 \cdot \boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s}) \succeq \mathbf{0}$*, where* $\boldsymbol{\Sigma} = \mathbf{X}^T \cdot \mathbf{X}$.

Proof:

Using Lemma A.5 for $\mathbf{G}$, we get that $\mathbf{G}$ is positive semidefinite if and only if
$\boldsymbol{\Sigma} - \big((\boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s})) \cdot \boldsymbol{\Sigma}^{-1} \cdot (\boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s}))\big) \succeq \mathbf{0}$.

And note:

$$\begin{aligned}
&\boldsymbol{\Sigma} - \big((\boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s})) \cdot \boldsymbol{\Sigma}^{-1} \cdot (\boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s}))\big) \\
&= \boldsymbol{\Sigma} - \left(\left(\underbrace{\boldsymbol{\Sigma} \cdot \boldsymbol{\Sigma}^{-1}}_{=\mathbf{I}} - \text{diag}(\boldsymbol{s}) \cdot \boldsymbol{\Sigma}^{-1} \right) \cdot (\boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s})) \right) \\
&= \boldsymbol{\Sigma} - \big(\big(\mathbf{I} - \text{diag}(\boldsymbol{s}) \cdot \boldsymbol{\Sigma}^{-1} \big) \cdot (\boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s})) \big) \\
&= \boldsymbol{\Sigma} - \left(\mathbf{I} \cdot \boldsymbol{\Sigma} - \mathbf{I} \cdot \text{diag}(\boldsymbol{s}) - \text{diag}(\boldsymbol{s}) \cdot \underbrace{\boldsymbol{\Sigma}^{-1} \cdot \boldsymbol{\Sigma}}_{=\mathbf{I}} + \text{diag}(\boldsymbol{s}) \cdot \boldsymbol{\Sigma}^{-1} \cdot \text{diag}(\boldsymbol{s}) \right)
\end{aligned}$$

$$
\begin{aligned}
&= \boldsymbol{\Sigma} - \left(\boldsymbol{\Sigma} - \operatorname{diag}(\boldsymbol{s}) - \operatorname{diag}(\boldsymbol{s}) + \operatorname{diag}(\boldsymbol{s}) \cdot \boldsymbol{\Sigma}^{-1} \cdot \operatorname{diag}(\boldsymbol{s})\right) \\
&= \boldsymbol{\Sigma} - \boldsymbol{\Sigma} + 2 \cdot \operatorname{diag}(\boldsymbol{s}) - \operatorname{diag}(\boldsymbol{s}) \cdot \boldsymbol{\Sigma}^{-1} \cdot \operatorname{diag}(\boldsymbol{s}) \\
&= 2 \cdot \operatorname{diag}(\boldsymbol{s}) - \operatorname{diag}(\boldsymbol{s}) \cdot \boldsymbol{\Sigma}^{-1} \cdot \operatorname{diag}(\boldsymbol{s}).
\end{aligned}
$$

Therefore,

$$
\mathbf{G} \succeq \mathbf{0} \quad \Longleftrightarrow \quad 2 \cdot \operatorname{diag}(\boldsymbol{s}) - \operatorname{diag}(\boldsymbol{s}) \cdot \boldsymbol{\Sigma}^{-1} \cdot \operatorname{diag}(\boldsymbol{s}) \succeq \mathbf{0}.
$$

Again by using Lemma A.5 for the positive semidefinite Schur complement on the right hand side, we get

$$
\mathbf{G} \succeq \mathbf{0} \quad \Longleftrightarrow \quad \begin{bmatrix} \boldsymbol{\Sigma} & \operatorname{diag}(\boldsymbol{s}) \\ \operatorname{diag}(\boldsymbol{s}) & 2 \cdot \operatorname{diag}(\boldsymbol{s}) \end{bmatrix} \succeq \mathbf{0}.
$$

By using Lemma A.6, we know that positive semidefiniteness of that matrix is equivalent to
i) $2 \cdot \operatorname{diag}(\boldsymbol{s}) \succeq \mathbf{0}$,
ii) $\operatorname{diag}(\boldsymbol{s}) = 2 \cdot \operatorname{diag}(\boldsymbol{s}) \cdot (2 \cdot \operatorname{diag}(\boldsymbol{s}))^{-} \cdot \operatorname{diag}(\boldsymbol{s})$,
iii) $\boldsymbol{\Sigma} - \operatorname{diag}(\boldsymbol{s}) \cdot (2 \cdot \operatorname{diag}(\boldsymbol{s}))^{-1} \cdot \operatorname{diag}(\boldsymbol{s}) \succeq \mathbf{0}$.

The second condition vanishes, because $2 \cdot \operatorname{diag}(\boldsymbol{s})$ is invertible and therefore, no special condition for generalized inverses is needed (condition ii) is always satisfied).

Note that i) is equivalent to $\operatorname{diag}(\boldsymbol{s}) \succeq \mathbf{0}$.
And iii) can be rewritten as $\boldsymbol{\Sigma} - \operatorname{diag}(\boldsymbol{s}) \cdot 2^{-1} \cdot \underbrace{(\operatorname{diag}(\boldsymbol{s}))^{-1} \cdot \operatorname{diag}(\boldsymbol{s})}_{=\mathbf{I}} = \boldsymbol{\Sigma} - \frac{1}{2} \cdot \operatorname{diag}(\boldsymbol{s}) \succeq \mathbf{0}$ which is equivalent to

$$
2 \cdot \boldsymbol{\Sigma} - \operatorname{diag}(\boldsymbol{s}) \succeq \mathbf{0}.
$$

Therefore,

$$
\mathbf{G} \succeq \mathbf{0} \quad \Longleftrightarrow \quad \operatorname{diag}(\boldsymbol{s}) \succeq \mathbf{0} \quad \text{and} \quad 2 \cdot \boldsymbol{\Sigma} - \operatorname{diag}(\boldsymbol{s}) \succeq \mathbf{0}.
$$

□

Comments:
i) A necessary and sufficient condition for $\tilde{\mathbf{X}}$ to exist ist that $\mathbf{G}$ is positive semidefinite (see Lemma 2.16). Lemma 2.14 shows that this is satisfied if and only if $\operatorname{diag}(\boldsymbol{s}) \succeq \mathbf{0}$ and $2 \cdot \boldsymbol{\Sigma} - \operatorname{diag}(\boldsymbol{s}) \succeq \mathbf{0}$.
ii) By Lemma 2.8 viii) and ix), we see that $\mathbf{X}_j^T \cdot \tilde{\mathbf{X}}_j = 1 - s_j$. To ensure that our method has high power, we should choose the entries of $\boldsymbol{s}$ as large as possible, so that a FX-Knockoff vector $\tilde{\mathbf{X}}_j$ is not too similar to the corresponding original variable $\mathbf{X}_j$.
iii) The assumption $\operatorname{diag}(\boldsymbol{s}) \succeq \mathbf{0}$ is satisfied if $\boldsymbol{s}$ is non-negative.
iv) The next example shows two methods to compute $\boldsymbol{s}$.

Example 2.15. Let $\mathbf{X}$ be a fixed design matrix with normalized columns $\mathbf{X}_j$; that is, $||\mathbf{X}_j||_2^2 = 1$ for all $j \in \{1, \ldots, p\}$. There are several ways to construct a vector $\boldsymbol{s}$ such that it obeys the assumptions $\operatorname{diag}(\boldsymbol{s}) \succeq \mathbf{0}$ and $2 \cdot \boldsymbol{\Sigma} - \operatorname{diag}(\boldsymbol{s}) \succeq \mathbf{0}$. Two possible constructions are:
i) ("equi-correlated construction"): for all j,

$$
s_j = \min\left\{2 \cdot \lambda_{\min}\left(\boldsymbol{\Sigma}\right), 1\right\}.
$$

ii) ("semidefinite program (SDP) construction"):

$$\text{minimize} \sum_j (1 - s_j) \quad \text{subject to} \quad 0 \leq s_j \leq 1,\ 2 \cdot \boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s}) \succeq \mathbf{0}.$$

Comments:
a) Note that $\lambda_{\min}(\boldsymbol{\Sigma})$ denotes the smallest eigenvalue of $\boldsymbol{\Sigma}$.
b) By Lemma A.11, we know that $\boldsymbol{\Sigma} - \lambda_{\min}(\boldsymbol{\Sigma}) \cdot \mathbf{I}_p \succeq \mathbf{0}$. Hence, $2 \cdot \boldsymbol{\Sigma} - 2 \cdot \lambda_{\min}(\boldsymbol{\Sigma}) \cdot \mathbf{I}_p \succeq \mathbf{0}$. This means that $\boldsymbol{s} = (2 \cdot \lambda_{\min}(\boldsymbol{\Sigma}), \ldots, 2 \cdot \lambda_{\min}(\boldsymbol{\Sigma}))^T$ would satisfy the assumption $2 \cdot \boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s}) \succeq \mathbf{0}$. And it would satisfy $\text{diag}(\boldsymbol{s}) \succeq \mathbf{0}$, because positive semidefiniteness of $\boldsymbol{\Sigma}$ implies $\lambda_{\min}(\boldsymbol{\Sigma}) \geq 0$. But to achieve higher power, we use a slightly different $\boldsymbol{s}$. By Lemma 2.8 viii) and ix), we know that $\tilde{\mathbf{X}}_j^T \cdot \mathbf{X}_j = 1 - s_j$. We would like to have $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ as orthogonal as possible (to get high power), which requires s_j be close to 1. This is why we use $s_j = \min\{2 \cdot \lambda_{\min}(\boldsymbol{\Sigma}), 1\}$ for all j. It implies that $\tilde{\mathbf{X}}_j^T \cdot \mathbf{X}_j$ is the same for all j which is why this construction is called "equi-correlated".
c) Computation of the equi-correlated construction is easy, because after finding $\lambda_{\min}(\boldsymbol{\Sigma})$ with some algorithm, we can calculate $s_j = \min\{2 \cdot \lambda_{\min}(\boldsymbol{\Sigma}), 1\}$ and set $\boldsymbol{s} = (s_j, \ldots, s_j)^T$.
d) The SDP construction minimizes the average correlation $\tilde{\mathbf{X}}_j^T \cdot \mathbf{X}_j = 1 - s_j$. This can be solved in practice by using SDP methods, e.g. the open source software DSDP as used in the R package *knockoff* (Sesia et al., 2018).
e) The next Lemma shows a method to construct valid FX-Knockoff matrices for a given vector $\boldsymbol{s}$.

Lemma 2.16. *Let $\tilde{\mathbf{U}} \in \mathbb{R}^{n \times p}$ be a matrix such that $\tilde{\mathbf{U}}^T \cdot \tilde{\mathbf{U}} = \mathbf{I}_p$ and $\tilde{\mathbf{U}}^T \cdot \mathbf{X} = \mathbf{0}$.*
Let $\mathbf{C}^T \cdot \mathbf{C} = 2 \cdot \text{diag}(\boldsymbol{s}) - \text{diag}(\boldsymbol{s}) \cdot \boldsymbol{\Sigma}^{-1} \cdot \text{diag}(\boldsymbol{s})$. Then

$$\tilde{\mathbf{X}} = \mathbf{X} - \mathbf{X} \cdot \boldsymbol{\Sigma}^{-1} \cdot \text{diag}(\boldsymbol{s}) + \tilde{\mathbf{U}} \cdot \mathbf{C}$$

satisfies the correlation structure given in Lemma 2.8 ix); that is, $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right] = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s}) \\ \boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s}) & \boldsymbol{\Sigma} \end{bmatrix}$,
where $\boldsymbol{\Sigma} = \mathbf{X}^T \cdot \mathbf{X}$.

Proof:

Prep1 wts $\tilde{\mathbf{U}}$ exists for $n \geq 2 \cdot p$:

Note that we always can apply QR-factorization to the rectangular matrix $\mathbf{X}$ (because $n \geq p$) such that $\mathbf{X} = [\mathbf{Q}_1, \mathbf{Q}_2] \cdot \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{bmatrix}$.
We are using the matrix $\mathbf{Q}_2$ (which has orthogonal columns) to construct $\tilde{\mathbf{U}}$:
- for $n = 2 \cdot p$ and if $\mathbf{X}$ has p linearly independent columns, $\mathbf{Q}_2$ is a $n \times (n - p) = n \times p$ matrix, so we take $\tilde{\mathbf{U}} := \mathbf{Q}_2$.
- for $n = 2 \cdot p$, and if $\mathbf{X}$ has less than p linearly independent columns, $\mathbf{Q}_2$ has n rows and more than p columns. We can choose p of them to construct $\tilde{\mathbf{U}}$. E.g. we can take the last p columns of $\mathbf{Q}_2$ as matrix $\tilde{\mathbf{U}}$.
- for $n > 2 \cdot p$, $\mathbf{Q}_2$ has n rows and more than p columns. We can again simply choose p of them to construct $\tilde{\mathbf{U}}$.

Of course, the second case is not relevant for our setting, because the assumption that $\mathbf{X}^T \cdot \mathbf{X}$ is invertible is equivalent to say that the $n \times p$ matrix $\mathbf{X}$ has p linearly independent columns. We discussed the second case just to point out, that invertibility of $\mathbf{X}^T \cdot \mathbf{X}$ is not needed to construct $\tilde{\mathbf{U}}$. In fact, the invertibility assumption is obstructive and the reason why the construction of such $\tilde{\mathbf{U}}$ is not possible for $n < 2 \cdot p$.

Prep2 wts $\mathbf{C}^T \cdot \mathbf{C}$ exists:

We proved in Lemma 2.14 that $2 \cdot \text{diag}(\boldsymbol{s}) - \text{diag}(\boldsymbol{s}) \cdot \boldsymbol{\Sigma}^{-1} \cdot \text{diag}(\boldsymbol{s}) \succeq \mathbf{0}$.
Then, by Lemma A.10, there is a Cholesky decomposition (not necessarily unique)

$$\mathbf{L} \cdot \mathbf{L}^T = 2 \cdot \text{diag}(\boldsymbol{s}) - \text{diag}(\boldsymbol{s}) \cdot \boldsymbol{\Sigma}^{-1} \cdot \text{diag}(\boldsymbol{s}).$$

By using QR-factorization on $\mathbf{L}^T$, this gives $\mathbf{L}^T = \mathbf{Q} \cdot \mathbf{R}$ and $\mathbf{L} = (\mathbf{Q} \cdot \mathbf{R})^T$ such that

$$(\mathbf{Q} \cdot \mathbf{R})^T \cdot \mathbf{Q} \cdot \mathbf{R} = 2 \cdot \text{diag}(\boldsymbol{s}) - \text{diag}(\boldsymbol{s}) \cdot \boldsymbol{\Sigma}^{-1} \cdot \text{diag}(\boldsymbol{s}).$$

We can now define $\mathbf{C} = \mathbf{Q} \cdot \mathbf{R}$ and get

$$\mathbf{C}^T \cdot \mathbf{C} = 2 \cdot \text{diag}(\boldsymbol{s}) - \text{diag}(\boldsymbol{s}) \cdot \boldsymbol{\Sigma}^{-1} \cdot \text{diag}(\boldsymbol{s}).$$

wts $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right] = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s}) \\ \boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s}) & \boldsymbol{\Sigma} \end{bmatrix}$, where $\boldsymbol{\Sigma} = \mathbf{X}^T \cdot \mathbf{X}$:

Note that

$$\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right] = \begin{bmatrix} \mathbf{X}^T \cdot \mathbf{X} & \mathbf{X}^T \cdot \tilde{\mathbf{X}} \\ \tilde{\mathbf{X}}^T \cdot \mathbf{X} & \tilde{\mathbf{X}}^T \cdot \tilde{\mathbf{X}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma} & \overbrace{\mathbf{X}^T \cdot \tilde{\mathbf{X}}}^{=:\mathbf{M}_1} \\ \underbrace{\tilde{\mathbf{X}}^T \cdot \mathbf{X}}_{=:\mathbf{M}_2} & \underbrace{\tilde{\mathbf{X}}^T \cdot \tilde{\mathbf{X}}}_{=:\mathbf{M}_3} \end{bmatrix}.$$

It remains to show that the matrices $\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3$ have the desired form.

Note first that because of $\tilde{\mathbf{U}}^T \cdot \mathbf{X} = \mathbf{0}$, by transposing both sides, we get $\mathbf{X}^T \cdot \tilde{\mathbf{U}} = \mathbf{0}$.
And note that because $\boldsymbol{\Sigma}$ is symmetric, $\boldsymbol{\Sigma}^{-1}$ is symmetric too.

wts $\mathbf{M}_1 = \boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s})$:

$$\begin{aligned}
&\mathbf{M}_1 \\
&= \mathbf{X}^T \cdot \tilde{\mathbf{X}} \\
&= \mathbf{X}^T \cdot \left(\mathbf{X} - \mathbf{X} \cdot \boldsymbol{\Sigma}^{-1} \cdot \text{diag}(\boldsymbol{s}) + \tilde{\mathbf{U}} \cdot \mathbf{C}\right) \\
&= \underbrace{\mathbf{X}^T \cdot \mathbf{X}}_{=\boldsymbol{\Sigma}} - \underbrace{\mathbf{X}^T \cdot \mathbf{X}}_{=\boldsymbol{\Sigma}} \cdot \boldsymbol{\Sigma}^{-1} \cdot \text{diag}(\boldsymbol{s}) + \underbrace{\mathbf{X}^T \cdot \tilde{\mathbf{U}}}_{=\mathbf{0}} \cdot \mathbf{C} \\
&= \boldsymbol{\Sigma} - \underbrace{\boldsymbol{\Sigma} \cdot \boldsymbol{\Sigma}^{-1}}_{=\mathbf{I}} \cdot \text{diag}(\boldsymbol{s}) \\
&= \boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s}).
\end{aligned}$$

wts $\mathbf{M}_2 = \boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s})$:

$$\begin{aligned}
&\mathbf{M}_2 \\
&= \tilde{\mathbf{X}}^T \cdot \mathbf{X}
\end{aligned}$$

$$
\begin{aligned}
&= \left(\mathbf{X} - \mathbf{X}\cdot\boldsymbol{\Sigma}^{-1}\cdot\operatorname{diag}(\boldsymbol{s}) + \tilde{\mathbf{U}}\cdot\mathbf{C}\right)^T\cdot\mathbf{X}\\
&= \left(\mathbf{X}^T - (\operatorname{diag}(\boldsymbol{s}))^T\cdot\left(\boldsymbol{\Sigma}^{-1}\right)^T\cdot\mathbf{X}^T + \mathbf{C}^T\cdot\tilde{\mathbf{U}}^T\right)\cdot\mathbf{X}\\
&\qquad\qquad\text{Note that } \boldsymbol{\Sigma}^{-1} \text{ is symmetric. And diagonal}\\
&\qquad\qquad\text{matrices are obviously symmetric.}\\
&= \left(\mathbf{X}^T - \operatorname{diag}(\boldsymbol{s})\cdot\boldsymbol{\Sigma}^{-1}\cdot\mathbf{X}^T + \mathbf{C}^T\cdot\tilde{\mathbf{U}}^T\right)\cdot\mathbf{X}\\
&= \underbrace{\mathbf{X}^T\cdot\mathbf{X}}_{=\boldsymbol{\Sigma}} - \operatorname{diag}(\boldsymbol{s})\cdot\boldsymbol{\Sigma}^{-1}\cdot\underbrace{\mathbf{X}^T\cdot\mathbf{X}}_{=\boldsymbol{\Sigma}} + \mathbf{C}^T\cdot\underbrace{\tilde{\mathbf{U}}^T\cdot\mathbf{X}}_{=\mathbf{0}}\\
&= \boldsymbol{\Sigma} - \operatorname{diag}(\boldsymbol{s})\cdot\underbrace{\boldsymbol{\Sigma}^{-1}\cdot\boldsymbol{\Sigma}}_{=\mathbf{I}}\\
&= \boldsymbol{\Sigma} - \operatorname{diag}(\boldsymbol{s}).
\end{aligned}
$$

wts $\mathbf{M}_3 = \boldsymbol{\Sigma}$:

$$
\begin{aligned}
&\mathbf{M}_3\\
&= \tilde{\mathbf{X}}^T\cdot\tilde{\mathbf{X}}\\
&= \left(\mathbf{X} - \mathbf{X}\cdot\boldsymbol{\Sigma}^{-1}\cdot\operatorname{diag}(\boldsymbol{s}) + \tilde{\mathbf{U}}\cdot\mathbf{C}\right)^T\cdot\left(\mathbf{X} - \mathbf{X}\cdot\boldsymbol{\Sigma}^{-1}\cdot\operatorname{diag}(\boldsymbol{s}) + \tilde{\mathbf{U}}\cdot\mathbf{C}\right)\\
&= \left(\mathbf{X}^T - (\operatorname{diag}(\boldsymbol{s}))^T\cdot\left(\boldsymbol{\Sigma}^{-1}\right)^T\cdot\mathbf{X}^T + \mathbf{C}^T\cdot\tilde{\mathbf{U}}^T\right)\cdot\left(\mathbf{X} - \mathbf{X}\cdot\boldsymbol{\Sigma}^{-1}\cdot\operatorname{diag}(\boldsymbol{s}) + \tilde{\mathbf{U}}\cdot\mathbf{C}\right)\\
&\qquad\qquad\text{Note that } \boldsymbol{\Sigma}^{-1} \text{ is symmetric. And diagonal}\\
&\qquad\qquad\text{matrices are obviously symmetric.}\\
&= \left(\mathbf{X}^T - \operatorname{diag}(\boldsymbol{s})\cdot\boldsymbol{\Sigma}^{-1}\cdot\mathbf{X}^T + \mathbf{C}^T\cdot\tilde{\mathbf{U}}^T\right)\cdot\left(\mathbf{X} - \mathbf{X}\cdot\boldsymbol{\Sigma}^{-1}\cdot\operatorname{diag}(\boldsymbol{s}) + \tilde{\mathbf{U}}\cdot\mathbf{C}\right)\\
&= \underbrace{\mathbf{X}^T\cdot\mathbf{X}}_{=\boldsymbol{\Sigma}} - \underbrace{\mathbf{X}^T\cdot\mathbf{X}}_{=\boldsymbol{\Sigma}}\cdot\boldsymbol{\Sigma}^{-1}\cdot\operatorname{diag}(\boldsymbol{s}) + \underbrace{\mathbf{X}^T\cdot\tilde{\mathbf{U}}}_{=\mathbf{0}}\cdot\mathbf{C}\\
&\quad - \operatorname{diag}(\boldsymbol{s})\cdot\boldsymbol{\Sigma}^{-1}\cdot\underbrace{\mathbf{X}^T\cdot\mathbf{X}}_{=\boldsymbol{\Sigma}} + \operatorname{diag}(\boldsymbol{s})\cdot\boldsymbol{\Sigma}^{-1}\cdot\underbrace{\mathbf{X}^T\cdot\mathbf{X}}_{=\boldsymbol{\Sigma}}\cdot\boldsymbol{\Sigma}^{-1}\cdot\operatorname{diag}(\boldsymbol{s}) - \operatorname{diag}(\boldsymbol{s})\cdot\boldsymbol{\Sigma}^{-1}\cdot\underbrace{\mathbf{X}^T\cdot\tilde{\mathbf{U}}}_{=\mathbf{0}}\cdot\mathbf{C}\\
&\quad + \mathbf{C}^T\cdot\underbrace{\tilde{\mathbf{U}}^T\cdot\mathbf{X}}_{=\mathbf{0}} - \mathbf{C}^T\cdot\underbrace{\tilde{\mathbf{U}}^T\cdot\mathbf{X}}_{=\mathbf{0}}\cdot\boldsymbol{\Sigma}^{-1}\cdot\operatorname{diag}(\boldsymbol{s}) + \mathbf{C}^T\cdot\underbrace{\tilde{\mathbf{U}}^T\cdot\tilde{\mathbf{U}}}_{=\mathbf{I}}\cdot\mathbf{C}\\
&= \boldsymbol{\Sigma} - \underbrace{\boldsymbol{\Sigma}\cdot\boldsymbol{\Sigma}^{-1}}_{=\mathbf{I}}\cdot\operatorname{diag}(\boldsymbol{s}) - \operatorname{diag}(\boldsymbol{s})\cdot\underbrace{\boldsymbol{\Sigma}^{-1}\cdot\boldsymbol{\Sigma}}_{=\mathbf{I}} + \operatorname{diag}(\boldsymbol{s})\cdot\boldsymbol{\Sigma}^{-1}\cdot\underbrace{\boldsymbol{\Sigma}\cdot\boldsymbol{\Sigma}^{-1}}_{=\mathbf{I}}\cdot\operatorname{diag}(\boldsymbol{s}) + \mathbf{C}^T\cdot\mathbf{C}\\
&= \boldsymbol{\Sigma} - \operatorname{diag}(\boldsymbol{s}) - \operatorname{diag}(\boldsymbol{s}) + \operatorname{diag}(\boldsymbol{s})\cdot\boldsymbol{\Sigma}^{-1}\cdot\operatorname{diag}(\boldsymbol{s}) + \mathbf{C}^T\cdot\mathbf{C}\\
&\qquad\qquad\text{Using } \mathbf{C}^T\cdot\mathbf{C} = 2\cdot\operatorname{diag}(\boldsymbol{s}) - \operatorname{diag}(\boldsymbol{s})\cdot\boldsymbol{\Sigma}^{-1}\cdot\operatorname{diag}(\boldsymbol{s}).\\
&= \boldsymbol{\Sigma}.
\end{aligned}
$$

□

Comments:
i) If we would be in a setting where $n < 2\cdot p$, there would be no guarantee to find a $n\times p$ matrix $\tilde{\mathbf{U}}$ with the desired properties. We would have to assume that $\mathbf{X}$ contains at least $2\cdot p - n$ linearly dependent columns to guarantee that such a $n\times p$ matrix $\tilde{\mathbf{U}}$ exists (but such an assumption would violate the assumption that $\mathbf{X}^T\cdot\mathbf{X}$ is invertible). Another possibility would be to add zero-rows to $\mathbf{X}$ such that we are in the setting $n \geq 2\cdot p$ again (we use this trick in the approximate construction below).
ii) To calculate $\mathbf{C}$, note that

$$
\begin{aligned}
&\mathbf{C}^T\cdot\mathbf{C}\\
&\qquad\qquad\text{We know that } \mathbf{C}^T\cdot\mathbf{C} \text{ is symmetric.}\\
&\qquad\qquad\text{Therefore, we can apply Lemma A.12}
\end{aligned}
$$

$$= \mathbf{U} \cdot \mathbf{\Lambda} \cdot \mathbf{U}^T$$

We know that $\mathbf{C}^T \cdot \mathbf{C}$ is positive semidefinite. This means that all eigenvalues are nonnegative. Therefore, we can take their square roots.

$$= \mathbf{U} \cdot \mathbf{\Lambda}^{1/2} \cdot \mathbf{\Lambda}^{1/2} \cdot \mathbf{U}^T$$

Note that $\mathbf{\Lambda}^{1/2}$ is a diagonal matrix.

$$= \mathbf{U} \cdot \mathbf{\Lambda}^{1/2} \cdot \left(\mathbf{\Lambda}^{1/2}\right)^T \cdot \mathbf{U}^T$$
$$= \left(\left(\mathbf{\Lambda}^{1/2}\right)^T \cdot \mathbf{U}^T\right)^T \cdot \left(\mathbf{\Lambda}^{1/2}\right)^T \cdot \mathbf{U}^T.$$

and we see that $\mathbf{C} = \left(\mathbf{\Lambda}^{1/2}\right)^T \cdot \mathbf{U}^T$.

Example 2.17. One way of calculating a FX-Knockoff matrix $\tilde{\mathbf{X}}$ for a corresponding design matrix $\mathbf{X}$ is:
1) Normalize the columns of $\mathbf{X}$.
2) Choose a procedure (e.g. equi or SDP) to build $\boldsymbol{s}$.
3) Calculate $\tilde{\mathbf{U}}$ with QR-decomposition (as described in Prep1 of Lemma 2.16).
4) Calculate $\mathbf{\Sigma}^{-1}$ (exists and can be calculated because $\mathbf{\Sigma} = \mathbf{X}^T \cdot \mathbf{X}$ is assumed to be invertible).
5) Calculate $\mathbf{C}$ (as explained above).
6) Calculate $\tilde{\mathbf{X}} = \mathbf{X} - \mathbf{X} \cdot \mathbf{\Sigma}^{-1} \cdot \mathrm{diag}(\boldsymbol{s}) + \tilde{\mathbf{U}} \cdot \mathbf{C}$.

Comments:
Normalizing columns can be done via $\mathbf{X}_j^{\text{normalized}} = \dfrac{1}{\sqrt{\sum_{i=1}^n X_{ij}^2}} \cdot \mathbf{X}_j$ for all j (where X_{ij} is the entry in row i and column j of the matrix $\mathbf{X}$), because

$$\begin{aligned}
&||\mathbf{X}_j^{\text{normalized}}||_2^2 \\
&= \left|\left| \frac{1}{\sqrt{\sum_{i=1}^n X_{ij}^2}} \cdot \mathbf{X}_j \right|\right|_2^2 \\
&= \left(\sqrt{ \left(\frac{1}{\sqrt{\sum_{i=1}^n X_{ij}^2}} \cdot X_{1j} \right)^2 + \cdots + \left(\frac{1}{\sqrt{\sum_{i=1}^n X_{ij}^2}} \cdot X_{nj} \right)^2 } \right)^2 \\
&= \frac{1}{\sum_{i=1}^n X_{ij}^2} \cdot X_{1j}^2 + \cdots + \frac{1}{\sum_{i=1}^n X_{ij}^2} \cdot X_{nj}^2 \\
&= \frac{1}{\sum_{i=1}^n X_{ij}^2} \cdot \left(X_{1j}^2 + \cdots + X_{nj}^2\right) \\
&= \frac{1}{\sum_{i=1}^n X_{ij}^2} \cdot \sum_{i=1}^n X_{ij}^2 \\
&= 1.
\end{aligned}$$

2.4.2 Approximate construction of FX Knockoffs (if $n < 2 \cdot p$)

For $n < 2 \cdot p$, $\tilde{\mathbf{U}}$ cannot be constructed as above (see proof of Lemma 2.16), but we can construct the FX-Knockoff matrix approximately for $p < n < 2 \cdot p$. In (1), σ^2 is not assumed to be known. But for such models, we know that the residual sum of squares is distributed as $\text{RSS} := ||\mathbf{Y} - \mathbf{X} \cdot \hat{\boldsymbol{\beta}}^{\text{LS}}||_2^2 \sim \sigma^2 \cdot \chi^2_{n-p}$, where $\hat{\boldsymbol{\beta}}^{\text{LS}}$ is the least squares estimator. And $\mathbb{E}(\text{RSS}) = \mathbb{E}\left(\sigma^2 \cdot \chi^2_{n-p}\right) = \sigma^2 \cdot \mathbb{E}\left(\chi^2_{n-p}\right) = \sigma^2 \cdot (n-p)$. Therefore,

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n-p}$$

is an unbiased estimator of σ^2 and we can draw a $(2 \cdot p - n) \times 1$ vector $\mathbf{Y}'$ with i.i.d. $\mathcal{N}\left(0, \hat{\sigma}^2\right)$ entries. If $n - p$ is large, $\hat{\sigma}^2$ will be close to σ^2 because $\mathbb{V}ar\left(\hat{\sigma}^2\right) = \mathbb{V}ar\left(\frac{\text{RSS}}{n-p}\right) = \frac{1}{(n-p)^2} \cdot \mathbb{V}ar\left(\text{RSS}\right)$ and we have approximately

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{Y}' \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{X} \\ \mathbf{0}_{(2\cdot p-n)\times p} \end{pmatrix} \cdot \boldsymbol{\beta},\ \sigma^2 \cdot \mathbf{I}_{2\cdot p}\right),$$

where the $2 \cdot p \times p$ matrix $\begin{pmatrix} \mathbf{X} \\ \mathbf{0}_{(2\cdot p-n)\times p} \end{pmatrix}$ is the rowwise concatenation of $\mathbf{X}$ and the zero-matrix $\mathbf{0}_{(2\cdot p-n)\times p}$. This means that we again have a linear model with response $\mathbf{Y}^* := \begin{pmatrix} \mathbf{Y} \\ \mathbf{Y}' \end{pmatrix}$, non-random design matrix $\mathbf{X}^* := \begin{pmatrix} \mathbf{X} \\ \mathbf{0}_{(2\cdot p-n)\times p} \end{pmatrix}$, coefficient vector $\boldsymbol{\beta}$ and error vector $\boldsymbol{z}^* \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \cdot \mathbf{I}_{2\cdot p}\right)$. For this linear model with $n^* := 2 \cdot p$ observations and p variables, we can apply the same FX-Knockoff methods we used for the $n \geq 2 \cdot p$ setting.

Alternative approaches for $p < n < 2 \cdot p$ are given in Barber and Candes (2015) and Barber and Candes (2019) show strategies for applying FX-Knockoff procedures in a setting where $p > n$.

2.4.3 Construction of FX-Knockoff feature statistics

We now show some examples of FX-Knockoff feature statistics W_j. Remember that the flip-sign and sufficiency properties have to be satisfied for valid FX-Knockoff feature statistics. The flip-sign property is easy to check and it is satisfied in a lot of situations, whereas the sufficiency property severely limits the choice of valid feature statistics.

Example 2.18. Statistics $W_j = |\mathbf{X}_j^T \cdot \mathbf{Y}| - |\tilde{\mathbf{X}}_j^T \cdot \mathbf{Y}|$ are valid FX-Knockoff feature statistics.

Proof:

Part1 wts w_j satisfies the flip-sign property:

If variables $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ are swapped,

$$w_j\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right]_{\text{swap}(\{j\})},\mathbf{Y}\right)=|\tilde{\mathbf{X}}_j^T\cdot\mathbf{Y}|-|\mathbf{X}_j^T\cdot\mathbf{Y}|=-\left(|\mathbf{X}_j^T\cdot\mathbf{Y}|-|\tilde{\mathbf{X}}_j^T\cdot\mathbf{Y}|\right)=-w_j\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right],\mathbf{Y}\right).$$

If other variables are swapped, w_j is unaffected since it only uses $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$.

Part2 wts w_j satisfies the sufficiency property:

Note that

$$\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\mathbf{Y}=\left[\mathbf{X}_1,\ldots,\mathbf{X}_p,\tilde{\mathbf{X}}_1,\ldots,\tilde{\mathbf{X}}_p\right]^T\cdot\mathbf{Y}=\begin{bmatrix}\mathbf{X}_1^T\\ \vdots\\ \mathbf{X}_p^T\\ \tilde{\mathbf{X}}_1^T\\ \vdots\\ \tilde{\mathbf{X}}_p^T\end{bmatrix}\cdot\mathbf{Y}=\begin{bmatrix}\mathbf{X}_1^T\cdot\mathbf{Y}\\ \vdots\\ \mathbf{X}_p^T\cdot\mathbf{Y}\\ \tilde{\mathbf{X}}_1^T\cdot\mathbf{Y}\\ \vdots\\ \tilde{\mathbf{X}}_p^T\cdot\mathbf{Y}\end{bmatrix}$$

and

$$\mathbf{W}=(W_1,\ldots,W_p)=\left(w_1\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right],\mathbf{Y}\right),\ldots,w_p\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right],\mathbf{Y}\right)\right)=\left(|\mathbf{X}_1^T\cdot\mathbf{Y}|-|\tilde{\mathbf{X}}_1^T\cdot\mathbf{Y}|,\ldots,|\mathbf{X}_p^T\cdot\mathbf{Y}|-|\tilde{\mathbf{X}}_p^T\cdot\mathbf{Y}|\right).$$

Therefore, w_j depends upon $\mathbf{X}$, $\tilde{\mathbf{X}}$ and $\mathbf{Y}$ only through $\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\mathbf{Y}$. □

Example 2.19. For the linear model $\mathbf{Y}=\left[\mathbf{X},\tilde{\mathbf{X}}\right]\cdot\boldsymbol{\beta}+\boldsymbol{z}$ we use least squares to estimate $\boldsymbol{\beta}$ and get the solution $\hat{\boldsymbol{\beta}}=\left(\hat{\beta}_1,\ldots,\hat{\beta}_p,\hat{\beta}_{1+p},\ldots,\hat{\beta}_{p+p}\right)$. Then the following statistics are valid FX-Knockoff feature statistics:
i) $W_j=|\hat{\beta}_j|-|\hat{\beta}_{j+p}|$,
ii) $W_j=|\hat{\beta}_j|^2-|\hat{\beta}_{j+p}|^2$.

Proof:

The sufficiency property for i) and ii) is trivially satisfied, because the LS-solution in our case is

$$\hat{\boldsymbol{\beta}}=\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]\right)^{-1}\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\mathbf{Y}.$$

And w_j uses two entries of $\hat{\boldsymbol{\beta}}$ to build statistics in both cases. Hence w_j depends on $\mathbf{X}$, $\tilde{\mathbf{X}}$ and $\mathbf{Y}$ only through $\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]$ and $\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\mathbf{Y}$.

ad i) wts w_j satisfies the flip-sign property:

If variables $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ are swapped,

$$w_j\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right]_{\text{swap}(\{j\})},\mathbf{Y}\right)=|\hat{\beta}_{j+p}|-|\hat{\beta}_j|=-\left(|\hat{\beta}_j|-|\hat{\beta}_{j+p}|\right)=-w_j\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right],\mathbf{Y}\right).$$

If other variables are swapped, w_j is unaffected since LS-solutions $\hat{\beta}_j$ and $\hat{\beta}_{j+p}$ remain the same as long as entries $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ are not swapped.

ad ii) wts w_j satisfies the flip-sign property:

If variables $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ are swapped,

$$w_j\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right]_{\text{swap}(\{j\})},\mathbf{Y}\right)=|\hat{\beta}_{j+p}|^2-|\hat{\beta}_j|^2=-\left(|\hat{\beta}_j|^2-|\hat{\beta}_{j+p}|^2\right)=-w_j\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right],\mathbf{Y}\right).$$

If other variables are swapped, w_j is unaffected since LS-solutions $\hat{\beta}_j$ and $\hat{\beta}_{j+p}$ remain the same as long as entries $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ are not swapped.

□

Example 2.20. For the linear model $\mathbf{Y}=\left[\mathbf{X},\tilde{\mathbf{X}}\right]\cdot\boldsymbol{\beta}+\boldsymbol{z}$ we use Lasso to estimate $\boldsymbol{\beta}$ and get the solution $\hat{\boldsymbol{\beta}}(\lambda)=\left(\hat{\beta}_1(\lambda),\ldots,\hat{\beta}_p(\lambda),\hat{\beta}_{1+p}(\lambda),\ldots,\hat{\beta}_{p+p}(\lambda)\right)$. We define $Z_j=\sup\left\{\lambda:\ \hat{\beta}_j(\lambda)\neq 0\right\}$ for $j\in\{1,\ldots,2\cdot p\}$ and end up with a $2\cdot p$-dimensional vector $(Z_1,\ldots,Z_p,Z_{1+p},\ldots,Z_{p+p})$. Then the following statistics are valid FX-Knockoff feature statistics:
i) Lasso Signed Max (LSM) statistics $W_j=(Z_j\vee Z_{j+p})\cdot\text{sign}(Z_j-Z_{j+p})$ for $j\in\{1,\ldots,p\}$,
ii) $W_j=Z_j-Z_{j+p}$ for $j\in\{1,\ldots,p\}$.
iii) If $Z_j=|\hat{\beta}_j(\lambda)|$ for all $j\in\{1,\ldots,2\cdot p\}$ is used instead, then the Lasso coefficient-difference (LCD) statistics $W_j=Z_j-Z_{j+p}$ for $j\in\{1,\ldots,p\}$ are valid FX-Knockoff feature statistics for a pre-specified value λ.

Proof:

Prep1 wts $||\mathbf{Y}-\left[\mathbf{X},\tilde{\mathbf{X}}\right]\cdot\mathbf{b}||_2^2=\mathbf{Y}^T\cdot\mathbf{Y}-2\cdot\mathbf{b}^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\mathbf{Y}+\mathbf{b}^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]\cdot\mathbf{b}$:

$$\begin{aligned}
&||\mathbf{Y}-\left[\mathbf{X},\tilde{\mathbf{X}}\right]\cdot\mathbf{b}||_2^2\\
&=\left(\sqrt{\left(\mathbf{Y}-\left[\mathbf{X},\tilde{\mathbf{X}}\right]\cdot\mathbf{b}\right)^T\cdot\left(\mathbf{Y}-\left[\mathbf{X},\tilde{\mathbf{X}}\right]\cdot\mathbf{b}\right)}\right)^2\\
&=\left(\mathbf{Y}-\left[\mathbf{X},\tilde{\mathbf{X}}\right]\cdot\mathbf{b}\right)^T\cdot\left(\mathbf{Y}-\left[\mathbf{X},\tilde{\mathbf{X}}\right]\cdot\mathbf{b}\right)\\
&=\left(\mathbf{Y}^T-\mathbf{b}^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\right)\cdot\left(\mathbf{Y}-\left[\mathbf{X},\tilde{\mathbf{X}}\right]\cdot\mathbf{b}\right)\\
&=\mathbf{Y}^T\cdot\mathbf{Y}-\mathbf{Y}^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]\cdot\mathbf{b}-\mathbf{b}^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\mathbf{Y}+\mathbf{b}^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]\cdot\mathbf{b}\\
&\qquad\qquad\text{Note that }\mathbf{Y}^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]\cdot\mathbf{b}\text{ is }1\times 1\text{. So we can}\\
&\qquad\qquad\text{write }\mathbf{Y}^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]\cdot\mathbf{b}=\left(\mathbf{Y}^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]\cdot\mathbf{b}\right)^T=\mathbf{b}^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\mathbf{Y}.\\
&=\mathbf{Y}^T\cdot\mathbf{Y}-2\cdot\mathbf{b}^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\mathbf{Y}+\mathbf{b}^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]\cdot\mathbf{b}.
\end{aligned}$$

To check the sufficiency property for i), ii) and iii), note that the Lasso-solution in our case is

$$
\begin{aligned}
&\hat{\boldsymbol{\beta}}(\lambda) \\
&= \underset{\mathbf{b}\in\mathbb{R}^{2\cdot p}}{\arg\min} \left[\frac{1}{2}\cdot||\mathbf{Y}-\left[\mathbf{X},\tilde{\mathbf{X}}\right]\cdot\mathbf{b}||_2^2+\lambda\cdot||\mathbf{b}||_1\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{By Prep1} \\
&= \underset{\mathbf{b}\in\mathbb{R}^{2\cdot p}}{\arg\min} \left[\frac{1}{2}\cdot\left(\mathbf{Y}^T\cdot\mathbf{Y}-2\cdot\mathbf{b}^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\mathbf{Y}+\mathbf{b}^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]\cdot\mathbf{b}\right)+\lambda\cdot||\mathbf{b}||_1\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{We are interested in the } \mathbf{b} \text{ which minimizes the whole expression.} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{Therefore, all terms which do not contain } \mathbf{b} \text{ can be dropped.} \\
&= \underset{\mathbf{b}\in\mathbb{R}^{2\cdot p}}{\arg\min} \left[\frac{1}{2}\cdot\left(-2\cdot\mathbf{b}^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\mathbf{Y}+\mathbf{b}^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]\cdot\mathbf{b}\right)+\lambda\cdot||\mathbf{b}||_1\right] \\
&= \underset{\mathbf{b}\in\mathbb{R}^{2\cdot p}}{\arg\min} \left[\frac{1}{2}\cdot\mathbf{b}^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]\cdot\mathbf{b}-\mathbf{b}^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\mathbf{Y}+\lambda\cdot||\mathbf{b}||_1\right].
\end{aligned}
$$

We see, that the Lasso-solution depends on $\mathbf{X}$, $\tilde{\mathbf{X}}$ and $\mathbf{Y}$ only through $\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]$ and $\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\mathbf{Y}$. In iii), w_j is created by using entries j and $j+p$ of $\hat{\boldsymbol{\beta}}(\lambda)$ for a pre-specified λ. Hence, w_j depends on $\mathbf{X}$, $\tilde{\mathbf{X}}$ and $\mathbf{Y}$ only through $\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]$ and $\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\mathbf{Y}$.

In i) and ii), w_j uses Lasso solutions $\hat{\boldsymbol{\beta}}(\lambda)$ for pre-specified λ and in a second step, takes the largest λ such that $\hat{\beta}_j(\lambda)\neq 0$. Hence, w_j depends on $\mathbf{X}$, $\tilde{\mathbf{X}}$ and $\mathbf{Y}$ only through $\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\left[\mathbf{X},\tilde{\mathbf{X}}\right]$ and $\left[\mathbf{X},\tilde{\mathbf{X}}\right]^T\cdot\mathbf{Y}$.

ad i) wts w_j satisfies the flip-sign property:

If variables $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ are swapped,

$$
\begin{aligned}
&w_j\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right]_{\text{swap}(\{j\})},\mathbf{Y}\right) \\
&= (Z_{j+p}\vee Z_j)\cdot\text{sign}(Z_{j+p}-Z_j) \\
&= (Z_j\vee Z_{j+p})\cdot\text{sign}\left((-1)\cdot(Z_j-Z_{j+p})\right) \\
&= -(Z_j\vee Z_{j+p})\cdot\text{sign}\left(Z_j-Z_{j+p}\right) \\
&= -w_j\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right],\mathbf{Y}\right).
\end{aligned}
$$

If other variables are swapped, w_j is unaffected since Lasso-solutions $\hat{\beta}_j(\lambda)$ and $\hat{\beta}_{j+p}(\lambda)$ remain the same as long as entries $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ are not swapped.

ad ii) and iii) wts w_j satisfies the flip-sign property:

If variables $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ are swapped,

$$
w_j\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right]_{\text{swap}(\{j\})},\mathbf{Y}\right) = Z_{j+p}-Z_j = -(Z_j-Z_{j+p}) = -w_j\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right],\mathbf{Y}\right).
$$

If other variables are swapped, w_j is unaffected since Lasso-solutions $\hat{\beta}_j(\lambda)$ and $\hat{\beta}_{j+p}(\lambda)$ remain the same as long as entries $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ are not swapped.

□

Comments:
a) Note that in i) and ii), Z_j is the largest penalty parameter λ at which $\mathbf{X}_j$ enters the model in the Lasso regression. Z_{j+p} is the largest penalty parameter λ at which $\tilde{\mathbf{X}}_j$ enters the model in the Lasso regression. A high value of λ indicates, that the variable seems to be important for the model. The statistics W_j in i) can be written as

$$W_j = \begin{cases} Z_j, & Z_j > Z_{j+p} \\ 0, & Z_j = Z_{j+p} \\ -Z_{j+p}, & Z_j < Z_{j+p} \end{cases}$$

and note that $\lambda \geq 0$ and therefore $Z_j, Z_{j+p} \geq 0$. This means for i) and ii) that large positive values of W_j indicate that variable $\mathbf{X}_j$ enters the Lasso model earlier than its fake version $\tilde{\mathbf{X}}_j$. So, large positive values of W_j give evidence that j is non-null. This is exactly the setup for which all our procedures are designed (see comments after Definition 2.4).
b) We can extend i) and ii) to all penalized likelihood estimation procedures of the form
$\hat{\boldsymbol{\beta}}(\lambda) = \underset{\mathbf{b} \in \mathbb{R}^{2 \cdot p}}{\arg\min} \left[\frac{1}{2} \cdot ||\mathbf{Y} - \left[\mathbf{X}, \tilde{\mathbf{X}}\right] \cdot \mathbf{b}||_2^2 + \lambda \cdot P(\mathbf{b}) \right]$, where $P(\cdot)$ is a penalty function. E.g. $P(\mathbf{b}) = ||\mathbf{b}||_2^2$ for Ridge regression.
c) More possibilities of creating FX Knockoff feature statistics (e.g. forward selection procedures) are given in Barber and Candes (2015).

3 Model-X Knockoff methods

Model-X Knockoff procedures were introduced first by Candes et al. (2018) and the whole chapter is based on that paper.

3.1 Introduction

The ideas of the previous chapter can be used to create Knockoff procedures for a different setting (with totally different assumptions). Again, abstractly speaking, the Knockoff copies can be seen as control group of the original variables. Feature statistics are used to tease apart important and unimportant variables in the final selection process. And variables get selected, if the corresponding feature statistic (importance measure) lies above some threshold. Because of this similarity to the fixed-X setting, the procedures itself are not explained in detail in this model-X part (we did that already in the fixed-X part). The symbols used in this chapter are the same as in the fixed-X part such that both settings are easy to compare.

In contrast to the fixed-X setting, assumptions for the response $\mathbf{Y}$ and the variables in $\mathbf{X}$ are totally different in the model-X setting.
First, instead of using a linear model and a fixed (non-random) $\mathbf{X}$, we consider a general conditional model with random $\mathbf{X}$.
Second, we assume that we have i.i.d. pairs $(X_{i1}, \dots, X_{ip}, Y_i) \in \mathbb{R}^p \times \mathbb{R},\ i = 1, \dots, n$, of covariates and responses, where

$$\mathbf{X} = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}.$$

And third (this is a major change compared to common setups in conditional models), we shift what authors call the "burden of knowledge". We consider a general conditional model, but instead of assuming that $\mathbf{Y}|\mathbf{X}$ follows a certain distribution, we assume exact knowledge of the distribution of $\mathbf{X}$ and nothing about the distribution of $\mathbf{Y}$. In other words, we model $\mathbf{X}$ instead of $\mathbf{Y}|\mathbf{X}$ (this is why we call the procedures of this chapter model-X Knockoff methods). Of course, this shift of knowledge changes the field of application of the model-X Knockoff methods. The paper lists several situations, when this third assumption is useful. Genetic case-control studies would be an example, where researchers know a lot about covariates but less about the response. The validity of model-X Knockoff procedures in such case-control studies is discussed in Barber and Candes (2018).

To summarize, the assumptions in the model-X setup are:

- $\mathbf{X}$ is random.
- We have n i.i.d. pairs $(X_{i1}, \dots, X_{ip}, Y_i) \in \mathbb{R}^p \times \mathbb{R}$ of covariates and responses.
- The distribution of $\mathbf{X}$ is known exactly.

We do not assume knowledge of the conditional distribution of $\mathbf{Y}|\mathbf{X}$ and we do not assume $n > p$. The number of covariates p can be much higher than the number of observations n.
The second assumption means that we assume that the observations (rows of $(\mathbf{X}, \mathbf{Y})$) are independent and that $(\mathbf{X}, \mathbf{Y})$ follows some arbitrary joint distribution. The i.i.d. property of $(\mathbf{X}, \mathbf{Y})$ will allow us to formulate and prove properties for an arbitrary row of $(\mathbf{X}, \mathbf{Y})$. Indeed, due to row independence, the cdf of $(\mathbf{X}, \mathbf{Y})$ can be written as product of cdfs of the rows, and all these cdfs are the same since all rows are identically distributed.
In the fixed-X part, we used the notation $\mathbf{X}_j$ for the j-th column of $\mathbf{X}$. Based on that notation, it is not easy to find a good and intuitive notation for the j-th row of $\mathbf{X}$. Of course, we could have used $\mathbf{X}_{\cdot j}$ for the j-th

column of $\mathbf{X}$ and then we would be able to use $\mathbf{X}_{j\cdot}$ for the j-th row of $\mathbf{X}$, but this notation is very annoying in lengthy and complicated proofs. Therefore, we simply use (X, Y) for an arbitrary row of $(\mathbf{X}, \mathbf{Y})$. And instead of writing $(X_{i1}, \ldots, X_{ip}, Y_i)$ for an arbitrary row i, we simply just write $(X_1, \ldots, X_p, Y)$.

We would like to know which of the many variables does Y depend on. In other words, we would like to find the "smallest" subset $\mathcal{S} \subseteq \{1, \ldots, p\}$, such that conditionally on $\{X_j\}_{j \in \mathcal{S}}$, the response Y is independent of all other variables (meaning that the other variables do not provide additional information about Y). Such a mininal set $\mathcal{S}$ is usually called *Markov boundary* in the literature on graphical models (see Definition B.1). For a lot of joint distributions of (X, Y), there exists a unique Markov boundary but there are pathological cases where it does not. The next example shows a case, where no unique Markov boundary exists.

Example 3.1. Let X_1 and X_2 be independent random variables with Gaussian distribution. Let $X_3 = X_1 - X_2$ and assume that the distribution of Y depends upon the vector $X = (X_1, X_2, X_3)$ only through $X_1 + X_2$, e.g., $Y|X \sim \mathcal{N}(X_1 + X_2, 1)$. Then the Markov boundary is ill defined (not unique).

Proof:

It is true that Y depends on X through (X_1, X_2), i.e., $(Y \perp\!\!\!\perp X_3)|\{X_1, X_2\}$.
Thus, the Markov blanket would be $S = \{X_1, X_2\}$.

But it is also true that Y depends on X through (X_1, X_3), i.e., $(Y \perp\!\!\!\perp X_2)|\{X_1, X_3\}$, because using $X_3 = X_1 - X_2$, we get $X_1 + X_2 = X_1 + X_1 - X_3 = 2 \cdot X_1 + X_3$ and this what we know by conditioning on $\{X_1, X_3\}$.
Thus, the Markov blanket would be $S = \{X_1, X_3\}$.

And it is also true that Y depends on X through (X_2, X_3), i.e., $(Y \perp\!\!\!\perp X_1)|\{X_2, X_3\}$, because using $X_3 = X_1 - X_2$, we get $X_1 + X_2 = X_3 + X_2 + X_2 = 2 \cdot X_2 + X_3$ and this what we know by conditioning on $\{X_2, X_3\}$.
Thus, the Markov blanket would be $S = \{X_2, X_3\}$.

All three sets S are equally good and cannot be "decreased" any further. Therefore, all three sets S are Markov boundaries.

□

Comments:
i) The definition of conditional independence is given in Definitions B.2 and B.3.
ii) Markov blanket is explained in Definition B.1 (note that some authors define Markov blanket the same way as we defined Markov boundary).
iii) In order to define a unique set of relevant variables, we use the notion of conditional pairwise independence:

Definition 3.2. (Definition 2.1 in Candes et al. (2018)). An index j or variable X_j is said to be *null* if and only if $(Y \perp\!\!\!\perp X_j)|X_{-j}$, where $X_{-j} = \{X_1, \ldots, X_p\} \setminus \{X_j\}$. The *subset of null indices* is denoted by $\mathcal{H}_0 \subseteq \{1, \ldots, p\}$ and we call an index j or a variable X_j *non-null* or *relevant* if $j \notin \mathcal{H}_0$.

In graphical model theory, independence of two variables, conditionally on all other variables, is known as *Pairwise Markov property*. That is what we have used in Definition 3.2. Our hope is, that the set of non-nulls with indices $\mathcal{H}_0^c := \{1, \ldots, p\} \setminus \mathcal{H}_0$ is exactly the desired unique Markov boundary, i.e., the unique smallest set $\{X_{-j}\}_{j \in \mathcal{H}_0^c}$ such that $(Y \perp\!\!\!\perp \{X_j\}_{j \in \mathcal{H}_0}) | \{X_{-j}\}_{j \in \mathcal{H}_0^c}$. The latter conditional independence is called *Local Markov property*. The Local Markov property is stronger than the Pairwise Markov property (see Part 4.3.2.2 in Koller and Friedman (2009)), but both properties are equivalent under quite general conditions (see pages 7 and 8 in Edwards (2000)). This means that if these general conditions hold, our definition of nulls and non-nulls leads to an unique Markov boundary (the non-null variables constitutes the Markov boundary).

Example 3.3. Consider the same setting as in Example 3.1. Then
i) all three variables X_1, X_2, X_3 would be classified as nulls.
ii) adding a bit of noise, e.g., $X_3 = X_1 - X_2 + Z$ where Z is Gaussian noise (independent of X_1 and X_2) however small, X_1 and X_2 are both non-null while X_3 is null.

Proof:

ad i):

Because of the perfect functional relationship $X_3 = X_1 - X_2$, we have seen that $(Y \perp\!\!\!\perp X_3) | \{X_1, X_2\}$, $(Y \perp\!\!\!\perp X_2) | \{X_1, X_3\}$ and $(Y \perp\!\!\!\perp X_1) | \{X_2, X_3\}$ holds in the previous example. Therefore, by Definition 3.2, all three variables X_1, X_2, X_3 are classified as nulls.

ad ii):

Instead of $X_3 = X_1 - X_2$, we have $X_3 = X_1 - X_2 + Z$ where Z is Gaussian noise (independent of X_1 and X_2).
So, X_1 is not determined by knowing X_2 and X_3.
And X_2 is not determined by knowing X_1 and X_3.

Therefore,
$(Y \perp\!\!\!\perp X_3) | \{X_1, X_2\}$ still holds, because Y depends upon the vector (X_1, X_2, X_3) only through $X_1 + X_2$, which is known because we condition on $\{X_1, X_2\}$.

$(Y \perp\!\!\!\perp X_2) | \{X_1, X_3\}$ is not true anymore. Note that $X_2 = X_1 + Z - X_3$. Therefore, $X_1 + X_2 = X_1 + X_1 + Z - X_3 = 2 \cdot X_1 + Z - X_3$. We know X_1 and X_3, but we have no information about Z and therefore no information about $X_1 + X_2$.

$(Y \perp\!\!\!\perp X_1) | \{X_2, X_3\}$ is not true anymore for the same reason. With $X_1 = X_2 + X_3 - Z$ we get $X_1 + X_2 = X_2 + X_3 - Z + X_2 = 2 \cdot X_2 + X_3 - Z$ and again, we have no information about Z and therefore no information about $X_1 + X_2$.

□

Usually, even if we know the distribution of X, exact recovery of the Markov boundary (perfectly teasing apart nulls and non-nulls and selecting all these non-nulls) is not possible. In other words, no perfect variable selection procedure exists for most of the problems. Variable selection procedures can produce two types of errors. Falsely selecting nulls (leads to type I error) and not selecting all non-nulls (leads to lower power). As in the fixed-X setting, we focus on controlling the FDR. We will present two variable selection procedures, which attempt to discover as many non-nulls as possible while keeping the FDR (or a modified version of it) under control.

The next Proposition shows, that for Generalized Linear Models (GLMs) with no redundant covariates, $\beta_j = 0$ is the same as $Y \perp\!\!\!\perp X_j | X_{-j}$. This means that testing the hypothesis that X_j is a null variable is the same as testing the hypothesis that $\beta_j = 0$ (which is what we did in the fixed-X setting where nulls are simply defined that way, see Definition 2.1).

Proposition 3.4. *(Proposition 2.2 in Candes et al. (2018)). Let $X_1, \ldots, X_p$ be a family of random variables such that one cannot perfectly predict any one of them from knowledge of the others. Consider a GLM (where the response Y follows a distribution in an exponential family). Then $Y \perp\!\!\!\perp X_j | X_{-j}$ if and only if $\beta_j = 0$. Hence, $\mathcal{H}_0$ from Definition 3.2 is exactly the set $\{j : \beta_j = 0\}$.*

Proof:

We prove this in the case of the logistic regression model. Candes et al. (2018) claim that the general case is similar.

For logistic regression, we have the setting:
We have a random vector $X = (X_1, \ldots, X_p)$ and random variable Y with $Y|X \sim \text{Bernoulli}(\theta)$.
Let $\beta_1, \ldots, \beta_p$ be unknown parameters.
The logit function is used as link function g, hence $g(x) = \ln\left(\frac{x}{1-x}\right)$.
The linear predictor is defined as $\eta = X \cdot \beta = \beta_1 \cdot X_1 + \cdots + \beta_p \cdot X_p$.
We set $g(\theta) = \eta$.
Exponentiating gives $\frac{\theta}{1-\theta} = e^\eta$ and solving for θ results in $\theta = \dfrac{e^\eta}{1+e^\eta}$.
And we know that $\mathbb{E}(Y|X) = g^{-1}(\eta)$, because

$$
\begin{aligned}
&\mathbb{E}(Y|X) \\
&\qquad \text{By definition of conditional expectation of a discrete random variable.} \\
&= \sum_{y \in \mathcal{Y}} y \cdot \mathbb{P}(Y = y \mid X) \\
&\qquad \text{Note that } \mathcal{Y} = \{0, 1\}. \\
&= \sum_{i=0}^{1} i \cdot \mathbb{P}(Y = i \mid X) \\
&= 0 \cdot \mathbb{P}(Y = 0 \mid X) + 1 \cdot \mathbb{P}(Y = 1 \mid X) \\
&= \mathbb{P}(Y = 1 \mid X) \\
&\qquad \text{The probability mass function of the Bernoulli distribution for } k \in \{0, 1\} \text{ is given by } \mathbb{P}(Y = k \mid X) = \theta^k \cdot (1-\theta)^{1-k}.
\end{aligned}
$$

$$
\begin{aligned}
&= \theta^1 \cdot (1-\theta)^{1-1} \\
&= \theta \\
&\qquad\qquad \text{We know that } g(\theta) = \eta. \\
&\qquad\qquad \text{Therefore, } \theta = g^{-1}(\eta). \\
&= g^{-1}(\eta).
\end{aligned}
$$

Prep1 wts $\mathbb{P}(A, B|C) = \mathbb{P}(A|B, C) \cdot \mathbb{P}(B|C)$ for events A, B, C:

$$
\begin{aligned}
&\mathbb{P}(A|B, C) \\
&\qquad\qquad \text{By definition of conditional probability} \\
&= \frac{\mathbb{P}(A, B, C)}{\mathbb{P}(B, C)} \\
&= \frac{\frac{\mathbb{P}(A,B,C)}{\mathbb{P}(C)}}{\frac{\mathbb{P}(B,C)}{\mathbb{P}(C)}} \\
&\qquad\qquad \text{By definition of conditional probability} \\
&= \frac{\mathbb{P}(A, B|C)}{\mathbb{P}(B|C)}.
\end{aligned}
$$

Of course, this only holds if $\mathbb{P}(C) > 0,\ \mathbb{P}(B, C) > 0,\ \mathbb{P}(B|C) > 0$.

Prep2 wts $\mathbb{P}(Y = y|X_j \leq x_j, X_{-j} \leq x_{-j}) = \dfrac{\exp(y \cdot X \cdot \beta)}{1 + \exp(X \cdot \beta)}$:

$$
\begin{aligned}
&\mathbb{P}(Y = y|X_j \leq x_j, X_{-j} \leq x_{-j}) \\
&= \mathbb{P}(Y = y|X) \\
&\qquad\qquad \text{We know that } Y|X \sim \text{Bernoulli}(\theta). \\
&= \theta^y \cdot (1-\theta)^{1-y} \\
&= \exp\left[\ln\left(\theta^y \cdot (1-\theta)^{1-y}\right)\right] \\
&= \exp\left[y \cdot \ln(\theta) + (1-y) \cdot \ln(1-\theta)\right] \\
&= \exp\left[y \cdot \ln(\theta) + \ln(1-\theta) - y \cdot \ln(1-\theta)\right] \\
&= \exp\left[y \cdot (\ln(\theta) - \ln(1-\theta)) + \ln(1-\theta)\right] \\
&= \exp\left[y \cdot \ln\left(\frac{\theta}{1-\theta}\right) + \ln(1-\theta)\right] \\
&\qquad\qquad \text{We know that } g(\theta) = \ln\left(\frac{\theta}{1-\theta}\right) = \eta. \\
&= \exp\left[y \cdot \eta + \ln(1-\theta)\right] \\
&= e^{y \cdot \eta} \cdot e^{\ln(1-\theta)} \\
&= e^{y \cdot \eta} \cdot (1-\theta) \\
&\qquad\qquad \text{We know that } \theta = \frac{e^\eta}{1+e^\eta}. \\
&= e^{y \cdot \eta} \cdot \left(1 - \frac{e^\eta}{1+e^\eta}\right) \\
&= e^{y \cdot \eta} \cdot \left(\frac{1+e^\eta}{1+e^\eta} - \frac{e^\eta}{1+e^\eta}\right)
\end{aligned}
$$

$$= e^{y\cdot\eta} \cdot \left(\frac{1}{1+e^{\eta}}\right)$$
$$= \frac{e^{y\cdot\eta}}{1+e^{\eta}}$$

We know that $\eta = X \cdot \beta$.

$$= \frac{\exp(y \cdot X \cdot \beta)}{1+\exp(X \cdot \beta)}.$$

Part1 wts $\beta_j = 0 \implies Y \perp\!\!\!\perp X_j | X_{-j}$:

$$\mathbb{P}(Y = y, X_j \leq x_j | X_{-j} \leq x_{-j})$$

By Prep1

$$= \mathbb{P}(Y = y | X_j \leq x_j, X_{-j} \leq x_{-j}) \cdot \mathbb{P}(X_j \leq x_j | X_{-j} \leq x_{-j})$$

By Prep2

$$= \frac{\exp(Y \cdot (\beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p))}{1+\exp(\beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p)} \cdot \mathbb{P}(X_j \leq x_j | X_{-j} \leq x_{-j})$$

We know that $\beta_j = 0$.

$$= \frac{\exp(Y \cdot (\beta_1 \cdot X_1 + \ldots + \beta_{j-1} \cdot X_{j-1} + \beta_{j+1} \cdot X_{j+1} + \ldots + \beta_p \cdot X_p))}{1+\exp(\beta_1 \cdot X_1 + \ldots + \beta_{j-1} \cdot X_{j-1} + \beta_{j+1} \cdot X_{j+1} + \ldots + \beta_p \cdot X_p)} \cdot \mathbb{P}(X_j \leq x_j | X_{-j} \leq x_{-j})$$

We see that conditioning on X_j can be omitted.

$$= \mathbb{P}(Y = y | X_{-j} \leq x_{-j}) \cdot \mathbb{P}(X_j \leq x_j | X_{-j} \leq x_{-j}).$$

By definition of conditional independence, we proved that $Y \perp\!\!\!\perp X_j | X_{-j}$.

Part2 wts $Y \perp\!\!\!\perp X_j | X_{-j} \implies \beta_j = 0$:

Assume that Y and X_j are conditionally independent given X_{-j}.
This means it has to hold

$$\mathbb{P}(Y = y, X_j \leq x_j | X_{-j} \leq x_{-j}) = \mathbb{P}(Y = y | X_{-j} \leq x_{-j}) \cdot \mathbb{P}(X_j \leq x_j | X_{-j} \leq x_{-j}).$$

We can apply Prep2 to the LHS and get

$$\mathbb{P}(Y = y | X_j \leq x_j, X_{-j} \leq x_{-j}) \cdot \mathbb{P}(X_j \leq x_j | X_{-j} \leq x_{-j}) = \mathbb{P}(Y = y | X_{-j} \leq x_{-j}) \cdot \mathbb{P}(X_j \leq x_j | X_{-j} \leq x_{-j}).$$

Hence, we know that

$$\mathbb{P}(Y = y | X_j \leq x_j, X_{-j} \leq x_{-j}) = \mathbb{P}(X_j \leq x_j | X_{-j} \leq x_{-j}).$$

By Prep2, this means that

$$\frac{\exp(y \cdot X \cdot \beta)}{1+\exp(X \cdot \beta)} = \mathbb{P}(X_j \leq x_j | X_{-j} \leq x_{-j}). \tag{12}$$

Note that

$$\frac{\exp(y \cdot X \cdot \beta)}{1+\exp(X \cdot \beta)}$$
$$= \frac{\exp(y \cdot (\beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p))}{1+\exp(\beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p)}$$

$$
\begin{aligned}
&= \frac{\exp\left(y \cdot \beta_1 \cdot X_1 + \ldots + y \cdot \beta_j \cdot X_j + \ldots + y \cdot \beta_p \cdot X_p\right)}{1 + \exp\left(\beta_1 \cdot X_1 + \ldots + \beta_j \cdot X_j + \ldots + \beta_p \cdot X_p\right)} \\
&= \frac{\exp\left(y \cdot \beta_j \cdot X_j\right) \cdot \exp\left(y \cdot \beta_1 \cdot X_1 + \ldots + y \cdot \beta_{j-1} \cdot X_{j-1} + y \cdot \beta_{j+1} \cdot X_{j+1} + \ldots + y \cdot \beta_p \cdot X_p\right)}{1 + \exp\left(\beta_j \cdot X_j\right) \cdot \exp\left(\beta_1 \cdot X_1 + \ldots + \beta_{j-1} \cdot X_{j-1} + \beta_{j+1} \cdot X_{j+1} + \ldots + \beta_p \cdot X_p\right)} \\
&= \frac{\exp\left(y \cdot \beta_j \cdot X_j\right)}{1 + \exp\left(\beta_j \cdot X_j\right) \cdot \exp\left(\beta_1 \cdot X_1 + \ldots + \beta_{j-1} \cdot X_{j-1} + \beta_{j+1} \cdot X_{j+1} + \ldots + \beta_p \cdot X_p\right)} \\
&\quad \cdot \exp\left(y \cdot \beta_1 \cdot X_1 + \ldots + y \cdot \beta_{j-1} \cdot X_{j-1} + y \cdot \beta_{j+1} \cdot X_{j+1} + \ldots + y \cdot \beta_p \cdot X_p\right).
\end{aligned}
$$

For (12) to hold, we see that

$$
\frac{\exp\left(y \cdot \beta_j \cdot X_j\right)}{1 + \exp\left(\beta_j \cdot X_j\right) \cdot \exp\left(\beta_1 \cdot X_1 + \ldots + \beta_{j-1} \cdot X_{j-1} + \beta_{j+1} \cdot X_{j+1} + \ldots + \beta_p \cdot X_p\right)}
$$

has to be independent of X_j. This is only possible if $\exp\left(y \cdot \beta_j \cdot X_j\right)$ and $\exp\left(\beta_j \cdot X_j\right)$ vanish, because by assumption, X_j is not determined by X_{-j} (conditioning on X_{-j} does not determine X_j). Therefore, it has to hold $\beta_j = 0$.

□

Definition 3.5. Let $X = (X_1, \ldots, X_p)$ and $\tilde{X} = \left(\tilde{X}_1, \ldots, \tilde{X}_p\right)$ be row vectors. The function swap is defined such that the row vector $\left(X, \tilde{X}\right)_{\text{swap}(S)}$ is obtained from $\left(X, \tilde{X}\right)$ by swapping entries X_j and $\tilde{X}_j$ for each $j \in S$.

Comments:
i) Example with $p = 3$ and $S = \{2, 3\}$:

$$
\left(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3\right)_{\text{swap}(\{2,3\})} = \left(X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3\right).
$$

ii) We will use this swap function even for truncated versions of concatenated row vectors. E.g.

$$
\left(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2\right)_{\text{swap}(\{2\})} = \left(X_1, \tilde{X}_2, X_3, \tilde{X}_1, X_2\right).
$$

Of course, this is not possible for every S (swapping of a pair $(X_j, \tilde{X}_j)$ is only possible, if both elements are given in the concatenated row vector).
iii) Our definition is similar to Definition 2.3 in the fixed-X setting. In fact, we will use the model-X version to swap columns of the concatenated matrix $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]$ by simply defining

$$
\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)} = \begin{bmatrix} \left(X_{11}, \ldots, X_{1p}, \tilde{X}_{11}, \ldots, \tilde{X}_{1p}\right)_{\text{swap}(S)} \\ \vdots \\ \left(X_{n1}, \ldots, X_{np}, \tilde{X}_{n1}, \ldots, \tilde{X}_{np}\right)_{\text{swap}(S)} \end{bmatrix}.
$$

Definition 3.6. (Definition 3.1 in Candes et al. (2018)) Let $X = (X_1, \dots, X_p)$ be a family of random variables. *MX-Knockoffs* are a new family of random variables $\tilde{X} = \left(\tilde{X}_1, \dots, \tilde{X}_p\right)$ constructed with the following properties:
i) for any subset $S \subseteq \{1, \dots, p\}$,

$$\left(X, \tilde{X}\right)_{\text{swap}(S)} \stackrel{d}{=} \left(X, \tilde{X}\right) \tag{13}$$

ii) $Y \perp\!\!\!\perp \tilde{X} | X$ if there is a response Y.

Comments:
a) Note that ii) is guaranteed if $\tilde{X}$ is constructed without looking at Y.
b) We see from (13) that original and knockoff variables are pairwise exchangeable: taking any subset of variables and swapping them with their knockoffs leaves the joint distribution invariant.

Definition 3.7. The *MX-Knockoff matrix*

$$\tilde{\mathbf{X}} = \begin{pmatrix} \tilde{X}_{11} & \cdots & \tilde{X}_{1p} \\ \vdots & \ddots & \vdots \\ \tilde{X}_{n1} & \cdots & \tilde{X}_{np} \end{pmatrix}$$

is created in such a way that for each observation label i, $\left(\tilde{X}_{i1}, \dots, \tilde{X}_{ip}\right)$ is a MX-Knockoff for $(X_{i1}, \dots, X_{ip})$ (see Definition 3.6).

Comments:
i) This construction is different from what we did in the FX-Knockoff setting (see Definition 2.2).
ii) By assumption, rows of $(\mathbf{X}, \mathbf{Y})$ are independent. MX-Knockoffs $\tilde{X}$ are constructed for every row (X, Y) of $(\mathbf{X}, \mathbf{Y})$. This is why the rows of $\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right], \mathbf{Y}\right)$ are again independent.
iii) Do not get confused by the notation. Usually, we use the word "Knockoff" for a column $\tilde{\mathbf{X}}_j$ of $\tilde{\mathbf{X}}$. In Definition 3.6, we defined MX-Knockoffs as row vectors (rows of $\tilde{\mathbf{X}}$). These are two different things. The word "MX-Knockoffs" is maybe a bit misleading. In some cases we abuse the notation again and call elements of $\tilde{X}$ MX-Knockoffs.

Definition 3.8. For each $j \in \{1, \dots, p\}$, statistics $W_j = w_j\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right], \mathbf{Y}\right)$ are said to be *MX-Knockoff feature statistics* for some real valued measurable function w_j which satisfies:
("flip-sign property"): For any $S \subseteq \{1, \dots, p\}$,

$$w_j\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}, \mathbf{Y}\right) = \begin{cases} w_j\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right], \mathbf{Y}\right), & j \notin S, \\ -w_j\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right], \mathbf{Y}\right), & j \in S. \end{cases} \tag{14}$$

Comments:
i) The flip-sign property says that swapping $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ has the effect of switching the sign of W_j.
ii) In contrast to Definition 2.4, we do not require the sufficiency property that w_j depend on $\mathbf{X}$, $\tilde{\mathbf{X}}$, and $\mathbf{Y}$ only through $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right]$ and $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \mathbf{Y}$. Getting rid of this sufficiency property is very convenient, because the sufficiency property intensively restricts the choice of feature statistics (see subsection 2.4.3).

We now present two variable selection procedures for the model-X setting.

Definition 3.9. (Part of Theorem 3.4 in Candes et al. (2018)). Let $\tilde{\mathbf{X}}$ be a MX-Knockoff matrix (see Definition 3.7), let $W_1, \ldots, W_p$ be MX-Knockoff feature statistics (see Definition 3.8). Let

$$\mathcal{W} = \{|W_j| : \ W_j \neq 0, \ j \in \{1, \ldots, p\}\}$$

be the set of unique nonzero values attained by the $|W_j|$'s. Let $q \in [0, 1]$ be the target FDR level and let

$$T = \min\left\{t \in \mathcal{W} : \ \frac{|\{j : \ W_j \leq -t\}|}{|\{j : \ W_j \geq t\}| \vee 1} \leq q\right\}$$

(or $T = +\infty$ if this set is empty) be the data-dependent threshold. Then the *MX-Knockoff method* selects variables with indices

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

Comments:
i) Note that T and therefore $\hat{S}$ both implicitly depend on the choice of the target FDR level q.
ii) Definition 2.5 is the corresponding method in the fixed-X setting. Note that the MX-Knockoff method is exactly the same as the FX-Knockoff method, except that MX-Knockoff matrix and MX-Knockoff feature statistics are used instead of FX-Knockoff matrix and FX-Knockoff feature statistics.

Definition 3.10. (Part of Theorem 3.4 in Candes et al. (2018)). Let $\tilde{\mathbf{X}}$ be a MX-Knockoff matrix (see Definition 3.7), let $W_1, \ldots, W_p$ be MX-Knockoff feature statistics (see Definition 3.8). Let

$$\mathcal{W} = \{|W_j| : \ W_j \neq 0, \ j \in \{1, \ldots, p\}\}$$

be the set of unique nonzero values attained by the $|W_j|$'s. Let $q \in [0, 1]$ be the target FDR level and let

$$T_+ = \min\left\{t \in \mathcal{W} : \ \frac{1 + |\{j : \ W_j \leq -t\}|}{|\{j : \ W_j \geq t\}| \vee 1} \leq q\right\}$$

(or $T_+ = +\infty$ if this set is empty) be the data-dependent threshold. Then the *MX-Knockoff+ method* selects variables with indices

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

Comments:
Definition 2.6 is the corresponding method in the fixed-X setting. Note that the MX-Knockoff+ method is exactly the same as the FX-Knockoff+ method, except that MX-Knockoffs and MX-Knockoff feature statistics are used instead of FX-Knockoffs and FX-Knockoff feature statistics.

3.2 FDR Control

Definition 3.11. In the MX-Knockoff setting where we have a selection procedure returning a subset $\hat{S} \subseteq \{1, \dots, p\}$ of indices of variables,
i) the *false discovery proportion* (FDP) is defined as

$$\text{FDP} = \frac{|\left\{j : \ j \in (\hat{S} \cap \mathcal{H}_0)\right\}|}{|\hat{S}| \vee 1}.$$

ii) the *false discovery rate* (FDR) is defined as

$$\text{FDR} = \mathbb{E}\left[\text{FDP}\right] = \mathbb{E}\left[\frac{|\left\{j : \ j \in (\hat{S} \cap \mathcal{H}_0)\right\}|}{|\hat{S}| \vee 1}\right].$$

The next two Lemmas will help to prove FDR-control.

Lemma 3.12. *(Lemma 3.2 in Candes et al. (2018)). For any subset $S \subseteq \mathcal{H}_0$ of nulls it holds that*

$$\left[\mathbf{X}, \tilde{\mathbf{X}}\right] | \mathbf{Y} \ \stackrel{d}{=} \ \left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)} | \mathbf{Y}.$$

Proof:

Let $S \subseteq \mathcal{H}_0$ be an arbitrary but fixed set of nulls.

We have to show that $\left[\mathbf{X}, \tilde{\mathbf{X}}\right] | \mathbf{Y} \ \stackrel{d}{=} \ \left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)} | \mathbf{Y}$.

In more detail, we have to show that $\mathbb{P}\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right] \leq [\mathbf{x}, \tilde{\mathbf{x}}] \mid \mathbf{Y} \leq \mathbf{y}\right) = \mathbb{P}\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)} \leq [\mathbf{x}, \tilde{\mathbf{x}}] \mid \mathbf{Y} \leq \mathbf{y}\right)$.

By using the definition of conditional probability, we can write that as

$$\frac{\mathbb{P}\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right] \leq [\mathbf{x}, \tilde{\mathbf{x}}], \ \mathbf{Y} \leq \mathbf{y}\right)}{\mathbb{P}\left(\mathbf{Y} \leq \mathbf{y}\right)} = \frac{\mathbb{P}\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)} \leq [\mathbf{x}, \tilde{\mathbf{x}}], \ \mathbf{Y} \leq \mathbf{y}\right)}{\mathbb{P}\left(\mathbf{Y} \leq \mathbf{y}\right)}.$$

The denominator on both sides is the same, so it suffices to show that

$$\mathbb{P}\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right] \leq [\mathbf{x},\tilde{\mathbf{x}}]\,,\ \mathbf{Y} \leq \mathbf{y}\right) = \mathbb{P}\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right]_{\text{swap}(S)} \leq [\mathbf{x},\tilde{\mathbf{x}}]\,,\ \mathbf{Y} \leq \mathbf{y}\right). \tag{15}$$

Assume wlog $S = \{1, 2, \ldots, m\}$ for some $m \leq p$.

We know that $\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right],\mathbf{Y}\right)$ has n independent rows.
Therefore, we know that the cdf of $\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right],\mathbf{Y}\right)$ can be achieved by multiplying cdfs of all rows of $\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right],\mathbf{Y}\right)$.
Hence, it suffices to show (15) for an arbitrary row. We can use the property

$$\left[\mathbf{X},\tilde{\mathbf{X}}\right]_{\text{swap}(S)} = \begin{bmatrix} \left(X_{11},\ldots,X_{1p},\tilde{X}_{11},\ldots,\tilde{X}_{1p}\right)_{\text{swap}(S)} \\ \vdots \\ \left(X_{n1},\ldots,X_{np},\tilde{X}_{n1},\ldots,\tilde{X}_{np}\right)_{\text{swap}(S)} \end{bmatrix}$$

for the RHS of (15).
So, it remains to show that $\left(\left(X,\tilde{X}\right),Y\right) \overset{d}{=} \left(\left(X,\tilde{X}\right)_{\text{swap}(S)},Y\right)$, where $\left(X,\tilde{X},Y\right)$ is an arbitrary row of $\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right],\mathbf{Y}\right)$.

In more detail, we have to show that $\mathbb{P}\left(\left(X,\tilde{X}\right) \leq (x,\tilde{x})\,,\ Y \leq y\right) = \mathbb{P}\left(\left(X,\tilde{X}\right)_{\text{swap}(S)} \leq (x,\tilde{x})\,,\ Y \leq y\right)$.

By switching terms, we get $\mathbb{P}\left(Y \leq y\,,\ \left(X,\tilde{X}\right) \leq (x,\tilde{x})\right) = \mathbb{P}\left(Y \leq y\,,\ \left(X,\tilde{X}\right)_{\text{swap}(S)} \leq (x,\tilde{x})\right)$ and applying the definition of conditional probability on both sides, we get

$$\begin{aligned} &\mathbb{P}\left(Y \leq y \mid \left(X,\tilde{X}\right) \leq (x,\tilde{x})\right) \cdot \mathbb{P}\left(\left(X,\tilde{X}\right) \leq (x,\tilde{x})\right) \\ &= \mathbb{P}\left(Y \leq y \mid \left(X,\tilde{X}\right)_{\text{swap}(S)} \leq (x,\tilde{x})\right) \cdot \mathbb{P}\left(\left(X,\tilde{X}\right)_{\text{swap}(S)} \leq (x,\tilde{x})\right). \end{aligned} \tag{16}$$

By Definition 3.6 i), we know that for any S, it holds that $\left(X,\tilde{X}\right)_{\text{swap}(S)} \overset{d}{=} \left(X,\tilde{X}\right)$.
In other words, $\mathbb{P}\left(\left(X,\tilde{X}\right)_{\text{swap}(S)} \leq (x,\tilde{x})\right) = \mathbb{P}\left(\left(X,\tilde{X}\right) \leq (x,\tilde{x})\right)$.
This means that (16) simplifies and it remains to show

$$\mathbb{P}\left(Y \leq y \mid \left(X,\tilde{X}\right) \leq (x,\tilde{x})\right) = \mathbb{P}\left(Y \leq y \mid \left(X,\tilde{X}\right)_{\text{swap}(S)} \leq (x,\tilde{x})\right).$$

Note that instead of $\left(X_{\cdot 1},\ldots X_{\cdot p},\tilde{X}_{\cdot 1},\ldots,\tilde{X}_{\cdot p}\right)$, we use $\left(X_1,\ldots,X_p,\tilde{X}_1,\ldots,\tilde{X}_p\right)$ for an arbitrary row $\left(X,\tilde{X}\right)$ in the remaining proof (for simplicity).

Prep1 wts $\left\{\left(X,\tilde{X}\right)_{\text{swap}(S)} \leq (x,\tilde{x})\right\} = \left\{\left(X,\tilde{X}\right) \leq (x,\tilde{x})_{\text{swap}(S)}\right\}$:

To see this, consider an easy example with $\left(X_1,X_2,X_3,\tilde{X}_1,\tilde{X}_2,\tilde{X}_3\right)$ and $S = \{2,3\}$:
Case1:

$$\left(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3\right)_{\text{swap}(\{2,3\})} \leq (x_1, x_2, x_3, \tilde{x}_1, \tilde{x}_2, \tilde{x}_3)$$

which results in

$$\left(X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3\right) \leq (x_1, x_2, x_3, \tilde{x}_1, \tilde{x}_2, \tilde{x}_3).$$

Case2:

$$\left(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3\right) \leq (x_1, x_2, x_3, \tilde{x}_1, \tilde{x}_2, \tilde{x}_3)_{\text{swap}(\{2,3\})}$$

which results in

$$\left(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3\right) \leq (x_1, \tilde{x}_2, \tilde{x}_3, \tilde{x}_1, x_2, x_3).$$

In both cases, we get $X_1 \leq x_1, X_2 \leq \tilde{x}_2, X_3 \leq \tilde{x}_3, \tilde{X}_1 \leq \tilde{x}_1, \tilde{X}_2 \leq x_2, \tilde{X}_3 \leq x_3$.

Prep2 wts $\mathbb{P}\left(Y \leq y | \tilde{X} \leq \tilde{x}, X \leq x\right) = \mathbb{P}\left(Y \leq y | X \leq x\right)$ for all $y, \tilde{x}, x$:

We know from Definition 3.6 ii) that $Y \perp\!\!\!\perp \tilde{X} | X$.
By definition of conditional independence (see Definition B.3), this is $\mathbb{P}\left(Y \leq y | \tilde{X} \leq \tilde{x}, X \leq x\right) = \mathbb{P}\left(Y \leq y | X \leq x\right)$ for all $y, \tilde{x}, x$.

Prep3 wts $\mathbb{P}\left(Y \leq y | X_i \leq x_i, X_{-i} \leq x_{-i}\right) = \mathbb{P}\left(Y \leq y | X_{-i} \leq x_{-i}\right)$ for all x_i if i is null:

We know that X_i is null. By Definition 3.2, this means that $Y \perp\!\!\!\perp X_i | X_{-i}$.
By Definition B.3, this means that
$\mathbb{P}\left(Y \leq y | X_i \leq x_i, X_{-i} \leq x_{-i}\right) = \mathbb{P}\left(Y \leq y | X_{-i} \leq x_{-i}\right)$ or $\mathbb{P}\left(X_i \leq x_i, X_{-i} \leq x_{-i}\right) = 0$.
Note that $\mathbb{P}\left(X_i \leq x_i, X_{-i} \leq x_{-i}\right) = \mathbb{P}\left(X \leq x\right)$.
If $\mathbb{P}\left(X \leq x\right)$ would be zero, the term in Prep2 would not be defined (conditioning on an event which has probability zero is not possible). Hence, we implicitly consider the case where $\mathbb{P}\left(X \leq x\right) > 0$ in this proof.
Therefore, we have $\mathbb{P}\left(Y \leq y | X_i \leq x_i, X_{-i} \leq x_{-i}\right) = \mathbb{P}\left(Y \leq y | X_{-i} \leq x_{-i}\right)$ for all x_i.

$$\mathbb{P}\left(Y \leq y \mid \left(X, \tilde{X}\right)_{\text{swap}(S)} \leq (x, \tilde{x})\right)$$

By Prep1

$$= \mathbb{P}\left(Y \leq y \mid \left(X, \tilde{X}\right) \leq (x, \tilde{x})_{\text{swap}(S)}\right)$$

$$= \mathbb{P}\left(Y \leq y \mid \left(X_1, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_p\right) \leq (x_1, \ldots, x_p, \tilde{x}_1, \ldots, \tilde{x}_p)_{\text{swap}(S)}\right)$$

Remember that $S = \{1, 2, \ldots, m\}$.

$$= \mathbb{P}\left(Y \leq y \mid \left(X_1, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_p\right) \leq (\tilde{x}_1, \ldots, \tilde{x}_m, x_{m+1}, \ldots, x_p, x_1, \ldots, x_m, \tilde{x}_{m+1}, \ldots, \tilde{x}_p)\right)$$

$$= \mathbb{P}\left(Y \leq y \mid (X_1, \ldots, X_p) \leq (\tilde{x}_1, \ldots, \tilde{x}_m, x_{m+1}, \ldots, x_p),\ \left(\tilde{X}_1, \ldots, \tilde{X}_p\right) \leq (x_1, \ldots, x_m, \tilde{x}_{m+1}, \ldots, \tilde{x}_p)\right)$$

By Prep2

$$= \mathbb{P}\left(Y \leq y \mid (X_1, \ldots, X_p) \leq (\tilde{x}_1, \ldots, \tilde{x}_m, x_{m+1}, \ldots, x_p)\right)$$

Define $x'_i = \begin{cases} \tilde{x}_i, & i \in S \\ x_i, & i \notin S \end{cases}$ for $i \in \{1, \dots, p\}$.

$$= \mathbb{P}\left(Y \leq y \mid (X_1, \dots, X_p) \leq (x'_1, \dots, x'_m, x'_{m+1}, \dots, x'_p)\right)$$

We know that $\{1\} \in S$.

$$= \mathbb{P}\left(Y \leq y \mid (X_1, \dots, X_p) \leq (\tilde{x}_1, x'_2, x'_3, \dots, x'_p)\right)$$
$$= \mathbb{P}\left(Y \leq y \mid X_1 \leq \tilde{x}_1, X_2 \leq x'_2, X_3 \leq x'_3, \dots, X_p \leq x'_p\right)$$

We know that 1 is null.
By Prep3

$$= \mathbb{P}\left(Y \leq y \mid X_2 \leq x'_2, X_3 \leq x'_3, \dots, X_p \leq x'_p\right)$$

We reverse what we just did. And because Prep3 holds for all possible x_i, we can choose x_1.

$$= \mathbb{P}\left(Y \leq y \mid X_1 \leq x_1, X_2 \leq x'_2, X_3 \leq x'_3, \dots, X_p \leq x'_p\right)$$
$$= \mathbb{P}\left(Y \leq y \mid (X_1, X_2, \dots, X_p) \leq \left(x_1, x'_2, \dots, x'_p\right)\right)$$

We repeat the whole argument with $\{2\} \in S$.

$$= \mathbb{P}\left(Y \leq y \mid (X_1, X_2, \dots, X_p) \leq \left(x_1, x_2, x'_3, \dots, x'_p\right)\right)$$

Repeating this argument until S is empty.

$$= \mathbb{P}\left(Y \leq y \mid (X_1, X_2, \dots, X_p) \leq (x_1, x_2, x_3, \dots, x_p)\right)$$
$$= \mathbb{P}\left(Y \leq y \mid X \leq x\right)$$

By Prep2

$$= \mathbb{P}\left(Y \leq y | \tilde{X} \leq \tilde{x}, X \leq x\right)$$
$$= \mathbb{P}\left(Y \leq y \mid \left(X, \tilde{X}\right) \leq (x, \tilde{x})\right).$$

□

Lemma 3.13. *(Lemma 3.3 in Candes et al. (2018)). Let* $\mathbf{W} = (W_1, \dots, W_p)$ *be a vector of MX-Knockoff feature statistics (see Definition 3.8). Let* $W_j \neq 0$ *for all* $j \in \{1, \dots, p\}$*. Then:*
i) For $j \in \mathcal{H}_0$*:* $\text{sign}(W_j)$ *are mutually (jointly) independent and* $\mathbb{P}(\text{sign}(W_j) = -1) = \mathbb{P}(\text{sign}(W_j) = +1) = \frac{1}{2}$*,*
ii) Part i) holds even conditionally on $\mathbf{Y}$*.*

Proof:

For any set $S \subseteq \{1, \dots, p\}$, let $\mathbf{W}_{\text{swap}(S)}$ be the statistic we would get if we had replaced $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]$ with $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}$ when calculating $\mathbf{W}$.

By (14), we know that

$$w_j\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}, \mathbf{Y}\right) = \begin{cases} w_j\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right], \mathbf{Y}\right), & j \notin S \\ -w_j\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right], \mathbf{Y}\right), & j \in S. \end{cases}$$

We can write that as

$$w_j\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}, \mathbf{Y}\right) = w_j\left(\left[\mathbf{X}, \tilde{\mathbf{X}}\right], \mathbf{Y}\right) \cdot \delta_j, \quad \text{where } \delta_j = \begin{cases} +1, & j \notin S \\ -1, & j \in S \end{cases}. \tag{17}$$

Define a measurable function $w:\ \mathbb{R}^{n\times(2\cdot p+1)} \to \mathbb{R}^p$ with $w\left[(u,v)\right] = (w_1(u,v),\ldots,w_p(u,v))$, where w_j are the functions from Definition 3.8.
Define $\mathbf{W}_{\mathrm{swap}(S)} = w\left[\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right]_{\mathrm{swap}(S)},\mathbf{Y}\right)\right]$.

Prep1 wts $\mathbf{W} \stackrel{d}{=} (W_1\cdot\delta_1, W_2\cdot\delta_2,\ldots,W_p\cdot\delta_p)$ for an arbitrary set of nulls S:

Let $S\subseteq\{1,\ldots,p\}$ be an arbitrary set of nulls.

Prep1.1 wts $\mathbf{W}_{\mathrm{swap}(S)} = (W_1\cdot\delta_1, W_2\cdot\delta_2,\ldots,W_p\cdot\delta_p)$:

$$
\begin{aligned}
&\mathbf{W}_{\mathrm{swap}(S)}\\
&= w\left[\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right]_{\mathrm{swap}(S)},\mathbf{Y}\right)\right]\\
&= \left(w_1\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right]_{\mathrm{swap}(S)},\mathbf{Y}\right), w_2\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right]_{\mathrm{swap}(S)},\mathbf{Y}\right),\ldots, w_p\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right]_{\mathrm{swap}(S)},\mathbf{Y}\right)\right)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{By (17)}\\
&= \left(w_1\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right],\mathbf{Y}\right)\cdot\delta_1, w_2\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right],\mathbf{Y}\right)\cdot\delta_2,\ldots, w_p\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right],\mathbf{Y}\right)\cdot\delta_p\right)\\
&= (W_1\cdot\delta_1, W_2\cdot\delta_2,\ldots,W_p\cdot\delta_p)\,.
\end{aligned}
$$

Prep1.2 wts $\mathbf{W}_{\mathrm{swap}(S)} \stackrel{d}{=} \mathbf{W}$:

Let $\mathbf{z} = (z_1,\ldots,z_p)$ be an arbitrary real row-vector of length p.

$$
\begin{aligned}
&\mathbb{P}\left(\mathbf{W}_{\mathrm{swap}(S)} \le \mathbf{z}\right)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{By Definition of } \mathbf{W}_{\mathrm{swap}(S)}\\
&= \mathbb{P}\left(w\left[\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right]_{\mathrm{swap}(S)},\mathbf{Y}\right)\right] \le \mathbf{z}\right)\\
&= \mathbb{P}\left(w\left[\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right]_{\mathrm{swap}(S)},\mathbf{Y}\right)\right] \in (-\infty,z_1]\times\ldots\times(-\infty,z_p]\right)\\
&= \mathbb{P}\left(\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right]_{\mathrm{swap}(S)},\mathbf{Y}\right) \in w^{-1}((-\infty,z_1]\times\ldots\times(-\infty,z_p])\right)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{In the proof of Lemma 3.12,}\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{we saw that for } S\subseteq\mathcal{H}_0,\text{ it holds}\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{that } \left[\mathbf{X},\tilde{\mathbf{X}}\right],\mathbf{Y} \stackrel{d}{=} \left[\mathbf{X},\tilde{\mathbf{X}}\right]_{\mathrm{swap}(S)},\mathbf{Y}.\\
&= \mathbb{P}\left(\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right],\mathbf{Y}\right) \in w^{-1}((-\infty,z_1]\times\ldots\times(-\infty,z_p])\right)\\
&= \mathbb{P}\left(w\left[\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right],\mathbf{Y}\right)\right] \in (-\infty,z_1]\times\ldots\times(-\infty,z_p]\right)\\
&= \mathbb{P}\left(w\left[\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right],\mathbf{Y}\right)\right] \le \mathbf{z}\right)\\
&= \mathbb{P}\left(\mathbf{W} \le \mathbf{z}\right).
\end{aligned}
$$

Prep1.1 and Prep1.2 together give $(W_1\cdot\delta_1, W_2\cdot\delta_2,\ldots,W_p\cdot\delta_p) = \mathbf{W}_{\mathrm{swap}(S)} \stackrel{d}{=} \mathbf{W}$.
Hence, $\mathbf{W} \stackrel{d}{=} (W_1\cdot\delta_1, W_2\cdot\delta_2,\ldots,W_p\cdot\delta_p)$.

For the rest of the proof, let wlog $\{1,\dots,m\}$ be the set of null indices and $\{m+1,\dots,p\}$ be the set of non-null indices.

ad i) wts $(\text{sign}(W_1),\dots,\text{sign}(W_m))$ are independent and $\mathbb{P}(\text{sign}(W_j)=-1)=\mathbb{P}(\text{sign}(W_j)=1)=\frac{1}{2}$ for all $j\in\{1,\dots,m\}$:

Part1.1 wts $(\text{sign}(W_1),\dots,\text{sign}(W_m))$ are independent:

We have to show that $\mathbb{P}(\text{sign}(W_1)=k_1,\dots,\text{sign}(W_m)=k_m)=\prod_{j=1}^m\mathbb{P}(\text{sign}(W_j)=k_j)$ for $k_1,\dots,k_m\in\{-1,0,+1\}$.

Since $W_j\neq 0$ for all $j\in\{1,\dots,p\}$ by assumption, it suffices to show this for all $k_1,\dots,k_m\in\{-1,+1\}$.

Note that $\{\text{sign}(W_j)=k_j\}=\{W_j\cdot k_j>0\}$ for $j\in\{1,\dots,p\}$, because:

Case1 $k_j=1$:

$$\begin{aligned}
&\{\text{sign}(W_j)=k_j\}\\
&\qquad\qquad\text{In this case, } k_j=1.\\
&=\{\text{sign}(W_j)=1\}\\
&\qquad\qquad\text{By Definition A.1 (sign-function)}\\
&=\{W_j>0\}\\
&=\{W_j\cdot 1>0\}\\
&\qquad\qquad\text{In this case, } k_j=1.\\
&=\{W_j\cdot k_j>0\}.
\end{aligned}$$

Case2 $k_j=-1$:

$$\begin{aligned}
&\{\text{sign}(W_j)=k_j\}\\
&\qquad\qquad\text{In this case, } k_j=-1.\\
&=\{\text{sign}(W_j)=-1\}\\
&\qquad\qquad\text{By Definition A.1 (sign-function)}\\
&=\{W_j<0\}\\
&=\{W_j\cdot(-1)>0\}\\
&\qquad\qquad\text{In this case, } k_j=-1.\\
&=\{W_j\cdot k_j>0\}.
\end{aligned}$$

The same holds for the whole random vector.
Hence, it suffices to show

$$\mathbb{P}(W_1\cdot k_1>0,\dots,W_m\cdot k_m>0)=\prod_{j=1}^m\mathbb{P}(W_j\cdot k_j>0)\quad\text{for all } k_1,\dots,k_m\in\{-1,+1\}. \tag{18}$$

By Prep1 we know that $(W_1,\dots,W_p)\overset{d}{=}(W_1\cdot\delta_1,\dots,W_p\cdot\delta_p)$ for $\delta_j=\begin{cases}+1, & j\notin S\\ -1, & j\in S\end{cases}$ and all $S\subseteq\{1,\dots,m\}$.
In particular, $\mathbb{P}(W_1\cdot k_1>0,\dots,W_m\cdot k_m>0)=\mathbb{P}(W_1>0,\dots,W_m>0)$ for all $k_1,\dots,k_m\in\{-1,+1\}$ (take $S=\{j\in\{1,\dots,m\}:\ k_j=-1\}$).
And of course, $\mathbb{P}(W_j\cdot k_j>0)=\mathbb{P}(W_j>0)$ for all $k_j\in\{-1,+1\}$, $j\in\{1,\dots,m\}$.

We now prove (18) by showing that both sides of (18) are equal to $\frac{1}{2^m}$.
First, consider the LHS of (18):

1

The sum of all possible probabilities has to be 1.

$$= \sum_{k_1,\ldots,k_m \in \{-1,1\}} \mathbb{P}\left(W_1 \cdot k_1 > 0, \ldots, W_m \cdot k_m > 0\right)$$

We know that $\mathbb{P}\left(W_1 \cdot k_1 > 0, \ldots, W_m \cdot k_m > 0\right) = \mathbb{P}\left(W_1 > 0, \ldots, W_m > 0\right)$.

$$= \sum_{k_1,\ldots,k_m \in \{-1,1\}} \mathbb{P}\left(W_1 > 0, \ldots, W_m > 0\right)$$

The probability term does not depend on $k_1, \ldots, k_m$ anymore.

$$= \mathbb{P}\left(W_1 > 0, \ldots, W_m > 0\right) \cdot \sum_{k_1,\ldots,k_m \in \{-1,1\}} 1$$

There are 2^m different combinations of $k_1, \ldots, k_m$.

$$= \mathbb{P}\left(W_1 > 0, \ldots, W_m > 0\right) \cdot 2^m.$$

Rearranged gives $\mathbb{P}\left(W_1 > 0, \ldots, W_m > 0\right) = \frac{1}{2^m}$.
And since we know $\mathbb{P}\left(W_1 \cdot k_1 > 0, \ldots, W_m \cdot k_m > 0\right) = \mathbb{P}\left(W_1 > 0, \ldots, W_m > 0\right)$ for all $k_1, \ldots, k_m \in \{-1, +1\}$, we get $\mathbb{P}\left(W_1 \cdot k_1 > 0, \ldots, W_m \cdot k_m > 0\right) = \frac{1}{2^m}$ for all $k_1, \ldots, k_m \in \{-1, +1\}$.

Second, consider the RHS of (18):

Let $j \in \{1, \ldots, m\}$ be arbitrary.
By assumption, $W_j \neq 0$.
So, we know that $\mathbb{P}\left(W_j = 0\right) = 0$ and hence it has to hold $\mathbb{P}\left(W_j > 0\right) + \mathbb{P}\left(W_j < 0\right) = 1$.
In other words, $\mathbb{P}\left(W_j > 0\right) + \mathbb{P}\left(W_j \cdot (-1) > 0\right) = 1$.
Since we know $\mathbb{P}\left(W_j \cdot k_j > 0\right) = \mathbb{P}\left(W_j > 0\right)$ for all $k_j \in \{-1, +1\}$ from above, we get $\mathbb{P}\left(W_j > 0\right) = \mathbb{P}\left(W_j \cdot (-1) > 0\right)$.
Hence, $\mathbb{P}\left(W_j > 0\right) = \mathbb{P}\left(W_j < 0\right) = \frac{1}{2}$.
Since $j \in \{1, \ldots, m\}$ was arbitrary, we have shown $\mathbb{P}\left(W_j > 0\right) = \mathbb{P}\left(W_j < 0\right) = \frac{1}{2}$ for all $j \in \{1, \ldots, m\}$.

Let $k_1, \ldots, k_m \in \{-1, +1\}$ be arbitrary.

$$\prod_{j=1}^{m} \mathbb{P}\left(W_j \cdot k_j > 0\right)$$

We know that $\mathbb{P}\left(W_j \cdot k_j > 0\right) = \mathbb{P}\left(W_j > 0\right)$ for all $k_j \in \{-1, +1\}$.

$$= \prod_{j=1}^{m} \mathbb{P}\left(W_j > 0\right)$$

We know that $\mathbb{P}\left(W_j > 0\right) = \frac{1}{2}$ for all $j \in \{1, \ldots, m\}$.

$$= \prod_{j=1}^{m} \frac{1}{2}$$
$$= \left(\frac{1}{2}\right)^m$$
$$= \frac{1^m}{2^m}$$
$$= \frac{1}{2^m}.$$

Part1.2 wts $\mathbb{P}(\text{sign}(W_j) = -1) = \mathbb{P}(\text{sign}(W_j) = 1) = \frac{1}{2}$ for all null j:

We know from Part1.1 that $\mathbb{P}(W_j < 0) = \mathbb{P}(W_j > 0) = \frac{1}{2}$ for all null j.
This is equivalent to $\mathbb{P}(\text{sign}(W_j) = -1) = \mathbb{P}(\text{sign}(W_j) = 1) = \frac{1}{2}$ for all null j.

ad ii) wts Conditionally on $\mathbf{Y}$, it holds that $\text{sign}(W_j) \overset{i.i.d.}{\sim} \{\pm 1\}$ for null j:

By Lemma 3.12, we know that $\left[\mathbf{X}, \tilde{\mathbf{X}}\right] | \mathbf{Y} \overset{d}{=} \left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)} | \mathbf{Y}$ for any subset S of nulls. We saw in the proof of that Lemma, that this implies that $\left[\mathbf{X}, \tilde{\mathbf{X}}\right], \mathbf{Y} \overset{d}{=} \left[\mathbf{X}, \tilde{\mathbf{X}}\right]_{\text{swap}(S)}, \mathbf{Y}$ for any subset S of nulls. That is what we used in Prep1.2 anyway. Hence, Prep1 holds even conditionally on $\mathbf{Y}$. Therefore, we can prove Part1.2 and Part1.2 even conditionally on $\mathbf{Y}$.

□

Comments:
Lemma 2.11 is the corresponding Lemma in the fixed-X setting. Note that the proof of i) in both Lemmas is almost identical.

Theorem 3.14. *(Part of Theorem 3.4 in Candes et al. (2018)). For any $q \in (0, 1)$, the MX-Knockoff method (see Definition 3.9) which is selecting variables with indices*

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

satisfies

$$\text{mFDR} := \mathbb{E}\left[\frac{|\left\{j : \ j \in (\hat{S} \cap \mathcal{H}_0)\right\}|}{|\hat{S}| + 1/q}\right] \leq q.$$

The result is non-asymptotic and holds no matter the dependence between the response and the covariates - in fact, it holds even conditionally on the response $\mathbf{Y}$.

Proof:

Note that this proof is the same as the proof of Theorem 2.12, except that we use properties of MX-Knockoff feature statistics instead of FX-Knockoff feature statistics and a different definition of nulls ($\mathcal{H}_0$ instead of $\{j : \ \beta_j = 0\}$).

We only prove the case where:
- $W_j \neq 0$ for all j (hence, the assumptions of Lemma 3.13 are satisfied).
- $|W_j|$ are distinct, i.e., $|W_i| \neq |W_j|$ for $i \neq j$.

- $T < \infty$.

The proof is split into five parts. Part2 and Part3 are martingale results (needed to use optional stopping theorem). Part1 and Part4 are inequalities we are going to use in Part5, where we finally prove FDR-control.

Let $q \in (0,1)$ be arbitrary but fixed.

We would like to assume that $|W_j|$ are sorted in increasing order to make the proof easier. Such sorting implies that null indices change too.
To address this problem, we define random indices $[1], \ldots, [p]$ through $|W_{[1]}| < |W_{[2]}| < \ldots < |W_{[p]}|$.
The concept is similar to order statistics, except that we use it for ordered absolute values of terms (e.g. for $W_1 = 3$, $W_2 = -1$ and $W_3 = -4$, we have $|W_{[1]}| < |W_{[2]}| < |W_{[3]}|$ where $W_{[1]} = W_2 = -1$, $W_{[2]} = W_1 = 3$ and $W_{[3]} = W_3 = -4$).

For each $k \in \{1, \ldots, p\}$, define

$$A(k) := |\left\{j : \ j \in \{k, \ldots, p\}\,, \ W_{[j]} < 0\right\}|,$$

$$B(k) := |\left\{j : \ j \in \{k, \ldots, p\}\,, \ W_{[j]} > 0\right\}|.$$

In other words, $A(k)$ is the number of negative W_j among the $p - (k-1)$ biggest $|W_j|$.
And $B(k)$ is the number of positive W_j among the $p - (k-1)$ biggest $|W_j|$.

Hence, we can write

$$A(k) = |\left\{j : \ |W_j| \geq |W_{[k]}|, \ W_j < 0\right\}|,$$

$$B(k) = |\left\{j : \ |W_j| \geq |W_{[k]}|, \ W_j > 0\right\}|.$$

In Definition 3.9, we have $\mathcal{W} = \{|W_j| : \ W_j \neq 0, \ j \in \{1, \ldots, p\}\}$.
Since we assume $W_j \neq 0$ in this proof, we can write that as $\mathcal{W} = \left\{|W_{[j]}| : \ j \in \{1, \ldots, p\}\right\}$.

Prep0.1 wts $\left\{j : \ W_j \leq -|W_{[k]}|\right\} = \left\{j : \ |W_j| \geq |W_{[k]}|, \ W_j < 0\right\}$ for $k \in \{1, \ldots, p\}$:

Let $k \in \{1, \ldots, p\}$ be arbitrary.
wts $\left\{j : \ W_j \leq -|W_{[k]}|\right\} \subseteq \left\{j : \ |W_j| \geq |W_{[k]}|, \ W_j < 0\right\}$:

Let $j \in \left\{j : \ W_j \leq -|W_{[k]}|\right\}$ be arbitrary.
This means that $W_j \leq -|W_{[k]}|$.
Since $W_j \neq 0$ for all j, this implies that $W_j < 0$.
Multiplying (-1) on both sides of $W_j \leq -|W_{[k]}|$ gives $-W_j \geq |W_{[k]}|$.
And since $W_j < 0$, we have $-W_j > 0$.
Therefore, $|-W_j| = -W_j$.
And of course (property of absolute value), $|-W_j| = |W_j|$.
Hence, $|W_j| \geq |W_{[k]}|$.
So, we have $|W_j| \geq |W_{[k]}|$ and $W_j < 0$.
Hence, $j \in \left\{j : \ |W_j| \geq |W_{[k]}|, \ W_j < 0\right\}$.

wts $\left\{j : \ W_j \leq -|W_{[k]}|\right\} \supseteq \left\{j : \ |W_j| \geq |W_{[k]}|, \ W_j < 0\right\}$:

Let $j \in \left\{j : \ |W_j| \geq |W_{[k]}|, \ W_j < 0\right\}$ be arbitrary.
This means that $|W_j| \geq |W_{[k]}|$ and $W_j < 0$.
Multiplying (-1) on both sides of $|W_j| \geq |W_{[k]}|$ gives $-|W_j| \leq -|W_{[k]}|$.
And since $W_j < 0$, we have $|W_j| = -W_j$ (by definition of the absolute value).
Hence, $-|W_j| = -(-W_j) = W_j$.
Therefore, $W_j \leq -|W_{[k]}|$.
Hence, $j \in \left\{j : \ W_j \leq -|W_{[k]}|\right\}$.

Both together give $\{j:\ W_j \le -|W_{[k]}|\} = \{j:\ |W_j| \ge |W_{[k]}|,\ W_j < 0\}$.
And since $k \in \{1,\dots,p\}$ was arbitrary, we have shown this result for all $k \in \{1,\dots,p\}$.

Prep0.2 wts $\{j:\ W_j \ge |W_{[k]}|\} = \{j:\ |W_j| \ge |W_{[k]}|,\ W_j > 0\}$ for $k \in \{1,\dots,p\}$:

Let $k \in \{1,\dots,p\}$ be arbitrary.
wts $\{j:\ W_j \ge |W_{[k]}|\} \subseteq \{j:\ |W_j| \ge |W_{[k]}|,\ W_j > 0\}$:

Let $j \in \{j:\ W_j \ge |W_{[k]}|\}$ be arbitrary.
This means that $W_j \ge |W_{[k]}|$.
Since $W_j \ne 0$ for all j, this implies that $W_j > 0$.
Therefore, $|W_j| = W_j$.
So, we have $|W_j| \ge |W_{[k]}|$.
Hence, $j \in \{j:\ |W_j| \ge |W_{[k]}|,\ W_j > 0\}$.

wts $\{j:\ W_j \ge |W_{[k]}|\} \supseteq \{j:\ |W_j| \ge |W_{[k]}|,\ W_j > 0\}$:

Let $j \in \{j:\ |W_j| \ge |W_{[k]}|,\ W_j > 0\}$ be arbitrary.
This means that $|W_j| \ge |W_{[k]}|$ and $W_j > 0$.
Since $W_j > 0$, we have $|W_j| = W_j$.
Therefore, $W_j \ge |W_{[k]}|$.
Hence, $j \in \{j:\ W_j \ge |W_{[k]}|\}$.

Both together give $\{j:\ W_j \ge |W_{[k]}|\} = \{j:\ |W_j| \ge |W_{[k]}|,\ W_j > 0\}$.
And since $k \in \{1,\dots,p\}$ was arbitrary, we have shown this result for all $k \in \{1,\dots,p\}$.

Then

$$
\begin{aligned}
& T \\
& \qquad \text{By Definition 3.9} \\
&= \min\left\{ t \in \mathcal{W}:\ \frac{|\{j:\ W_j \le -t\}|}{|\{j:\ W_j \ge t\}| \vee 1} \le q \right\} \\
& \qquad \text{We know that } \mathcal{W} = \left\{ |W_{[j]}|:\ j \in \{1,\dots,p\} \right\}. \\
&= \min\left\{ |W_{[k]}|:\ k \in \{1,\dots,p\},\ \frac{|\{j:\ W_j \le -|W_{[k]}|\}|}{|\{j:\ W_j \ge |W_{[k]}|\}| \vee 1} \le q \right\} \\
& \qquad \text{By Prep0.1 and Prep0.2, we know} \\
& \qquad \text{that } \{j:\ W_j \le -|W_{[k]}|\} = \{j:\ |W_j| \ge |W_{[k]}|,\ W_j < 0\} \\
& \qquad \text{and } \{j:\ W_j \ge |W_{[k]}|\} = \{j:\ |W_j| \ge |W_{[k]}|,\ W_j > 0\}. \\
&= \min\left\{ |W_{[k]}|:\ k \in \{1,\dots,p\},\ \frac{|\{j:\ |W_j| \ge |W_{[k]}|,\ W_j < 0\}|}{|\{j:\ |W_j| \ge |W_{[k]}|,\ W_j > 0\}| \vee 1} \le q \right\} \\
& \qquad \text{By definition of } A(k) \text{ and } B(k) \\
&= \min\left\{ |W_{[k]}|:\ k \in \{1,\dots,p\},\ \frac{A(k)}{B(k) \vee 1} \le q \right\}.
\end{aligned}
$$

Set

$$K := \min\left\{ k:\ k \in \{1,\dots,p\},\ \frac{A(k)}{B(k) \vee 1} \le q \right\}.$$

Since $T < \infty$, we know that at least one $k \in \{1,\dots,p\}$ satisfies the inequality in the definition of K.
We know from above that T is equal to the smallest $|W_{[k]}|$ such that this inequality is satisfied.
And we know that $|W_{[1]}| < |W_{[2]}| < \ldots < |W_{[p]}|$.
Therefore,

$$T = |W_{[K]}|.$$

Side note: If we would relax the assumption $|W_{[1]}| < |W_{[2]}| < \ldots < |W_{[p]}|$ to $|W_{[1]}| \leq |W_{[2]}| \leq \ldots \leq |W_{[p]}|$, then $A(k)$, $B(k)$ and therefore K would not be unique in general, as a small example shows:

Take $p = 3$ with $W_1 = 3$, $W_2 = -1$ and $W_3 = -3$. Then we can choose if we write $|W_2| \leq |W_1| \leq |W_3|$ or $|W_2| \leq |W_3| \leq |W_1|$.

Case1 $W_{[1]} = W_2$, $W_{[2]} = W_1$, $W_{[3]} = W_3$:

Here, we have
$A(3) = |\{j : j \in \{3\}, W_{[j]} < 0\}| = |\{3\}| = 1$ and
$B(3) = |\{j : j \in \{3\}, W_{[j]} > 0\}| = |\emptyset| = 0$.
Hence, in particular,

$$\frac{A(3)}{B(3) \vee 1} = \frac{1}{0 \vee 1} = 1.$$

Case2 $W_{[1]} = W_2$, $W_{[2]} = W_3$, $W_{[3]} = W_1$:

Here, we have
$A(3) = |\{j : j \in \{3\}, W_{[j]} < 0\}| = |\emptyset| = 0$ and
$B(3) = |\{j : j \in \{3\}, W_{[j]} > 0\}| = |\{3\}| = 1$.
Hence, in particular,

$$\frac{A(3)}{B(3) \vee 1} = \frac{0}{1 \vee 1} = 0.$$

To continue the proof, note that

$$\begin{aligned}
&\hat{S} \\
&= \{j : W_j \geq T\} \\
&\qquad\qquad \text{We know that } T = |W_{[K]}|. \\
&= \{j : W_j \geq |W_{[K]}|\} \\
&\qquad\qquad \text{By the same argument as in Prep0.2} \\
&= \{j : |W_j| \geq |W_{[K]}|, W_j > 0\} \\
&\qquad\qquad \text{By the same argument as above, where} \\
&\qquad\qquad \text{we introduced a different formula for } B(k). \\
&= \{j : j \in \{K, \ldots, p\}, W_{[j]} > 0\}.
\end{aligned}$$

Prep1.1 wts $1 \leq \dfrac{1 + |\{j : j \in \{K, \ldots, p\}, W_{[j]} < 0\}|}{1 + |\{j : j \in \{K, \ldots, p\}, W_{[j]} < 0, [j] \in \mathcal{H}_0\}|}$:

It is clear, that $\{j : j \in \{K, \ldots, p\}, W_{[j]} < 0\} \supseteq \{j : j \in \{K, \ldots, p\}, W_{[j]} < 0, [j] \in \mathcal{H}_0\}$.
Hence, $|\{j : j \in \{K, \ldots, p\}, W_{[j]} < 0\}| \geq |\{j : j \in \{K, \ldots, p\}, W_{[j]} < 0, [j] \in \mathcal{H}_0\}|$.
So, $1 + |\{j : j \in \{K, \ldots, p\}, W_{[j]} < 0\}| \geq 1 + |\{j : j \in \{K, \ldots, p\}, W_{[j]} < 0, [j] \in \mathcal{H}_0\}|$.
Rearranged gives

$$1 \leq \frac{1 + |\{j : j \in \{K, \ldots, p\}, W_{[j]} < 0\}|}{1 + |\{j : j \in \{K, \ldots, p\}, W_{[j]} < 0, [j] \in \mathcal{H}_0\}|}.$$

Prep1.2 wts $\dfrac{1 + A(K)}{B(K) + 1/q} \leq q$:

We know that

$$K = \min\left\{k : \ k \in \{1, \ldots, p\}, \ \frac{A(k)}{B(k) \vee 1} \leq q\right\}.$$

Therefore, K is one such $k \in \{1, \ldots, p\}$ satisfying this inequality. So, it holds

$$\frac{A(K)}{B(K) \vee 1} \leq q.$$

Note that $B(K) > 0$, because:

Assume otherwise, $B(K) = 0$.
Note that $A(k) + B(k) = p - (k-1)$ for all $k \in \{1, \ldots, p\}$.
So, $A(K) = p - (K-1)$.
And $B(K) = 0$ also implies $\frac{A(K)}{B(K) \vee 1} = \frac{A(K)}{1}$. Hence by above inequality, $A(K) \leq q$.
Since $q < 1$, this implies $A(K) = 0$.
But that is impossible, because $A(K) = p - (K-1)$.
Even vor $K = p$, we would have $A(K) = p - (p-1) = 1$.
Contradiction!
Hence, the assumption $B(K) = 0$ is wrong.
Therefore (since $B(\cdot)$ is nonnegative by definition), $B(K) > 0$.

Hence we can write

$$\frac{A(K)}{B(K)} \leq q.$$

Rearranged gives (note that $B(K) > 0$)

$$A(K) \leq q \cdot B(K). \tag{19}$$

Therefore,

$$\begin{aligned}
&\frac{1 + A(K)}{B(K) + 1/q} \\
&\qquad\qquad \text{Multiplied by } q \text{ in numerator and denominator} \\
&\qquad\qquad \text{(possible because } q > 0) \\
&= q \cdot \frac{1 + A(K)}{q \cdot B(K) + 1} \\
&\qquad\qquad \text{Note that all terms are nonnegative. So, we can use (19).} \\
&\leq q \cdot \frac{1 + A(K)}{A(K) + 1} \\
&= q \cdot 1 \\
&= q.
\end{aligned}$$

Define

$$V^+(K) := |\left\{j : \ j \in \{K, \ldots, p\}, \ W_{[j]} > 0, \ [j] \in \mathcal{H}_0\right\}|,$$

$$V^-(K) := |\left\{j : \ j \in \{K, \ldots, p\}, \ W_{[j]} < 0, \ [j] \in \mathcal{H}_0\right\}|.$$

Part1 wts $\dfrac{|\left\{j : \ j \in \hat{S}, \ [j] \in \mathcal{H}_0\right\}|}{|\hat{S}| + 1/q} \leq q \cdot \dfrac{V^+(K)}{1 + V^-(K)}$:

$$\frac{|\left\{j:\ j\in\hat{S},\ [j]\in\mathcal{H}_0\right\}|}{|\hat{S}|+1/q}$$

We know that $\hat{S}=\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}>0\right\}$.

$$=\frac{|\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}>0,\ [j]\in\mathcal{H}_0\right\}|}{|\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}>0\right\}|+1/q}$$

$$=\frac{|\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}>0,\ [j]\in\mathcal{H}_0\right\}|}{|\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}>0\right\}|+1/q}\cdot 1$$

The first term is nonnegative.
So, by Prep1.1

$$\leq\frac{|\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}>0,\ [j]\in\mathcal{H}_0\right\}|}{|\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}>0\right\}|+1/q}\cdot\frac{1+|\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}<0\right\}|}{1+|\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}<0,\ [j]\in\mathcal{H}_0\right\}|}$$

$$=\frac{1+|\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}<0\right\}|}{|\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}>0\right\}|+1/q}\cdot\frac{|\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}>0,\ [j]\in\mathcal{H}_0\right\}|}{1+|\left\{j:\ j\in\{K,\dots,p\},\ W_{[j]}<0,\ [j]\in\mathcal{H}_0\right\}|}$$

$$=\frac{1+A(K)}{B(K)+1/q}\cdot\frac{V^+(K)}{1+V^-(K)}$$

The second term is nonnegative.
So, by Prep1.2

$$\leq q\cdot\frac{V^+(K)}{1+V^-(K)}.$$

We now introduce martingale results.

Let $\mathfrak{F}=\{\mathcal{F}_k:\ k\in\{1,\dots,p\}\}$ be a filtration with sub-σ-algebras

$$\mathcal{F}_k=\sigma\left(A(l),B(l),V^+(l),V^-(l)\ \forall l\in\{1,\dots,k\};\ \mathbb{1}_{\{[j]\in\mathcal{H}_0\}}\ \forall j\in\{1,\dots,p\}\right).$$

Note that $\mathfrak{F}$ is a filtration, because $\mathcal{F}_1\subseteq\ldots\subseteq\mathcal{F}_p$.

Part2 wts K is a bounded stopping time w.r.t. $\mathfrak{F}$:

Note that $K\in\{1,\dots,p\}$. Therefore, K is bounded.
It remains to show that K is a stopping time.
This means we have to show $\{K=t\}\in\mathcal{F}_t$ for all $t\in\{1,\dots,p\}$.

Remember that

$$K=\min\left\{k:\ k\in\{1,\dots,p\},\ \frac{A(k)}{B(k)\vee 1}\leq q\right\}.$$

We see that K is the minimal $k\in\{1,\dots,p\}$, such that a specific inequality is satisfied.

First, observe that $\{K=1\}\in\mathcal{F}_1$:

Note that by definition of $\mathcal{F}_1$, we know $A(1)$, $B(1)$.
Hence, we can tell if the inequality in the definition of K is satisfied.
In other words, we can tell if $K=1$.
Therefore, $\{K=1\}\in\mathcal{F}_1$.

Second, observe that $\{K=1\},\{K=2\}\in\mathcal{F}_2$:

Note that by definition of $\mathcal{F}_2$, we know $A(l)$, $B(l)$ for $l \in \{1, \ldots, 2\}$.
So, we know $A(1)$, $B(1)$, $A(2)$ and $B(2)$.
Knowing $A(1)$ and $B(1)$ tells us if the inequality in the definition of K is satisfied for $k = 1$.
Therefore, $\{K = 1\} \in \mathcal{F}_2$.
Knowing $A(2)$ and $B(2)$ tells us if the inequality in the definition of K is satisfied for $k = 2$.
Therefore, $\{K = 2\} \in \mathcal{F}_2$.

By the same arguments for $3, \ldots, p$, we get overall: $\{K = t\} \in \mathcal{F}_t$ for all $t \in \{1, \ldots, p\}$.

Define $M(k) = \dfrac{V^+(k)}{1 + V^-(k)}$ for $k \in \{1, \ldots, p\}$.

Part3 wts $M(k)$ is a martingale w.r.t. $\mathfrak{F}$:

Part3.1 wts $M(k)$ is adapted to $\mathfrak{F}$:

Let $k \in \{1, \ldots, p\}$ be arbitrary.
We have to show that $M(k)$ is measurable w.r.t. $\mathcal{F}_k$.
By definition of $\mathcal{F}_k$, $V^+(k)$ and $V^-(k)$ are $\mathcal{F}_k$-measurable.
Note that constant functions are measureable.
We know that the sum of two measurable functions is measurable.
Hence, $1 + V^-(k)$ is $\mathcal{F}_k$-measurable.
And we know that the quotient of two measurable functions is measurable.
Therefore, $M(k)$ is $\mathcal{F}_k$-measurable.

Part3.2 wts $\mathbb{E}(|M(k)|) < \infty$ for all $k \in \{1, \ldots, p\}$:

Let $k \in \{1, \ldots, p\}$.

$$
\begin{aligned}
&\mathbb{E}(|M(k)|) && \\
&&& \text{By definition of } M(k) \\
&= \mathbb{E}\left(\left|\frac{V^+(k)}{1 + V^-(k)}\right|\right) && \\
&&& \text{We can omit the absolute value, because} \\
&&& \text{the term inside is always nonnegative} \\
&= \mathbb{E}\left(\frac{V^+(k)}{1 + V^-(k)}\right) && \\
&&& \text{We know that } j \in \{1, \ldots, p\}. \text{ Therefore,} \\
&&& V^+(k) \leq p \text{ and } V^-(k) \geq 0 \text{ for all } k. \\
&&& \text{So, } \frac{V^+(k)}{1 + V^-(k)} \leq \frac{p}{1 + 0} = p. \\
&\leq \mathbb{E}(p) && \\
&= p && \\
&< \infty. &&
\end{aligned}
$$

Part3.3 wts $\mathbb{E}[M(k+1) \mid \mathcal{F}_k] = M(k)$ for all $k \in \{1, \ldots, p-1\}$:

Let $k \in \{1, \ldots, p-1\}$.

Prep3.3.1 wts $M(k+1) = M(k)$ on the event $\{[k] \notin \mathcal{H}_0\}$:

$$
\begin{aligned}
& M(k+1) \\
& \qquad \text{By definition of } M(k+1) \\
&= \frac{V^+(k+1)}{1+V^-(k+1)} \\
& \qquad \text{By definition of } V^+(k+1) \text{ and } V^-(k+1) \\
&= \frac{|\left\{j:\ j \in \{k+1, \ldots, p\},\ W_{[j]} > 0,\ [j] \in \mathcal{H}_0\right\}|}{1+|\left\{j:\ j \in \{k+1, \ldots, p\},\ W_{[j]} < 0,\ [j] \in \mathcal{H}_0\right\}|} \\
& \qquad \text{The counts in numerator and denominator} \\
& \qquad \text{do not change if we start with } k \text{ instead of } k+1 \\
& \qquad \text{since we know that } [k] \notin \mathcal{H}_0. \\
&= \frac{|\left\{j:\ j \in \{k, \ldots, p\},\ W_{[j]} > 0,\ [j] \in \mathcal{H}_0\right\}|}{1+|\left\{j:\ j \in \{k, \ldots, p\},\ W_{[j]} < 0,\ [j] \in \mathcal{H}_0\right\}|} \\
& \qquad \text{By definition of } V^+(k) \text{ and } V^-(k) \\
&= \frac{V^+(k)}{1+V^-(k)} \\
& \qquad \text{By definition of } M(k) \\
&= M(k).
\end{aligned}
$$

Prep3.3.2 wts $M(k+1) = \dfrac{V^+(k)-1}{1+V^-(k)}$ on the event $\left\{[k] \in \mathcal{H}_0,\ W_{[k]} > 0\right\}$:

$$
\begin{aligned}
& M(k+1) \\
& \qquad \text{By definition of } M(k+1) \\
&= \frac{V^+(k+1)}{1+V^-(k+1)} \\
& \qquad \text{By definition of } V^+(k+1) \text{ and } V^-(k+1) \\
&= \frac{|\left\{j:\ j \in \{k+1, \ldots, p\},\ W_{[j]} > 0,\ [j] \in \mathcal{H}_0\right\}|}{1+|\left\{j:\ j \in \{k+1, \ldots, p\},\ W_{[j]} < 0,\ [j] \in \mathcal{H}_0\right\}|} \\
& \qquad \text{Since we consider the event } \left\{[k] \in \mathcal{H}_0,\ W_{[k]} > 0\right\}, \\
& \qquad \text{the numerator will increase by 1 if we start} \\
& \qquad \text{with } k \text{ instead of } k+1. \text{ The denominator is not affected.} \\
&= \frac{|\left\{j:\ j \in \{k, \ldots, p\},\ W_{[j]} > 0,\ [j] \in \mathcal{H}_0\right\}| - 1}{1+|\left\{j:\ j \in \{k, \ldots, p\},\ W_{[j]} < 0,\ [j] \in \mathcal{H}_0\right\}|} \\
& \qquad \text{By definition of } V^+(k) \text{ and } V^-(k) \\
&= \frac{V^+(k)-1}{1+V^-(k)}.
\end{aligned}
$$

Prep3.3.3 wts $M(k+1) = \dfrac{V^+(k)}{V^-(k)}$ on the event $\left\{[k] \in \mathcal{H}_0,\ W_{[k]} < 0\right\}$:

$$
\begin{aligned}
& M(k+1) \\
& \qquad \text{By definition of } M(k+1) \\
&= \frac{V^+(k+1)}{1+V^-(k+1)}
\end{aligned}
$$

By definition of $V^+(k+1)$ and $V^-(k+1)$

$$= \frac{|\{j : j \in \{k+1,\dots,p\},\ W_{[j]} > 0,\ [j] \in \mathcal{H}_0\}|}{1 + |\{j : j \in \{k+1,\dots,p\},\ W_{[j]} < 0,\ [j] \in \mathcal{H}_0\}|}$$

Since we consider the event $\{[k] \in \mathcal{H}_0,\ W_{[k]} < 0\}$, the denominator will increase by 1 if we start with k instead of $k+1$. The numerator is not affected.

$$= \frac{|\{j : j \in \{k,\dots,p\},\ W_{[j]} > 0,\ [j] \in \mathcal{H}_0\}|}{1 + |\{j : j \in \{k,\dots,p\},\ W_{[j]} < 0,\ [j] \in \mathcal{H}_0\}| - 1}$$

By definition of $V^+(k)$ and $V^-(k)$

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

Note that $V^-(k) > 0$ in this case, because we consider the event $\{[k] \in \mathcal{H}_0,\ W_{[k]} < 0\}$ and hence, $V^-(k) \geq 1$.

Prep3.3.4 wts $\mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] = \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \frac{V^+(k)}{V^+(k)+V^-(k)}$:

Note that we can write $V^+(k) = \sum_{j=k}^{p} \mathbb{1}_{\{[j]\in\mathcal{H}_0\}} \cdot \mathbb{1}_{\{W_{[j]}>0\}}$.
Conditional expected value applied on both sides gives

$$\mathbb{E}\left[V^+(k) \mid \mathcal{F}_k\right] = \mathbb{E}\left[\sum_{j=k}^{p} \mathbb{1}_{\{[j]\in\mathcal{H}_0\}} \cdot \mathbb{1}_{\{W_{[j]}>0\}} \mid \mathcal{F}_k\right].$$

Since $V^+(k), \{[j] \in \mathcal{H}_0\} \in \mathcal{F}_k$ for all $j \in \{1,\dots,p\}$, we can pull these terms out and get

$$V^+(k) \cdot \mathbb{E}\left[1 \mid \mathcal{F}_k\right] = \sum_{j=k}^{p} \mathbb{1}_{\{[j]\in\mathcal{H}_0\}} \cdot \mathbb{E}\left[\mathbb{1}_{\{W_{[j]}>0\}} \mid \mathcal{F}_k\right].$$

Hence,

$$V^+(k) = \sum_{j=k}^{p} \mathbb{1}_{\{[j]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[j]} > 0 \mid \mathcal{F}_k\right].$$

By Lemma 3.13 i) we know for $[j] \in \mathcal{H}_0$: $\operatorname{sign}(W_{[j]})$ are independent and $\mathbb{P}\left(\operatorname{sign}(W_{[j]}) = -1\right) = \mathbb{P}\left(\operatorname{sign}(W_{[j]}) = +1\right) = \frac{1}{2}$.
The latter can be written as $\mathbb{P}\left(W_{[j]} < 0\right) = \mathbb{P}\left(W_{[j]} > 0\right) = \frac{1}{2}$.
In other words, the signs of the null MX-Knockoff feature statistics are i.i.d. and hence exchangeable.
By definition of $\mathcal{F}_k$, all nulls are known.
Therefore,

$$\mathbb{1}_{\{[j]\in\mathcal{H}_0\}} \cdot \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] = \mathbb{1}_{\{[j]\in\mathcal{H}_0\}} \cdot \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[j]} > 0 \mid \mathcal{F}_k\right]$$

for all $j \in \{k,\dots,p\}$.

And note that

$\mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot V^+(k)$

We know from above that $V^+(k) = \sum_{j=k}^{p} \mathbb{1}_{\{[j]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[j]} > 0 \mid \mathcal{F}_k\right]$.

$$= \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \sum_{j=k}^{p} \mathbb{1}_{\{[j]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[j]} > 0 \mid \mathcal{F}_k\right]$$

$$= \sum_{j=k}^{p} \mathbb{1}_{\{[j]\in\mathcal{H}_0\}} \cdot \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[j]} > 0 \mid \mathcal{F}_k\right]$$

We know that $\mathbb{1}_{\{[j]\in\mathcal{H}_0\}} \cdot \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] = \mathbb{1}_{\{[j]\in\mathcal{H}_0\}} \cdot \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[j]} > 0 \mid \mathcal{F}_k\right]$.

$$= \sum_{j=k}^{p} \mathbb{1}_{\{[j]\in\mathcal{H}_0\}} \cdot \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right]$$

$$= \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] \cdot \sum_{j=k}^{p} \mathbb{1}_{\{[j]\in\mathcal{H}_0\}}$$

$$= \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] \cdot \left|\left\{j :\ j \in \{k,\dots,p\},\ [j] \in \mathcal{H}_0\right\}\right|$$

Note that $V^+(k) + V^-(k) = \left|\left\{j :\ j \in \{k,\dots,p\},\ [j] \in \mathcal{H}_0\right\}\right|$.

$$= \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] \cdot \left(V^+(k) + V^-(k)\right).$$

Rearranged gives

$$\mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] = \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \frac{V^+(k)}{V^+(k) + V^-(k)}.$$

Therefore,

$$\mathbb{E}\left[M(k+1) \mid \mathcal{F}_k\right]$$

$$= \mathbb{E}\left[M(k+1) \cdot \mathbb{1}_{\{[k]\notin\mathcal{H}_0\}\cup\{[k]\in\mathcal{H}_0\}} \mid \mathcal{F}_k\right]$$

Note that $\mathbb{1}_{C\cup D} = \mathbb{1}_C + \mathbb{1}_D$ for disjoint sets C, D.

$$= \mathbb{E}\left[M(k+1) \cdot \left(\mathbb{1}_{\{[k]\notin\mathcal{H}_0\}} + \mathbb{1}_{\{[k]\in\mathcal{H}_0\}}\right) \mid \mathcal{F}_k\right]$$

$$= \mathbb{E}\left[M(k+1) \cdot \left(\mathbb{1}_{\{[k]\notin\mathcal{H}_0\}} + \mathbb{1}_{\left\{[k]\in\mathcal{H}_0,\, W_{[k]}>0\right\}\cup\left\{[k]\in\mathcal{H}_0,\, W_{[k]}<0\right\}}\right) \mid \mathcal{F}_k\right]$$

Note that $\mathbb{1}_{C\cup D} = \mathbb{1}_C + \mathbb{1}_D$ for disjoint sets C, D.

$$= \mathbb{E}\left[M(k+1) \cdot \left(\mathbb{1}_{\{[k]\notin\mathcal{H}_0\}} + \mathbb{1}_{\left\{[k]\in\mathcal{H}_0,\, W_{[k]}>0\right\}} + \mathbb{1}_{\left\{[k]\in\mathcal{H}_0,\, W_{[k]}<0\right\}}\right) \mid \mathcal{F}_k\right]$$

$$= \mathbb{E}\left[M(k+1) \cdot \mathbb{1}_{\{[k]\notin\mathcal{H}_0\}} + M(k+1) \cdot \mathbb{1}_{\left\{[k]\in\mathcal{H}_0,\, W_{[k]}>0\right\}} + M(k+1) \cdot \mathbb{1}_{\left\{[k]\in\mathcal{H}_0,\, W_{[k]}<0\right\}} \mid \mathcal{F}_k\right]$$

$$= \mathbb{E}\left[M(k+1) \cdot \mathbb{1}_{\{[k]\notin\mathcal{H}_0\}} \mid \mathcal{F}_k\right]$$
$$+ \mathbb{E}\left[M(k+1) \cdot \mathbb{1}_{\left\{[k]\in\mathcal{H}_0,\, W_{[k]}>0\right\}} \mid \mathcal{F}_k\right]$$
$$+ \mathbb{E}\left[M(k+1) \cdot \mathbb{1}_{\left\{[k]\in\mathcal{H}_0,\, W_{[k]}<0\right\}} \mid \mathcal{F}_k\right]$$

By Prep3.3.1, Prep3.3.2 and Prep3.3.3

$$= \mathbb{E}\left[M(k) \cdot \mathbb{1}_{\{[k]\notin\mathcal{H}_0\}} \mid \mathcal{F}_k\right]$$
$$+ \mathbb{E}\left[\frac{V^+(k)-1}{1+V^-(k)} \cdot \mathbb{1}_{\left\{[k]\in\mathcal{H}_0,\, W_{[k]}>0\right\}} \mid \mathcal{F}_k\right]$$
$$+ \mathbb{E}\left[\frac{V^+(k)}{V^-(k)} \cdot \mathbb{1}_{\left\{[k]\in\mathcal{H}_0,\, W_{[k]}<0\right\}} \mid \mathcal{F}_k\right]$$

Note that $\mathbb{1}_{C\cap D} = \mathbb{1}_C \cdot \mathbb{1}_D$ for sets C, D.

$$= \mathbb{E}\left[M(k) \cdot \mathbb{1}_{\{[k]\notin\mathcal{H}_0\}} \mid \mathcal{F}_k\right]$$
$$+ \mathbb{E}\left[\frac{V^+(k)-1}{1+V^-(k)} \cdot \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{1}_{\left\{W_{[k]}>0\right\}} \mid \mathcal{F}_k\right]$$
$$+ \mathbb{E}\left[\frac{V^+(k)}{V^-(k)} \cdot \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{1}_{\left\{W_{[k]}<0\right\}} \mid \mathcal{F}_k\right]$$

By definition of $\mathcal{F}_k$, we know that $V^+(k), V^-(k), \{[k] \in \mathcal{H}_0\} \in \mathcal{F}_k$.
Since complements of elements of a σ-algebra are
again in that σ-algebra, $\{[k] \in \mathcal{H}_0\}^c = \{[k] \notin \mathcal{H}_0\} \in \mathcal{F}_k$.
Hence we can pull out these terms.

$$
\begin{aligned}
&= M(k) \cdot \mathbb{1}_{\{[k] \notin \mathcal{H}_0\}} \cdot \mathbb{E}\left[1 \mid \mathcal{F}_k\right] \\
&\quad + \frac{V^+(k)-1}{1+V^-(k)} \cdot \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \mathbb{E}\left[\mathbb{1}_{\{W_{[k]}>0\}} \mid \mathcal{F}_k\right] \\
&\quad + \frac{V^+(k)}{V^-(k)} \cdot \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \mathbb{E}\left[\mathbb{1}_{\{W_{[k]}<0\}} \mid \mathcal{F}_k\right] \\
&= M(k) \cdot \mathbb{1}_{\{[k] \notin \mathcal{H}_0\}} \\
&\quad + \frac{V^+(k)-1}{1+V^-(k)} \cdot \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] \\
&\quad + \frac{V^+(k)}{V^-(k)} \cdot \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[k]} < 0 \mid \mathcal{F}_k\right]
\end{aligned}
$$

By Prep3.3.4

$$
\begin{aligned}
&= M(k) \cdot \mathbb{1}_{\{[k] \notin \mathcal{H}_0\}} \\
&\quad + \frac{V^+(k)-1}{1+V^-(k)} \cdot \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \frac{V^+(k)}{V^+(k)+V^-(k)} \\
&\quad + \frac{V^+(k)}{V^-(k)} \cdot \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \left(1 - \frac{V^+(k)}{V^+(k)+V^-(k)}\right) \\
&= M(k) \cdot \mathbb{1}_{\{[k] \notin \mathcal{H}_0\}} \\
&\quad + \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \frac{V^+(k)-1}{1+V^-(k)} \cdot \frac{V^+(k)}{V^+(k)+V^-(k)} \\
&\quad + \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \frac{V^+(k)}{V^-(k)} \cdot \left(\frac{V^+(k)+V^-(k)}{V^+(k)+V^-(k)} - \frac{V^+(k)}{V^+(k)+V^-(k)}\right) \\
&= M(k) \cdot \mathbb{1}_{\{[k] \notin \mathcal{H}_0\}} \\
&\quad + \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \frac{V^+(k)-1}{1+V^-(k)} \cdot \frac{V^+(k)}{V^+(k)+V^-(k)} \\
&\quad + \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \frac{V^+(k)}{V^-(k)} \cdot \frac{V^-(k)}{V^+(k)+V^-(k)} \\
&= M(k) \cdot \mathbb{1}_{\{[k] \notin \mathcal{H}_0\}} + \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \left(\frac{V^+(k)-1}{1+V^-(k)} \cdot \frac{V^+(k)}{V^+(k)+V^-(k)} + \frac{V^-(k)}{V^-(k)} \cdot \frac{V^+(k)}{V^+(k)+V^-(k)}\right) \\
&= M(k) \cdot \mathbb{1}_{\{[k] \notin \mathcal{H}_0\}} + \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \left(\frac{V^+(k)}{V^+(k)+V^-(k)} \cdot \left(\frac{V^+(k)-1}{1+V^-(k)} + 1\right)\right) \\
&= M(k) \cdot \mathbb{1}_{\{[k] \notin \mathcal{H}_0\}} + \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \left(\frac{V^+(k)}{V^+(k)+V^-(k)} \cdot \left(\frac{V^+(k)-1}{1+V^-(k)} + \frac{1+V^-(k)}{1+V^-(k)}\right)\right) \\
&= M(k) \cdot \mathbb{1}_{\{[k] \notin \mathcal{H}_0\}} + \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \left(\frac{V^+(k)}{V^+(k)+V^-(k)} \cdot \frac{V^+(k)+V^-(k)}{1+V^-(k)}\right) \\
&= M(k) \cdot \mathbb{1}_{\{[k] \notin \mathcal{H}_0\}} + \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \frac{V^+(k)}{1+V^-(k)}
\end{aligned}
$$

By definition of $M(k)$

$$
\begin{aligned}
&= M(k) \cdot \mathbb{1}_{\{[k] \notin \mathcal{H}_0\}} + \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot M(k) \\
&= M(k) \cdot \left(\mathbb{1}_{\{[k] \notin \mathcal{H}_0\}} + \mathbb{1}_{\{[k] \in \mathcal{H}_0\}}\right)
\end{aligned}
$$

Note that $\mathbb{1}_{C \cup D} = \mathbb{1}_C + \mathbb{1}_D$ for disjoint sets C, D.

$$
\begin{aligned}
&= M(k) \cdot \mathbb{1}_{\{[k] \notin \mathcal{H}_0\} \cup \{[k] \in \mathcal{H}_0\}} \\
&= M(k).
\end{aligned}
$$

Since $k \in \{1, \ldots, p-1\}$ was arbitrary, we have shown $\mathbb{E}\left[M(k+1) \mid \mathcal{F}_k\right] = M(k)$ for all $k \in \{1, \ldots, p-1\}$.

Define $N_0 = |\mathcal{H}_0|$ as the total number of nulls.

Prep4.1 wts $V^+(1) \sim \text{Binomial}\left(N_0, \frac{1}{2}\right)$:

Note that we can write

$$V^+(1) = |\left\{j:\ j \in \{1,\dots,p\},\ W_{[j]} > 0,\ [j] \in \mathcal{H}_0\right\}| = |\left\{j:\ j \in \{1,\dots,p\},\ \text{sign}(W_{[j]}) = +1,\ [j] \in \mathcal{H}_0\right\}|.$$

Define $U_j = \mathbb{1}_{\left\{\text{sign}(W_{[j]})=+1\right\}}$.
Note that $\mathbb{P}\left(U_j = 1\right) = \mathbb{P}\left(\text{sign}(W_{[j]}) = +1\right) = \frac{1}{2}$ for all j where $[j] \in \mathcal{H}_0$ by Lemma 3.13 i).
By Lemma 3.13 i), $\text{sign}(W_{[j]})$ are independent for j where $[j] \in \mathcal{H}_0$.
This means that U_j are independent random variables with $U_j \sim \text{Bernoulli}\left(\frac{1}{2}\right)$ for all j where $[j] \in \mathcal{H}_0$.
Short: $U_j \overset{i.i.d.}{\sim} \text{Bernoulli}\left(\frac{1}{2}\right)$ for j where $[j] \in \mathcal{H}_0$.
Note that $\sum_{\substack{j \in \{1,\dots,p\},\\ [j] \in \mathcal{H}_0}} U_j = V^+(1)$.
By definition of the Binomial-distribution, $V^+(1) \sim \text{Binomial}\left(N_0, \frac{1}{2}\right)$.

Prep4.2 wts $V^-(1) = N_0 - V^+(1)$:

$$\begin{aligned}
&V^-(1)\\
&= |\left\{j:\ j \in \{1,\dots,p\},\ W_{[j]} < 0,\ [j] \in \mathcal{H}_0\right\}| \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{Note that } N_0 = |\mathcal{H}_0|.\\
&= N_0 - |\left\{j:\ j \in \{1,\dots,p\},\ W_{[j]} > 0,\ [j] \in \mathcal{H}_0\right\}| \\
&= N_0 - V^+(1).
\end{aligned}$$

Part4 wts $\mathbb{E}\left[\frac{V^+(K)}{1+V^-(K)}\right] \leq 1$:

$$\begin{aligned}
&\mathbb{E}\left[\frac{V^+(K)}{1+V^-(K)}\right]\\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{By Part2, we know that } K \text{ is a bounded stopping time.}\\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{By Part3, we know that } \frac{V^+(k)}{1+V^-(k)} \text{ is a martingale.}\\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{Therefore, we can use Lemma A.4 i)}\\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{(optional stopping theorem for martingales).}\\
&= \mathbb{E}\left[\frac{V^+(1)}{1+V^-(1)}\right]\\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{By Prep4.2}
\end{aligned}$$

$$= \mathbb{E}\left[\frac{V^+(1)}{1+N_0-V^+(1)}\right]$$

In general, it holds for a discrete random variable Y and function g that $\mathbb{E}(g(Y)) = \sum_y g(y)\cdot\mathbb{P}(Y=y)$. In our case, $g(x) := \frac{x}{1+N_0-x}$ and $Y := V^+(1)$. Note that $N_0 \le p$ and $Y \in \{0,\dots,N_0\}$.

$$= \sum_{i=0}^{N_0} g(i)\cdot\mathbb{P}(Y=i)$$

By definition of g

$$= \sum_{i=0}^{N_0} \frac{i}{1+N_0-i}\cdot\mathbb{P}(Y=i)$$

By Prep4.1, $V^+(1) \sim \text{Binomial}\left(N_0, \frac{1}{2}\right)$.

$$= \sum_{i=0}^{N_0} \frac{i}{1+N_0-i}\cdot\binom{N_0}{i}\cdot\left(\frac{1}{2}\right)^i\cdot\left(1-\frac{1}{2}\right)^{N_0-i}$$

We can start the sum at $i=1$.

$$= \sum_{i=1}^{N_0} \frac{i}{1+N_0-i}\cdot\binom{N_0}{i}\cdot\left(\frac{1}{2}\right)^i\cdot\left(1-\frac{1}{2}\right)^{N_0-i}$$
$$= \sum_{i=1}^{N_0} \frac{i}{1+N_0-i}\cdot\frac{N_0!}{i!\cdot(N_0-i)!}\cdot\left(\frac{1}{2}\right)^i\cdot\left(\frac{1}{2}\right)^{N_0-i}$$
$$= \sum_{i=1}^{N_0} \frac{1}{1+N_0-i}\cdot\frac{N_0!}{(i-1)!\cdot(N_0-i)!}\cdot\left(\frac{1}{2}\right)^{i-1}\cdot\left(\frac{1}{2}\right)^1\cdot\left(\frac{1}{2}\right)^{N_0-(i-1)}\cdot\left(\frac{1}{2}\right)^{-1}$$

Note that $1+N_0-i = N_0-(i-1)$.

$$= \sum_{i=1}^{N_0} \frac{N_0!}{(i-1)!\cdot(N_0-(i-1))!}\cdot\left(\frac{1}{2}\right)^{i-1}\cdot\left(\frac{1}{2}\right)^{N_0-(i-1)}$$
$$= \sum_{i=1}^{N_0} \binom{N_0}{i-1}\cdot\left(\frac{1}{2}\right)^{i-1}\cdot\left(\frac{1}{2}\right)^{N_0-(i-1)}$$

Substitution: set $j = i-1$.

$$= \sum_{j=0}^{N_0-1} \underbrace{\binom{N_0}{j}\cdot\left(\frac{1}{2}\right)^j\cdot\left(\frac{1}{2}\right)^{N_0-j}}_{=\mathbb{P}(Y=j)}$$

The sum of probabilities the random variable Y can achieve has to be smaller than 1.

$$\le 1.$$

Part5 wts mFDR $\le q$:

mFDR

$$= \mathbb{E}\left[\frac{|\left\{j:\ j\in\hat{S},\ [j]\in\mathcal{H}_0\right\}|}{|\hat{S}|+1/q}\right]$$

By Part1, we know that $\frac{|\left\{j:\ j\in\hat{S},\ [j]\in\mathcal{H}_0\right\}|}{|\hat{S}|+1/q} \leq q\cdot\frac{V^+(K)}{1+V^-(K)}$.

$$\leq \mathbb{E}\left[q\cdot\frac{V^+(K)}{1+V^-(K)}\right]$$
$$= q\cdot\mathbb{E}\left[\frac{V^+(K)}{1+V^-(K)}\right]$$

By Part4, we know that $\mathbb{E}\left[\frac{V^+(K)}{1+V^-(K)}\right]\leq 1$.

$$\leq q\cdot 1$$
$$= q.$$

As $q\in(0,1)$ was arbitrary but fixed, we have shown the result for all $q\in(0,1)$.

And note that this result holds even conditionally on $\mathbf{Y}$, because we can use Lemma 3.13 ii) (instead of just using Lemma 3.13 i)) in Prep4.1.

□

Theorem 3.15. *(Part of Theorem 3.4 in Candes et al. (2018)). For any $q\in[0,1]$, the MX-Knockoff+ method (see Definition 3.10) which is selecting variables with indices*

$$\hat{S}=\{j:\ W_j\geq T_+\}$$

satisfies

$$\text{FDR}\leq q.$$

The result is non-asymptotic and holds no matter the dependence between the response and the covariates - in fact, it holds even conditionally on the response $\mathbf{Y}$.

Proof:

Note that this proof is the same as the proof of Theorem 2.13, except that we use properties of MX-Knockoff feature statistics instead of FX-Knockoff feature statistics and a different definition of nulls ($\mathcal{H}_0$ instead of $\{j:\ \beta_j=0\}$).

We only prove the case where:
- $W_j\neq 0$ for all j (hence, the assumptions of Lemma 3.13 are satisfied).
- $|W_j|$ are distinct, i.e., $|W_i|\neq|W_j|$ for $i\neq j$.
- $T_+<\infty$.

Let $q\in[0,1]$ be arbitrary but fixed.

We would like to assume that $|W_j|$ are sorted in increasing order to make the proof easier. Such sorting implies that null indices change too.
To address this problem, we define random indices $[1],\ldots,[p]$ through $|W_{[1]}|<|W_{[2]}|<\ldots<|W_{[p]}|$.

The concept is similar to order statistics, except that we use it for ordered absolute values of terms (e.g. for $W_1 = 3$, $W_2 = -1$ and $W_3 = -4$, we have $|W_{[1]}| < |W_{[2]}| < |W_{[3]}|$ where $W_{[1]} = W_2 = -1$, $W_{[2]} = W_1 = 3$ and $W_{[3]} = W_3 = -4$).

For each $k \in \{1, \ldots, p\}$, define

$$A(k) := |\left\{j : \ j \in \{k, \ldots, p\}\,, \ W_{[j]} < 0\right\}|,$$

$$B(k) := |\left\{j : \ j \in \{k, \ldots, p\}\,, \ W_{[j]} > 0\right\}|.$$

In other words, $A(k)$ is the number of negative W_j among the $p - (k-1)$ biggest $|W_j|$.
And $B(k)$ is the number of positive W_j among the $p - (k-1)$ biggest $|W_j|$.

Hence, we can write

$$A(k) = |\left\{j : \ |W_j| \geq |W_{[k]}|, \ W_j < 0\right\}|,$$

$$B(k) = |\left\{j : \ |W_j| \geq |W_{[k]}|, \ W_j > 0\right\}|.$$

In Definition 3.10, we have $\mathcal{W} = \{|W_j| : \ W_j \neq 0, \ j \in \{1, \ldots, p\}\}$.
Since we assume $W_j \neq 0$ in this proof, we can write that as $\mathcal{W} = \left\{|W_{[j]}| : \ j \in \{1, \ldots, p\}\right\}$.

Prep0.1 wts $\left\{j : \ W_j \leq -|W_{[k]}|\right\} = \left\{j : \ |W_j| \geq |W_{[k]}|, \ W_j < 0\right\}$ for $k \in \{1, \ldots, p\}$:

> Let $k \in \{1, \ldots, p\}$ be arbitrary.
> wts $\left\{j : \ W_j \leq -|W_{[k]}|\right\} \subseteq \left\{j : \ |W_j| \geq |W_{[k]}|, \ W_j < 0\right\}$:
>
> > Let $j \in \left\{j : \ W_j \leq -|W_{[k]}|\right\}$ be arbitrary.
> > This means that $W_j \leq -|W_{[k]}|$.
> > Since $W_j \neq 0$ for all j, this implies that $W_j < 0$.
> > Multiplying (-1) on both sides of $W_j \leq -|W_{[k]}|$ gives $-W_j \geq |W_{[k]}|$.
> > And since $W_j < 0$, we have $-W_j > 0$.
> > Therefore, $|-W_j| = -W_j$.
> > And of course (property of absolute value), $|-W_j| = |W_j|$.
> > Hence, $|W_j| \geq |W_{[k]}|$.
> > So, we have $|W_j| \geq |W_{[k]}|$ and $W_j < 0$.
> > Hence, $j \in \left\{j : \ |W_j| \geq |W_{[k]}|, \ W_j < 0\right\}$.
>
> wts $\left\{j : \ W_j \leq -|W_{[k]}|\right\} \supseteq \left\{j : \ |W_j| \geq |W_{[k]}|, \ W_j < 0\right\}$:
>
> > Let $j \in \left\{j : \ |W_j| \geq |W_{[k]}|, \ W_j < 0\right\}$ be arbitrary.
> > This means that $|W_j| \geq |W_{[k]}|$ and $W_j < 0$.
> > Multiplying (-1) on both sides of $|W_j| \geq |W_{[k]}|$ gives $-|W_j| \leq -|W_{[k]}|$.
> > And since $W_j < 0$, we have $|W_j| = -W_j$ (by definition of the absolute value).
> > Hence, $-|W_j| = -(-W_j) = W_j$.
> > Therefore, $W_j \leq -|W_{[k]}|$.
> > Hence, $j \in \left\{j : \ W_j \leq -|W_{[k]}|\right\}$.
>
> Both together give $\left\{j : \ W_j \leq -|W_{[k]}|\right\} = \left\{j : \ |W_j| \geq |W_{[k]}|, \ W_j < 0\right\}$.
> And since $k \in \{1, \ldots, p\}$ was arbitrary, we have shown this result for all $k \in \{1, \ldots, p\}$.

Prep0.2 wts $\left\{j : \ W_j \geq |W_{[k]}|\right\} = \left\{j : \ |W_j| \geq |W_{[k]}|, \ W_j > 0\right\}$ for $k \in \{1, \ldots, p\}$:

> Let $k \in \{1, \ldots, p\}$ be arbitrary.
> wts $\left\{j : \ W_j \geq |W_{[k]}|\right\} \subseteq \left\{j : \ |W_j| \geq |W_{[k]}|, \ W_j > 0\right\}$:

Let $j \in \{j : W_j \geq |W_{[k]}|\}$ be arbitrary.
This means that $W_j \geq |W_{[k]}|$.
Since $W_j \neq 0$ for all j, this implies that $W_j > 0$.
Therefore, $|W_j| = W_j$.
So, we have $|W_j| \geq |W_{[k]}|$.
Hence, $j \in \{j : |W_j| \geq |W_{[k]}|,\ W_j > 0\}$.

wts $\{j : W_j \geq |W_{[k]}|\} \supseteq \{j : |W_j| \geq |W_{[k]}|,\ W_j > 0\}$:

Let $j \in \{j : |W_j| \geq |W_{[k]}|,\ W_j > 0\}$ be arbitrary.
This means that $|W_j| \geq |W_{[k]}|$ and $W_j > 0$.
Since $W_j > 0$, we have $|W_j| = W_j$.
Therefore, $W_j \geq |W_{[k]}|$.
Hence, $j \in \{j : W_j \geq |W_{[k]}|\}$.

Both together give $\{j : W_j \geq |W_{[k]}|\} = \{j : |W_j| \geq |W_{[k]}|,\ W_j > 0\}$.
And since $k \in \{1, \ldots, p\}$ was arbitrary, we have shown this result for all $k \in \{1, \ldots, p\}$.

Then

$$
\begin{aligned}
& T_+ \\
& \qquad \text{By Definition 3.10} \\
&= \min\left\{t \in \mathcal{W} : \ \frac{1 + |\{j : W_j \leq -t\}|}{|\{j : W_j \geq t\}| \vee 1} \leq q\right\} \\
& \qquad \text{We know that } \mathcal{W} = \left\{|W_{[j]}| : j \in \{1, \ldots, p\}\right\}. \\
&= \min\left\{|W_{[k]}| : k \in \{1, \ldots, p\}, \ \frac{1 + |\{j : W_j \leq -|W_{[k]}|\}|}{|\{j : W_j \geq |W_{[k]}|\}| \vee 1} \leq q\right\} \\
& \qquad \text{By Prep0.1 and Prep0.2, we know} \\
& \qquad \text{that } \{j : W_j \leq -|W_{[k]}|\} = \{j : |W_j| \geq |W_{[k]}|,\ W_j < 0\} \\
& \qquad \text{and } \{j : W_j \geq |W_{[k]}|\} = \{j : |W_j| \geq |W_{[k]}|,\ W_j > 0\}. \\
&= \min\left\{|W_{[k]}| : k \in \{1, \ldots, p\}, \ \frac{1 + |\{j : |W_j| \geq |W_{[k]}|,\ W_j < 0\}|}{|\{j : |W_j| \geq |W_{[k]}|,\ W_j > 0\}| \vee 1} \leq q\right\} \\
& \qquad \text{By definition of } A(k) \text{ and } B(k) \\
&= \min\left\{|W_{[k]}| : k \in \{1, \ldots, p\}, \ \frac{1 + A(k)}{B(k) \vee 1} \leq q\right\}.
\end{aligned}
$$

Set

$$K := \min\left\{k : k \in \{1, \ldots, p\}, \ \frac{1 + A(k)}{B(k) \vee 1} \leq q\right\}.$$

Since $T_+ < \infty$, we know that at least one $k \in \{1, \ldots, p\}$ satisfies the inequality in the definition of K.
We know from above that T_+ is equal to the smallest $|W_{[k]}|$ such that this inequality is satisfied.
And we know that $|W_{[1]}| < |W_{[2]}| < \ldots < |W_{[p]}|$.
Therefore,

$$T_+ = |W_{[K]}|.$$

Side note: If we would relax the assumption $|W_{[1]}| < |W_{[2]}| < \ldots < |W_{[p]}|$ to $|W_{[1]}| \leq |W_{[2]}| \leq \ldots \leq |W_{[p]}|$, then $A(k)$, $B(k)$ and therefore K would not be unique in general, as a small example shows:

Take $p = 3$ with $W_1 = 3$, $W_2 = -1$ and $W_3 = -3$. Then we can choose if we write $|W_2| \leq |W_1| \leq |W_3|$ or $|W_2| \leq |W_3| \leq |W_1|$.

Case1 $W_{[1]} = W_2$, $W_{[2]} = W_1$, $W_{[3]} = W_3$:

Here, we have
$A(3) = |\left\{j:\ j \in \{3\},\ W_{[j]} < 0\right\}| = |\{3\}| = 1$ and
$B(3) = |\left\{j:\ j \in \{3\},\ W_{[j]} > 0\right\}| = |\emptyset| = 0.$
Hence, in particular,

$$\frac{1 + A(3)}{B(3) \vee 1} = \frac{1+1}{0 \vee 1} = 2.$$

Case2 $W_{[1]} = W_2$, $W_{[2]} = W_3$, $W_{[3]} = W_1$:

Here, we have
$A(3) = |\left\{j:\ j \in \{3\},\ W_{[j]} < 0\right\}| = |\emptyset| = 0$ and
$B(3) = |\left\{j:\ j \in \{3\},\ W_{[j]} > 0\right\}| = |\{3\}| = 1.$
Hence, in particular,

$$\frac{1 + A(3)}{B(3) \vee 1} = \frac{1+0}{1 \vee 1} = 1.$$

To continue the proof, note that

$$\begin{aligned}
&\hat{S} \\
&= \{j:\ W_j \geq T_+\} \\
&\qquad\qquad \text{We know that } T_+ = |W_{[K]}|. \\
&= \left\{j:\ W_j \geq |W_{[K]}|\right\} \\
&\qquad\qquad \text{By the same argument as in Prep0.2} \\
&= \left\{j:\ |W_j| \geq |W_{[K]}|,\ W_j > 0\right\} \\
&\qquad\qquad \text{By the same argument as above, where} \\
&\qquad\qquad \text{we introduced a different formula for } B(k). \\
&= \left\{j:\ j \in \{K, \ldots, p\},\ W_{[j]} > 0\right\}.
\end{aligned}$$

Prep1.1 wts $\dfrac{1 + A(K)}{B(K) \vee 1} \leq q$:

We know that

$$K = \min\left\{k:\ k \in \{1, \ldots, p\},\ \frac{1 + A(k)}{B(k) \vee 1} \leq q\right\}.$$

Therefore, K is one such $k \in \{1, \ldots, p\}$ satisfying this inequality.

Define

$$V^+(K) := |\left\{j:\ j \in \{K, \ldots, p\},\ W_{[j]} > 0,\ [j] \in \mathcal{H}_0\right\}|,$$

$$V^-(K) := |\left\{j:\ j \in \{K, \ldots, p\},\ W_{[j]} < 0,\ [j] \in \mathcal{H}_0\right\}|.$$

Part1 wts FDP $\leq q \cdot \dfrac{V^+(K)}{1 + V^-(K)}$:

FDP

By Definition 3.11 i)

$$= \frac{|\left\{j:\ j \in \hat{S},\ [j] \in \mathcal{H}_0\right\}|}{|\hat{S}| \vee 1}$$

We know that $\hat{S} = \left\{j :\ j \in \{K, \dots, p\},\ W_{[j]} > 0\right\}$.

$$= \frac{|\left\{j :\ j \in \{K, \dots, p\},\ W_{[j]} > 0,\ [j] \in \mathcal{H}_0\right\}|}{|\left\{j :\ j \in \{K, \dots, p\},\ W_{[j]} > 0\right\}| \vee 1}$$

$$= \frac{|\left\{j :\ j \in \{K, \dots, p\},\ W_{[j]} > 0,\ [j] \in \mathcal{H}_0\right\}|}{|\left\{j :\ j \in \{K, \dots, p\},\ W_{[j]} > 0\right\}| \vee 1} \cdot 1$$

We multiply with $\frac{1 + A(K)}{1 + A(K)}$.

$$= \frac{|\left\{j :\ j \in \{K, \dots, p\},\ W_{[j]} > 0,\ [j] \in \mathcal{H}_0\right\}|}{|\left\{j :\ j \in \{K, \dots, p\},\ W_{[j]} > 0\right\}| \vee 1} \cdot \frac{1 + |\left\{j :\ j \in \{K, \dots, p\},\ W_{[j]} < 0\right\}|}{1 + |\left\{j :\ j \in \{K, \dots, p\},\ W_{[j]} < 0\right\}|}$$

$$= \frac{1 + |\left\{j :\ j \in \{K, \dots, p\},\ W_{[j]} < 0\right\}|}{|\left\{j :\ j \in \{K, \dots, p\},\ W_{[j]} > 0\right\}| \vee 1} \cdot \frac{|\left\{j :\ j \in \{K, \dots, p\},\ W_{[j]} > 0,\ [j] \in \mathcal{H}_0\right\}|}{1 + |\left\{j :\ j \in \{K, \dots, p\},\ W_{[j]} < 0\right\}|}$$

$$= \frac{1 + A(K)}{B(K) \vee 1} \cdot \frac{V^+(K)}{1 + |\left\{j :\ j \in \{K, \dots, p\},\ W_{[j]} < 0\right\}|}$$

All terms are nonnegative and note that $|\left\{j :\ j \in \{K, \dots, p\},\ W_{[j]} < 0\right\}| \geq |\left\{j :\ j \in \{K, \dots, p\},\ W_{[j]} < 0,\ [j] \in \mathcal{H}_0\right\}|$.

$$\leq \frac{1 + A(K)}{B(K) \vee 1} \cdot \frac{V^+(K)}{1 + |\left\{j :\ j \in \{K, \dots, p\},\ W_{[j]} < 0,\ [j] \in \mathcal{H}_0\right\}|}$$

$$= \frac{1 + A(K)}{B(K) \vee 1} \cdot \frac{V^+(K)}{1 + V^-(K)}$$

By Prep1.1

$$\leq q \cdot \frac{V^+(K)}{1 + V^-(K)}.$$

We now introduce martingale results.

Let $\mathfrak{F} = \{\mathcal{F}_k :\ k \in \{1, \dots, p\}\}$ be a filtration with sub-σ-algebras

$$\mathcal{F}_k = \sigma\left(A(l), B(l), V^+(l), V^-(l)\ \ \forall l \in \{1, \dots, k\};\ \mathbb{1}_{\{[j] \in \mathcal{H}_0\}}\ \ \forall j \in \{1, \dots, p\}\right).$$

Note that $\mathfrak{F}$ is a filtration, because $\mathcal{F}_1 \subseteq \dots \subseteq \mathcal{F}_p$.

Part2 wts K is a bounded stopping time w.r.t. $\mathfrak{F}$:

Note that $K \in \{1, \dots, p\}$. Therefore, K is bounded.
It remains to show that K is a stopping time.
This means we have to show $\{K = t\} \in \mathcal{F}_t$ for all $t \in \{1, \dots, p\}$.

Remember that

$$K = \min\left\{k :\ k \in \{1, \dots, p\},\ \frac{1 + A(k)}{B(k) \vee 1} \leq q\right\}.$$

We see that K is the minimal $k \in \{1, \dots, p\}$, such that a specific inequality is satisfied.

First, observe that $\{K = 1\} \in \mathcal{F}_1$:

Note that by definition of $\mathcal{F}_1$, we know $A(1)$, $B(1)$.
Hence, we can tell if the inequality in the definition of K is satisfied.
In other words, we can tell if $K = 1$.
Therefore, $\{K = 1\} \in \mathcal{F}_1$.

Second, observe that $\{K = 1\}, \{K = 2\} \in \mathcal{F}_2$:

Note that by definition of $\mathcal{F}_2$, we know $A(l)$, $B(l)$ for $l \in \{1, \ldots, 2\}$.
So, we know $A(1)$, $B(1)$, $A(2)$ and $B(2)$.
Knowing $A(1)$ and $B(1)$ tells us if the inequality in the definition of K is satisfied for $k = 1$.
Therefore, $\{K = 1\} \in \mathcal{F}_2$.
Knowing $A(2)$ and $B(2)$ tells us if the inequality in the definition of K is satisfied for $k = 2$.
Therefore, $\{K = 2\} \in \mathcal{F}_2$.

By the same arguments for $3, \ldots, p$, we get overall: $\{K = t\} \in \mathcal{F}_t$ for all $t \in \{1, \ldots, p\}$.

Define $M(k) = \dfrac{V^+(k)}{1 + V^-(k)}$ for $k \in \{1, \ldots, p\}$.

Part3 wts $M(k)$ is a martingale w.r.t. $\mathfrak{F}$:

Part3.1 wts $M(k)$ is adapted to $\mathfrak{F}$:

Let $k \in \{1, \ldots, p\}$ be arbitrary.
We have to show that $M(k)$ is measurable w.r.t. $\mathcal{F}_k$.
By definition of $\mathcal{F}_k$, $V^+(k)$ and $V^-(k)$ are $\mathcal{F}_k$-measurable.
Note that constant functions are measureable.
We know that the sum of two measurable functions is measurable.
Hence, $1 + V^-(k)$ is $\mathcal{F}_k$-measurable.
And we know that the quotient of two measurable functions is measurable.
Therefore, $M(k)$ is $\mathcal{F}_k$-measurable.

Part3.2 wts $\mathbb{E}(|M(k)|) < \infty$ for all $k \in \{1, \ldots, p\}$:

Let $k \in \{1, \ldots, p\}$.

$$
\begin{aligned}
&\mathbb{E}(|M(k)|) && \\
&&& \text{By definition of } M(k) \\
&= \mathbb{E}\left(\left|\frac{V^+(k)}{1 + V^-(k)}\right|\right) && \\
&&& \text{We can omit the absolute value, because} \\
&&& \text{the term inside is always nonnegative} \\
&= \mathbb{E}\left(\frac{V^+(k)}{1 + V^-(k)}\right) && \\
&&& \text{We know that } j \in \{1, \ldots, p\}. \text{ Therefore,} \\
&&& V^+(k) \leq p \text{ and } V^-(k) \geq 0 \text{ for all } k. \\
&&& \text{So, } \frac{V^+(k)}{1 + V^-(k)} \leq \frac{p}{1 + 0} = p. \\
&\leq \mathbb{E}(p) && \\
&= p && \\
&< \infty. &&
\end{aligned}
$$

Part3.3 wts $\mathbb{E}[M(k+1) \mid \mathcal{F}_k] = M(k)$ for all $k \in \{1, \ldots, p-1\}$:

Let $k \in \{1, \ldots, p-1\}$.

Prep3.3.1 wts $M(k+1) = M(k)$ on the event $\{[k] \notin \mathcal{H}_0\}$:

$$
\begin{aligned}
& M(k+1) \\
& \qquad \text{By definition of } M(k+1) \\
&= \frac{V^+(k+1)}{1+V^-(k+1)} \\
& \qquad \text{By definition of } V^+(k+1) \text{ and } V^-(k+1) \\
&= \frac{|\left\{j:\ j \in \{k+1,\ldots,p\},\ W_{[j]} > 0,\ [j] \in \mathcal{H}_0\right\}|}{1+|\left\{j:\ j \in \{k+1,\ldots,p\},\ W_{[j]} < 0,\ [j] \in \mathcal{H}_0\right\}|} \\
& \qquad \text{The counts in numerator and denominator} \\
& \qquad \text{do not change if we start with } k \text{ instead of } k+1 \\
& \qquad \text{since we know that } [k] \notin \mathcal{H}_0. \\
&= \frac{|\left\{j:\ j \in \{k,\ldots,p\},\ W_{[j]} > 0,\ [j] \in \mathcal{H}_0\right\}|}{1+|\left\{j:\ j \in \{k,\ldots,p\},\ W_{[j]} < 0,\ [j] \in \mathcal{H}_0\right\}|} \\
& \qquad \text{By definition of } V^+(k) \text{ and } V^-(k) \\
&= \frac{V^+(k)}{1+V^-(k)} \\
& \qquad \text{By definition of } M(k) \\
&= M(k).
\end{aligned}
$$

Prep3.3.2 wts $M(k+1) = \dfrac{V^+(k)-1}{1+V^-(k)}$ on the event $\left\{[k] \in \mathcal{H}_0,\ W_{[k]} > 0\right\}$:

$$
\begin{aligned}
& M(k+1) \\
& \qquad \text{By definition of } M(k+1) \\
&= \frac{V^+(k+1)}{1+V^-(k+1)} \\
& \qquad \text{By definition of } V^+(k+1) \text{ and } V^-(k+1) \\
&= \frac{|\left\{j:\ j \in \{k+1,\ldots,p\},\ W_{[j]} > 0,\ [j] \in \mathcal{H}_0\right\}|}{1+|\left\{j:\ j \in \{k+1,\ldots,p\},\ W_{[j]} < 0,\ [j] \in \mathcal{H}_0\right\}|} \\
& \qquad \text{Since we consider the event } \left\{[k] \in \mathcal{H}_0,\ W_{[k]} > 0\right\}, \\
& \qquad \text{the numerator will increase by 1 if we start} \\
& \qquad \text{with } k \text{ instead of } k+1. \text{ The denominator is not affected.} \\
&= \frac{|\left\{j:\ j \in \{k,\ldots,p\},\ W_{[j]} > 0,\ [j] \in \mathcal{H}_0\right\}| - 1}{1+|\left\{j:\ j \in \{k,\ldots,p\},\ W_{[j]} < 0,\ [j] \in \mathcal{H}_0\right\}|} \\
& \qquad \text{By definition of } V^+(k) \text{ and } V^-(k) \\
&= \frac{V^+(k)-1}{1+V^-(k)}.
\end{aligned}
$$

Prep3.3.3 wts $M(k+1) = \dfrac{V^+(k)}{V^-(k)}$ on the event $\left\{[k] \in \mathcal{H}_0,\ W_{[k]} < 0\right\}$:

$$
\begin{aligned}
& M(k+1) \\
& \qquad \text{By definition of } M(k+1) \\
&= \frac{V^+(k+1)}{1+V^-(k+1)}
\end{aligned}
$$

By definition of $V^+(k+1)$ and $V^-(k+1)$

$$= \frac{|\{j : j \in \{k+1,\dots,p\}, W_{[j]} > 0, [j] \in \mathcal{H}_0\}|}{1 + |\{j : j \in \{k+1,\dots,p\}, W_{[j]} < 0, [j] \in \mathcal{H}_0\}|}$$

Since we consider the event $\{[k] \in \mathcal{H}_0, W_{[k]} < 0\}$, the denominator will increase by 1 if we start with k instead of $k+1$. The numerator is not affected.

$$= \frac{|\{j : j \in \{k,\dots,p\}, W_{[j]} > 0, [j] \in \mathcal{H}_0\}|}{1 + |\{j : j \in \{k,\dots,p\}, W_{[j]} < 0, [j] \in \mathcal{H}_0\}| - 1}$$

By definition of $V^+(k)$ and $V^-(k)$

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

Note that $V^-(k) > 0$ in this case, because we consider the event $\{[k] \in \mathcal{H}_0, W_{[k]} < 0\}$ and hence, $V^-(k) \geq 1$.

Prep3.3.4 wts $\mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] = \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \frac{V^+(k)}{V^+(k)+V^-(k)}$:

Note that we can write $V^+(k) = \sum_{j=k}^{p} \mathbb{1}_{\{[j]\in\mathcal{H}_0\}} \cdot \mathbb{1}_{\{W_{[j]}>0\}}$.
Conditional expected value applied on both sides gives

$$\mathbb{E}\left[V^+(k) \mid \mathcal{F}_k\right] = \mathbb{E}\left[\sum_{j=k}^{p} \mathbb{1}_{\{[j]\in\mathcal{H}_0\}} \cdot \mathbb{1}_{\{W_{[j]}>0\}} \mid \mathcal{F}_k\right].$$

Since $V^+(k), \{[j] \in \mathcal{H}_0\} \in \mathcal{F}_k$ for all $j \in \{1,\dots,p\}$, we can pull these terms out and get

$$V^+(k) \cdot \mathbb{E}\left[1 \mid \mathcal{F}_k\right] = \sum_{j=k}^{p} \mathbb{1}_{\{[j]\in\mathcal{H}_0\}} \cdot \mathbb{E}\left[\mathbb{1}_{\{W_{[j]}>0\}} \mid \mathcal{F}_k\right].$$

Hence,

$$V^+(k) = \sum_{j=k}^{p} \mathbb{1}_{\{[j]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[j]} > 0 \mid \mathcal{F}_k\right].$$

By Lemma 3.13 i) we know for $[j] \in \mathcal{H}_0$: $\operatorname{sign}(W_{[j]})$ are independent and $\mathbb{P}\left(\operatorname{sign}(W_{[j]}) = -1\right) = \mathbb{P}\left(\operatorname{sign}(W_{[j]}) = +1\right) = \frac{1}{2}$.
The latter can be written as $\mathbb{P}\left(W_{[j]} < 0\right) = \mathbb{P}\left(W_{[j]} > 0\right) = \frac{1}{2}$.
In other words, the signs of the null MX-Knockoff feature statistics are i.i.d. and hence exchangeable.
By definition of $\mathcal{F}_k$, all nulls are known.
Therefore,

$$\mathbb{1}_{\{[j]\in\mathcal{H}_0\}} \cdot \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] = \mathbb{1}_{\{[j]\in\mathcal{H}_0\}} \cdot \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[j]} > 0 \mid \mathcal{F}_k\right]$$

for all $j \in \{k,\dots,p\}$.

And note that

$\mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot V^+(k)$

We know from above that $V^+(k) = \sum_{j=k}^{p} \mathbb{1}_{\{[j]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[j]} > 0 \mid \mathcal{F}_k\right]$.

$$= \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \sum_{j=k}^{p} \mathbb{1}_{\{[j]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[j]} > 0 \mid \mathcal{F}_k\right]$$

$$= \sum_{j=k}^{p} \mathbb{1}_{\{[j]\in\mathcal{H}_0\}} \cdot \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[j]} > 0 \mid \mathcal{F}_k\right]$$

We know that $\mathbb{1}_{\{[j]\in\mathcal{H}_0\}} \cdot \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] = \mathbb{1}_{\{[j]\in\mathcal{H}_0\}} \cdot \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[j]} > 0 \mid \mathcal{F}_k\right]$.

$$= \sum_{j=k}^{p} \mathbb{1}_{\{[j]\in\mathcal{H}_0\}} \cdot \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right]$$

$$= \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] \cdot \sum_{j=k}^{p} \mathbb{1}_{\{[j]\in\mathcal{H}_0\}}$$

$$= \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] \cdot \left|\left\{j :\ j \in \{k,\dots,p\},\ [j] \in \mathcal{H}_0\right\}\right|$$

Note that $V^+(k) + V^-(k) = \left|\left\{j :\ j \in \{k,\dots,p\},\ [j] \in \mathcal{H}_0\right\}\right|$.

$$= \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] \cdot \left(V^+(k) + V^-(k)\right).$$

Rearranged gives

$$\mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] = \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \frac{V^+(k)}{V^+(k) + V^-(k)}.$$

Therefore,

$$\mathbb{E}\left[M(k+1) \mid \mathcal{F}_k\right]$$
$$= \mathbb{E}\left[M(k+1) \cdot \mathbb{1}_{\{[k]\notin\mathcal{H}_0\}\cup\{[k]\in\mathcal{H}_0\}} \mid \mathcal{F}_k\right]$$

Note that $\mathbb{1}_{C\cup D} = \mathbb{1}_C + \mathbb{1}_D$ for disjoint sets C, D.

$$= \mathbb{E}\left[M(k+1) \cdot \left(\mathbb{1}_{\{[k]\notin\mathcal{H}_0\}} + \mathbb{1}_{\{[k]\in\mathcal{H}_0\}}\right) \mid \mathcal{F}_k\right]$$
$$= \mathbb{E}\left[M(k+1) \cdot \left(\mathbb{1}_{\{[k]\notin\mathcal{H}_0\}} + \mathbb{1}_{\left\{[k]\in\mathcal{H}_0,\, W_{[k]}>0\right\}\cup\left\{[k]\in\mathcal{H}_0,\, W_{[k]}<0\right\}}\right) \mid \mathcal{F}_k\right]$$

Note that $\mathbb{1}_{C\cup D} = \mathbb{1}_C + \mathbb{1}_D$ for disjoint sets C, D.

$$= \mathbb{E}\left[M(k+1) \cdot \left(\mathbb{1}_{\{[k]\notin\mathcal{H}_0\}} + \mathbb{1}_{\left\{[k]\in\mathcal{H}_0,\, W_{[k]}>0\right\}} + \mathbb{1}_{\left\{[k]\in\mathcal{H}_0,\, W_{[k]}<0\right\}}\right) \mid \mathcal{F}_k\right]$$
$$= \mathbb{E}\left[M(k+1) \cdot \mathbb{1}_{\{[k]\notin\mathcal{H}_0\}} + M(k+1) \cdot \mathbb{1}_{\left\{[k]\in\mathcal{H}_0,\, W_{[k]}>0\right\}} + M(k+1) \cdot \mathbb{1}_{\left\{[k]\in\mathcal{H}_0,\, W_{[k]}<0\right\}} \mid \mathcal{F}_k\right]$$
$$= \mathbb{E}\left[M(k+1) \cdot \mathbb{1}_{\{[k]\notin\mathcal{H}_0\}} \mid \mathcal{F}_k\right]$$
$$+ \mathbb{E}\left[M(k+1) \cdot \mathbb{1}_{\left\{[k]\in\mathcal{H}_0,\, W_{[k]}>0\right\}} \mid \mathcal{F}_k\right]$$
$$+ \mathbb{E}\left[M(k+1) \cdot \mathbb{1}_{\left\{[k]\in\mathcal{H}_0,\, W_{[k]}<0\right\}} \mid \mathcal{F}_k\right]$$

By Prep3.3.1, Prep3.3.2 and Prep3.3.3

$$= \mathbb{E}\left[M(k) \cdot \mathbb{1}_{\{[k]\notin\mathcal{H}_0\}} \mid \mathcal{F}_k\right]$$
$$+ \mathbb{E}\left[\frac{V^+(k)-1}{1+V^-(k)} \cdot \mathbb{1}_{\left\{[k]\in\mathcal{H}_0,\, W_{[k]}>0\right\}} \mid \mathcal{F}_k\right]$$
$$+ \mathbb{E}\left[\frac{V^+(k)}{V^-(k)} \cdot \mathbb{1}_{\left\{[k]\in\mathcal{H}_0,\, W_{[k]}<0\right\}} \mid \mathcal{F}_k\right]$$

Note that $\mathbb{1}_{C\cap D} = \mathbb{1}_C \cdot \mathbb{1}_D$ for sets C, D.

$$= \mathbb{E}\left[M(k) \cdot \mathbb{1}_{\{[k]\notin\mathcal{H}_0\}} \mid \mathcal{F}_k\right]$$
$$+ \mathbb{E}\left[\frac{V^+(k)-1}{1+V^-(k)} \cdot \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{1}_{\left\{W_{[k]}>0\right\}} \mid \mathcal{F}_k\right]$$
$$+ \mathbb{E}\left[\frac{V^+(k)}{V^-(k)} \cdot \mathbb{1}_{\{[k]\in\mathcal{H}_0\}} \cdot \mathbb{1}_{\left\{W_{[k]}<0\right\}} \mid \mathcal{F}_k\right]$$

By definition of $\mathcal{F}_k$, we know that $V^+(k), V^-(k), \{[k] \in \mathcal{H}_0\} \in \mathcal{F}_k$.
Since complements of elements of a σ-algebra are
again in that σ-algebra, $\{[k] \in \mathcal{H}_0\}^c = \{[k] \notin \mathcal{H}_0\} \in \mathcal{F}_k$.
Hence we can pull out these terms.

$$
\begin{aligned}
&= M(k) \cdot \mathbb{1}_{\{[k] \notin \mathcal{H}_0\}} \cdot \mathbb{E}\left[1 \mid \mathcal{F}_k\right] \\
&\quad + \frac{V^+(k)-1}{1+V^-(k)} \cdot \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \mathbb{E}\left[\mathbb{1}_{\{W_{[k]}>0\}} \mid \mathcal{F}_k\right] \\
&\quad + \frac{V^+(k)}{V^-(k)} \cdot \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \mathbb{E}\left[\mathbb{1}_{\{W_{[k]}<0\}} \mid \mathcal{F}_k\right] \\
&= M(k) \cdot \mathbb{1}_{\{[k] \notin \mathcal{H}_0\}} \\
&\quad + \frac{V^+(k)-1}{1+V^-(k)} \cdot \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[k]} > 0 \mid \mathcal{F}_k\right] \\
&\quad + \frac{V^+(k)}{V^-(k)} \cdot \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \mathbb{P}\left[W_{[k]} < 0 \mid \mathcal{F}_k\right]
\end{aligned}
$$

By Prep3.3.4

$$
\begin{aligned}
&= M(k) \cdot \mathbb{1}_{\{[k] \notin \mathcal{H}_0\}} \\
&\quad + \frac{V^+(k)-1}{1+V^-(k)} \cdot \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \frac{V^+(k)}{V^+(k)+V^-(k)} \\
&\quad + \frac{V^+(k)}{V^-(k)} \cdot \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \left(1 - \frac{V^+(k)}{V^+(k)+V^-(k)}\right) \\
&= M(k) \cdot \mathbb{1}_{\{[k] \notin \mathcal{H}_0\}} \\
&\quad + \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \frac{V^+(k)-1}{1+V^-(k)} \cdot \frac{V^+(k)}{V^+(k)+V^-(k)} \\
&\quad + \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \frac{V^+(k)}{V^-(k)} \cdot \left(\frac{V^+(k)+V^-(k)}{V^+(k)+V^-(k)} - \frac{V^+(k)}{V^+(k)+V^-(k)}\right) \\
&= M(k) \cdot \mathbb{1}_{\{[k] \notin \mathcal{H}_0\}} \\
&\quad + \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \frac{V^+(k)-1}{1+V^-(k)} \cdot \frac{V^+(k)}{V^+(k)+V^-(k)} \\
&\quad + \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \frac{V^+(k)}{V^-(k)} \cdot \frac{V^-(k)}{V^+(k)+V^-(k)} \\
&= M(k) \cdot \mathbb{1}_{\{[k] \notin \mathcal{H}_0\}} + \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \left(\frac{V^+(k)-1}{1+V^-(k)} \cdot \frac{V^+(k)}{V^+(k)+V^-(k)} + \frac{V^-(k)}{V^-(k)} \cdot \frac{V^+(k)}{V^+(k)+V^-(k)}\right) \\
&= M(k) \cdot \mathbb{1}_{\{[k] \notin \mathcal{H}_0\}} + \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \left(\frac{V^+(k)}{V^+(k)+V^-(k)} \cdot \left(\frac{V^+(k)-1}{1+V^-(k)} + 1\right)\right) \\
&= M(k) \cdot \mathbb{1}_{\{[k] \notin \mathcal{H}_0\}} + \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \left(\frac{V^+(k)}{V^+(k)+V^-(k)} \cdot \left(\frac{V^+(k)-1}{1+V^-(k)} + \frac{1+V^-(k)}{1+V^-(k)}\right)\right) \\
&= M(k) \cdot \mathbb{1}_{\{[k] \notin \mathcal{H}_0\}} + \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \left(\frac{V^+(k)}{V^+(k)+V^-(k)} \cdot \frac{V^+(k)+V^-(k)}{1+V^-(k)}\right) \\
&= M(k) \cdot \mathbb{1}_{\{[k] \notin \mathcal{H}_0\}} + \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot \frac{V^+(k)}{1+V^-(k)}
\end{aligned}
$$

By definition of $M(k)$

$$
\begin{aligned}
&= M(k) \cdot \mathbb{1}_{\{[k] \notin \mathcal{H}_0\}} + \mathbb{1}_{\{[k] \in \mathcal{H}_0\}} \cdot M(k) \\
&= M(k) \cdot \left(\mathbb{1}_{\{[k] \notin \mathcal{H}_0\}} + \mathbb{1}_{\{[k] \in \mathcal{H}_0\}}\right)
\end{aligned}
$$

Note that $\mathbb{1}_{C \cup D} = \mathbb{1}_C + \mathbb{1}_D$ for disjoint sets C, D.

$$
\begin{aligned}
&= M(k) \cdot \mathbb{1}_{\{[k] \notin \mathcal{H}_0\} \cup \{[k] \in \mathcal{H}_0\}} \\
&= M(k).
\end{aligned}
$$

Since $k \in \{1, \ldots, p-1\}$ was arbitrary, we have shown $\mathbb{E}\left[M(k+1) \mid \mathcal{F}_k\right] = M(k)$ for all $k \in \{1, \ldots, p-1\}$.

Define $N_0 = |\mathcal{H}_0|$ as the total number of nulls.

Prep4.1 wts $V^+(1) \sim \text{Binomial}\left(N_0, \frac{1}{2}\right)$:

Note that we can write

$$V^+(1) = |\left\{j:\ j \in \{1,\ldots,p\},\ W_{[j]} > 0,\ [j] \in \mathcal{H}_0\right\}| = |\left\{j:\ j \in \{1,\ldots,p\},\ \text{sign}(W_{[j]}) = +1,\ [j] \in \mathcal{H}_0\right\}|.$$

Define $U_j = \mathbb{1}_{\left\{\text{sign}(W_{[j]})=+1\right\}}$.
Note that $\mathbb{P}\left(U_j = 1\right) = \mathbb{P}\left(\text{sign}(W_{[j]}) = +1\right) = \frac{1}{2}$ for all j where $[j] \in \mathcal{H}_0$ by Lemma 3.13 i).
By Lemma 3.13 i), $\text{sign}(W_{[j]})$ are independent for j where $[j] \in \mathcal{H}_0$.
This means that U_j are independent random variables with $U_j \sim \text{Bernoulli}\left(\frac{1}{2}\right)$ for all j where $[j] \in \mathcal{H}_0$.
Short: $U_j \overset{i.i.d.}{\sim} \text{Bernoulli}\left(\frac{1}{2}\right)$ for j where $[j] \in \mathcal{H}_0$.
Note that $\sum_{\substack{j \in \{1,\ldots,p\},\\ [j] \in \mathcal{H}_0}} U_j = V^+(1)$.
By definition of the Binomial-distribution, $V^+(1) \sim \text{Binomial}\left(N_0, \frac{1}{2}\right)$.

Prep4.2 wts $V^-(1) = N_0 - V^+(1)$:

$$\begin{aligned}
&V^-(1)\\
&= |\left\{j:\ j \in \{1,\ldots,p\},\ W_{[j]} < 0,\ [j] \in \mathcal{H}_0\right\}| \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{Note that } N_0 = |\mathcal{H}_0|.\\
&= N_0 - |\left\{j:\ j \in \{1,\ldots,p\},\ W_{[j]} > 0,\ [j] \in \mathcal{H}_0\right\}| \\
&= N_0 - V^+(1).
\end{aligned}$$

Part4 wts $\mathbb{E}\left[\frac{V^+(K)}{1+V^-(K)}\right] \leq 1$:

$$\begin{aligned}
&\mathbb{E}\left[\frac{V^+(K)}{1+V^-(K)}\right]\\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{By Part2, we know that } K \text{ is a bounded stopping time.}\\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{By Part3, we know that } \frac{V^+(k)}{1+V^-(k)} \text{ is a martingale.}\\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{Therefore, we can use Lemma A.4 i)}\\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{(optional stopping theorem for martingales).}\\
&= \mathbb{E}\left[\frac{V^+(1)}{1+V^-(1)}\right]\\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{By Prep4.2}
\end{aligned}$$

$$= \mathbb{E}\left[\frac{V^+(1)}{1+N_0-V^+(1)}\right]$$

In general, it holds for a discrete random variable Y and function g that $\mathbb{E}(g(Y)) = \sum_y g(y)\cdot\mathbb{P}(Y=y)$. In our case, $g(x) := \frac{x}{1+N_0-x}$ and $Y := V^+(1)$. Note that $N_0 \le p$ and $Y \in \{0,\dots,N_0\}$.

$$= \sum_{i=0}^{N_0} g(i)\cdot\mathbb{P}(Y=i)$$

By definition of g

$$= \sum_{i=0}^{N_0} \frac{i}{1+N_0-i}\cdot\mathbb{P}(Y=i)$$

By Prep4.1, $V^+(1) \sim \text{Binomial}\left(N_0, \frac{1}{2}\right)$.

$$= \sum_{i=0}^{N_0} \frac{i}{1+N_0-i}\cdot\binom{N_0}{i}\cdot\left(\frac{1}{2}\right)^i\cdot\left(1-\frac{1}{2}\right)^{N_0-i}$$

We can start the sum at $i=1$.

$$= \sum_{i=1}^{N_0} \frac{i}{1+N_0-i}\cdot\binom{N_0}{i}\cdot\left(\frac{1}{2}\right)^i\cdot\left(1-\frac{1}{2}\right)^{N_0-i}$$

$$= \sum_{i=1}^{N_0} \frac{i}{1+N_0-i}\cdot\frac{N_0!}{i!\cdot(N_0-i)!}\cdot\left(\frac{1}{2}\right)^i\cdot\left(\frac{1}{2}\right)^{N_0-i}$$

$$= \sum_{i=1}^{N_0} \frac{1}{1+N_0-i}\cdot\frac{N_0!}{(i-1)!\cdot(N_0-i)!}\cdot\left(\frac{1}{2}\right)^{i-1}\cdot\left(\frac{1}{2}\right)^1\cdot\left(\frac{1}{2}\right)^{N_0-(i-1)}\cdot\left(\frac{1}{2}\right)^{-1}$$

Note that $1+N_0-i = N_0-(i-1)$.

$$= \sum_{i=1}^{N_0} \frac{N_0!}{(i-1)!\cdot(N_0-(i-1))!}\cdot\left(\frac{1}{2}\right)^{i-1}\cdot\left(\frac{1}{2}\right)^{N_0-(i-1)}$$

$$= \sum_{i=1}^{N_0} \binom{N_0}{i-1}\cdot\left(\frac{1}{2}\right)^{i-1}\cdot\left(\frac{1}{2}\right)^{N_0-(i-1)}$$

Substitution: set $j = i-1$.

$$= \sum_{j=0}^{N_0-1} \underbrace{\binom{N_0}{j}\cdot\left(\frac{1}{2}\right)^j\cdot\left(\frac{1}{2}\right)^{N_0-j}}_{=\mathbb{P}(Y=j)}$$

The sum of probabilities the random variable Y can achieve has to be smaller than 1.

$$\le 1.$$

Part5 wts FDR $\le q$:

FDR

By Definition 3.11 ii)

$$
\begin{aligned}
&= \mathbb{E}\left[\text{FDP}\right] \\
&\qquad\qquad \text{By Part1, we know that } \text{FDP} \leq q \cdot \frac{V^+(K)}{1+V^-(K)} \\
&\leq \mathbb{E}\left[q \cdot \frac{V^+(K)}{1+V^-(K)}\right] \\
&= q \cdot \mathbb{E}\left[\frac{V^+(K)}{1+V^-(K)}\right] \\
&\qquad\qquad \text{By Part4, we know that } \mathbb{E}\left[\frac{V^+(K)}{1+V^-(K)}\right] \leq 1. \\
&\leq q \cdot 1 \\
&= q.
\end{aligned}
$$

As $q \in [0,1]$ was arbitrary but fixed, we have shown the result for all $q \in [0,1]$.

And note that this result holds even conditionally on $\mathbf{Y}$, because we can use Lemma 3.13 ii) (instead of just using Lemma 3.13 i)) in Prep4.1.

□

3.3 Power

Candes et al. (2018) do not present theoretical work on power, but simulations were used to compare MX-Knockoff methods to existing alternatives. The problem is that there are no existing FDR-controlling procedures for the setting for which MX-Knockoffs are designed for. Mainly, BHq procedures were used for comparison and only MX-Knockoff method (instead of MX-Knockoff+ method) was considered, because comparable BHq procedures did not provide exact FDR-control.
Authors created a conditional randomization test as an alternative to the controlled variable selection problem. Together with BHq (conditional randomization test was used to produce p-values and BHq was used to select variables based on those p-values), it resulted in higher power than the MX-Knockoff method, but computational cost was huge. Even for $n = 400$, $p = 600$ and after application of several speed-ups/shortcuts.
Authors also used simulations in settings where $\mathbf{Y}$ comes from a Gaussian or Binomial linear model and the design matrix $\mathbf{X}$ is "i.i.d. $\mathcal{N}(0, 1/n)$". Power of MX-Knockoff method was superior to BHq procedures for $n = 3000$ and $p = 1000$ (and $p = 6000$), except for cases where BHq procedures lost FDR-control (e.g. in settings with high correlation between covariates). For $n = 3000$ and $p = 1000$ and Gaussian response, MX-Knockoff method had higher power than FX-Knockoff method.

For Gaussian variables, MX-Knockoff method was robust to errors in the estimation of the true joint distribution of the covariates. Even for covariance estimates very far from the truth (covariance estimates were used to generate the MX-Knockoff matrix), FDR-control was never violated. And power remained roughly constant until a mix of 50% exact covariance and 50% empirical covariance was reached.

Fan et al. (2017) provide theoretical power analyses for methods similar to MX-Knockoff procedures in high-dimensional linear models.

Gimenez and Zou (2018) introduces a multi-MX-Knockoff method, where several multi-MX-Knockoffs $\tilde{X}$ are sampled simultaneously. This method seems to have higher power (than MX-Knockoff procedures) in settings where the number of non-nulls is small. For a higher number of non-nulls, the power decreases with the number of sampled multi-MX-Knockoffs.

3.4 Examples

Note again, that we abuse notation in this section and denote random variables by $(X_1, \dots, X_p)$ instead of $(X_{\cdot 1}, \dots, X_{\cdot p})$.

3.4.1 Exact construction of MX Knockoffs

Construction of MX-Knockoffs in practice was, except for special cases, left to further research. We will show the construction for a case where $\mathbf{X}$ has Gaussian rows. But first, we present an abstract algorithm which shows how valid MX-Knockoffs can be produced (in theory) for an arbitrary distribution of $\mathbf{X}$. To prove that this algorithm produces valid MX-Knockoffs, the next Proposition will be helpful.

Proposition 3.16. *(Proposition 3.5 in Candes et al. (2018)). The random variables* $\left(\tilde{X}_1, \dots, \tilde{X}_p\right)$ *are MX-Knockoffs for* $(X_1, \dots, X_p)$ *if and only if*
i) for any $j \in \{1, \dots, p\}$,

$$\left(X_j, \tilde{X}_j\right) \mid \left(X_{-j}, \tilde{X}_{-j}\right) \overset{d}{=} \left(\tilde{X}_j, X_j\right) \mid \left(X_{-j}, \tilde{X}_{-j}\right), \quad \textit{and}$$

ii) $Y \perp\!\!\!\perp \tilde{X} | X$ *holds.*

Proof:

Note that by Definition 3.6 ii) we know that $Y \perp\!\!\!\perp \tilde{X} | X$.
Therefore, equivalence of ii) is trivially satisfied (it is exactly the same).

So, it remains to show that i) is equivalent to Definition 3.6 i).
More explicit, we have to show that

$$\forall S \subseteq \{1, \dots, p\} : \left(X, \tilde{X}\right)_{\text{swap}(S)} \overset{d}{=} \left(X, \tilde{X}\right) \iff \forall j \in \{1, \dots, p\} : \left(X_j, \tilde{X}_j\right) \mid \left(X_{-j}, \tilde{X}_{-j}\right) \overset{d}{=} \left(\tilde{X}_j, X_j\right) \mid \left(X_{-j}, \tilde{X}_{-j}\right).$$

Part1 wts $\Longrightarrow$:

> Let $j \in \{1, \dots, p\}$ be arbitrary but fixed.
> We have to show that
>
> $$\mathbb{P}\left(X_j \leq a, \tilde{X}_j \leq b \mid X_{-j} \leq x_{-j}, \tilde{X}_{-j} \leq \tilde{x}_{-j}\right) = \mathbb{P}\left(\tilde{X}_j \leq a, X_j \leq b \mid X_{-j} \leq x_{-j}, \tilde{X}_{-j} \leq \tilde{x}_{-j}\right).$$
>
> By definition of conditional probability, this is equal to
>
> $$\frac{\mathbb{P}\left(X_j \leq a, \tilde{X}_j \leq b, X_{-j} \leq x_{-j}, \tilde{X}_{-j} \leq \tilde{x}_{-j}\right)}{\mathbb{P}\left(X_{-j} \leq x_{-j}, \tilde{X}_{-j} \leq \tilde{x}_{-j}\right)} = \frac{\mathbb{P}\left(\tilde{X}_j \leq a, X_j \leq b, X_{-j} \leq x_{-j}, \tilde{X}_{-j} \leq \tilde{x}_{-j}\right)}{\mathbb{P}\left(X_{-j} \leq x_{-j}, \tilde{X}_{-j} \leq \tilde{x}_{-j}\right)}.$$

Note that both denominators are the same, therefore it remains to show that

$$\mathbb{P}\left(X_j \leq a, \tilde{X}_j \leq b, X_{-j} \leq x_{-j}, \tilde{X}_{-j} \leq \tilde{x}_{-j}\right) = \mathbb{P}\left(\tilde{X}_j \leq a, X_j \leq b, X_{-j} \leq x_{-j}, \tilde{X}_{-j} \leq \tilde{x}_{-j}\right).$$

In other words, we have to show that

$$\left(X_j, \tilde{X}_j, X_{-j}, \tilde{X}_{-j}\right) \stackrel{d}{=} \left(\tilde{X}_j, X_j, X_{-j}, \tilde{X}_{-j}\right).$$

We can use a measurable function which changes the position of entries on both sides (move the first entry to the position between $\tilde{X}_{j-1}$ and $\tilde{X}_{j+1}$ and move the second entry to the position between X_{j-1} and X_{j+1}) and use Lemma A.13 i) to get

$$\left(X_1, \dots X_{j-1}, \tilde{X}_j, X_{j+1}, \dots, X_p, \tilde{X}_1, \dots, \tilde{X}_{j-1}, X_j, \tilde{X}_{j+1}, \dots, \tilde{X}_p\right) \stackrel{d}{=} \left(X_1, \dots, X_p, \tilde{X}_1, \dots, \tilde{X}_p\right).$$

Short,

$$\left(X, \tilde{X}\right)_{\text{swap}(\{j\})} \stackrel{d}{=} \left(X, \tilde{X}\right).$$

By Definition 3.6 i), we know that for any subset $S \subseteq \{1, \dots, p\}$, it holds that $\left(X, \tilde{X}\right)_{\text{swap}(S)} \stackrel{d}{=} \left(X, \tilde{X}\right)$.
We can choose $S = \{j\}$ and get that $\left(X, \tilde{X}\right)_{\text{swap}(\{j\})} \stackrel{d}{=} \left(X, \tilde{X}\right)$.

Because j was arbitrary but fixed, we proved the claim for all $j \in \{1, \dots, p\}$.

Part2 wts $\Longleftarrow$:

We know from Part1, that

$$\left(X_j, \tilde{X}_j\right) | \left(X_{-j}, \tilde{X}_{-j}\right) \stackrel{d}{=} \left(\tilde{X}_j, X_j\right) | \left(X_{-j}, \tilde{X}_{-j}\right)$$

is the same as

$$\left(X, \tilde{X}\right)_{\text{swap}(\{j\})} \stackrel{d}{=} \left(X, \tilde{X}\right). \tag{20}$$

Let $S \subseteq \{1, \dots, p\}$ be arbitrary but fixed.

The claim can now be proved by repeatedly using (20) for all indices in S.
To visualize this argument, let wlog $S = \{1, \dots, m\}$ for $m \leq p$.
We can use (20) for $j = 1$ to get

$$\left(X, \tilde{X}\right)_{\text{swap}(\{1\})} \stackrel{d}{=} \left(X, \tilde{X}\right).$$

Because the LHS is equal in distribution to the RHS, we can apply (20) to the LHS, this time with the next index in S; that is $j = 2$. Therefore, we get

$$\left(\left(X, \tilde{X}\right)_{\text{swap}(\{1\})}\right)_{\text{swap}(\{2\})} \stackrel{d}{=} \left(X, \tilde{X}\right)_{\text{swap}(\{1\})} \stackrel{d}{=} \left(X, \tilde{X}\right).$$

Note that $\left(\left(X, \tilde{X}\right)_{\text{swap}(\{1\})}\right)_{\text{swap}(\{2\})} = \left(X, \tilde{X}\right)_{\text{swap}(\{1,2\})}$, and therefore, we have

$$\left(X, \tilde{X}\right)_{\text{swap}(\{1,2\})} \stackrel{d}{=} \left(X, \tilde{X}\right).$$

Repeating this argument for $j = 3, \dots, m$ gives

$$\left(X, \tilde{X}\right)_{\text{swap}(S)} \stackrel{d}{=} \left(X, \tilde{X}\right).$$

Because S was arbitrary but fixed, we proved the claim for all $S \subseteq \{1, \ldots, p\}$.

□

The next Example shows that construction of MX-Knockoffs is easy, if the components of the vector $X = (X_1, \ldots, X_p)$ are independent.

Example 3.17. Let $X = (X_1, \ldots, X_p)$ be a vector with independent components. Let $\tilde{X} = \left(\tilde{X}_1, \ldots, \tilde{X}_p\right)$ be any vector independently sampled from the joint distribution as X. Then $\left(\tilde{X}_1, \ldots, \tilde{X}_p\right)$ are valid MX-Knockoffs for $(X_1, \ldots, X_p)$.

Proof:

By construction, $X, \tilde{X}$ are independent and therefore, $X_1, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_p$ are independent random variables.
Also by construction, $X \stackrel{d}{=} \tilde{X}$ (because $\tilde{X}$ is sampled from the same joint distribution as X). In particular, $X_j \stackrel{d}{=} \tilde{X}_j$ for all j. Together with the independence of $X_j, \tilde{X}_j$, this means that $X_j, \tilde{X}_j$ are i.i.d. for all j.

Let $j \in \{1, \ldots, p\}$ be arbitrary but fixed.

$$\begin{aligned}
&\left(X_j, \tilde{X}_j\right) \mid \left(X_{-j}, \tilde{X}_{-j}\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{Note that all random} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{variables are independent.} \\
&\stackrel{d}{=} \left(X_j, \tilde{X}_j\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad X_j, \tilde{X}_j \text{ are i.i.d. for all } j \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{and therefore exchangeable.} \\
&\stackrel{d}{=} \left(\tilde{X}_j, X_j\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{Note that all random} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{variables are independent.} \\
&\stackrel{d}{=} \left(\tilde{X}_j, X_j\right) \mid \left(X_{-j}, \tilde{X}_{-j}\right).
\end{aligned}$$

Because j was arbitrary but fixed, we proved the claim for all $j \in \{1, \ldots, p\}$.

□

Comments:
i) Of course, the assumption that the variables $\mathbf{X}_j$ (and therefore the components of a row-vector X) will be independent is very strong and often not satisfied in practice.

ii) If the variables $\mathbf{X}_j$ (and therefore the components of the vector X) are not necessarily independent, the next algorithm shows an abstract way of producing MX-Knockoffs:

3.4.1.1 SCIP Algorithm

Algorithm 1: Sequential Conditional Independent Pairs (SCIP)

Sample $\tilde{X}_1$ from $\mathcal{L}(X_1|X_{-1})$ such that $\tilde{X}_1 \perp\!\!\!\perp X_1|X_{-1}$.
for $j = 2$ **to** p **do**
    Sample $\tilde{X}_j$ from $\mathcal{L}\left(X_j|X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)$ such that $\left(\tilde{X}_j \perp\!\!\!\perp X_j\right) | \left(X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)$.
end

Comments:
i) Note that $\mathcal{L}$ denotes the law/distribution.
ii) Sampling $\tilde{X}_j$ conditionally independent of X_j is done to make sure that X_j and $\tilde{X}_j$ (given all other variables) are as different as possible.

Example 3.18. Let $X = (X_1, X_2, X_3)$. Then Algorithm 1 works as follows:
Sample $\tilde{X}_1$ from $\mathcal{L}(X_1|X_2, X_3)$.
Then $\mathcal{L}\left(X_1, X_2, X_3, \tilde{X}_1\right)$ is available and therefore, $\mathcal{L}\left(X_2|X_1, X_3, \tilde{X}_1\right)$ is known.
Sample $\tilde{X}_2$ from $\mathcal{L}\left(X_2|X_1, X_3, \tilde{X}_1\right)$.
Then $\mathcal{L}\left(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2\right)$ is available and therefore, $\mathcal{L}\left(X_3|X_1, X_2, \tilde{X}_1, \tilde{X}_2\right)$ is known.
Sample $\tilde{X}_3$ from $\mathcal{L}\left(X_3|X_1, X_2, \tilde{X}_1, \tilde{X}_2\right)$.

Comments:
i) The first problem using this algorithm is that it is computationally hard. The set of variables grows at each step which means that at each step, the conditional distribution has to be "recalculated".
ii) The second problem is that it is not clear how specifying these distributions can be done in practice. And it is not clear how to sample from all these conditional distributions. The algorithm is only given abstractly.
iii) The ordering in which MX-Knockoffs are created is not important. We could have started with $\tilde{X}_2$ or $\tilde{X}_3$ in the last example. The results would differ, but Algorithm 1 would still satisfy Definition 3.6. The next Lemma helps to understand why this is the case. The fact that ordering is irrelevant was used by Candes et al. (2018) in some of their applications: Columns of $\mathbf{X}$ were permuted and Algorithm 1 was used to create MX-Knockoffs for each row to get a MX-Knockoff matrix with permuted columns. And at the last step, the permutation of columns was reverted again to get $\tilde{\mathbf{X}}$.

Lemma 3.19. *Algorithm 1 produces valid MX-Knockoffs (as defined in Definition 3.6).*

Proof:

By Proposition 3.16, we have to show that two properties hold. The second property, $\tilde{X} \perp\!\!\!\perp Y|X$, is satisfied because Algorithm 1 produces $\tilde{X}$ without using Y.
Therefore, it remains to show that Algorithm 1 satisfies the first property, namely that for any $j \in \{1, \dots, p\}$,

$$\left(X_j, \tilde{X}_j\right) | \left(X_{-j}, \tilde{X}_{-j}\right) \stackrel{d}{=} \left(\tilde{X}_j, X_j\right) | \left(X_{-j}, \tilde{X}_{-j}\right).$$

In Part1 of the Proof of Proposition 3.16, we have shown that this is equivalent to

$$\left(X, \tilde{X}\right)_{\text{swap}(\{j\})} \stackrel{d}{=} \left(X, \tilde{X}\right)$$

for any $j \in \{1, \dots, p\}$.

Therefore more explicit, we have to show for all $j \in \{1, \dots, p\}$:

$$\left(X_1, \dots, X_p, \tilde{X}_1, \dots, \tilde{X}_p\right)_{\text{swap}(\{j\})} \stackrel{d}{=} \left(X_1, \dots, X_p, \tilde{X}_1, \dots, \tilde{X}_p\right).$$

We will prove this by induction on j for the case where X_i are discrete random variables.
Candes et al. (2018) claim that the general case follows the same argument with a slightly more careful measure-theoretic treatment using Radon-Nikodym derivatives instead of probability mass functions.

We will use a sloppy notation throughout this proof. Instead of explicitly writing expressions like $\mathbb{P}\left(X_1 = x_1, \dots, X_p = x_p, \tilde{X}_1 = \tilde{x}_1, \dots, \tilde{X}_p = \tilde{x}_p\right)$, we only denote values of "noteworthy" random variables, e.g. $\mathbb{P}\left(X_1, \dots, X_p = x_p, \tilde{X}_1, \dots, \tilde{X}_p\right)$.

Induction hypothesis: After j steps, every pair $\left(X_k, \tilde{X}_k\right)$ is exchangeable in the joint distribution of $\left(X_1, \dots, X_p, \tilde{X}_1, \dots, \tilde{X}_j\right)$ for $k = 1, \dots, j$.

1) base case $j = 1$:

The algorithm samples $\tilde{X}_1$ from $\mathcal{L}(X_1|X_{-1})$. This means that $\tilde{X}_1|X_{-1} \stackrel{d}{=} X_1|X_{-1}$.
Furthermore, the algorithm samples $\tilde{X}_1$ such that $\left(\tilde{X}_1 \perp\!\!\!\perp X_1\right)|X_{-1}$.
Both properties together can be phraised as: conditional on X_{-1}, the random variables X_1 and $\tilde{X}_1$ are i.i.d.
We know in general that i.i.d. implies exchangeability, therefore, conditional on X_{-1},
we have $\left(X_1, \tilde{X}_1\right) \stackrel{d}{=} \left(\tilde{X}_1, X_1\right)$. Short,

$$\left(X_1, \tilde{X}_1\right)|X_{-1} \stackrel{d}{=} \left(\tilde{X}_1, X_1\right)|X_{-1}.$$

By using the definition of conditional probability and then recognizing that denominators are the same, this is equivalent to

$$\left(X_1, \tilde{X}_1, X_{-1}\right) \stackrel{d}{=} \left(\tilde{X}_1, X_1, X_{-1}\right).$$

More explicit,

$$\left(X_1, \tilde{X}_1, X_2, \dots, X_p\right) \stackrel{d}{=} \left(\tilde{X}_1, X_1, X_2, \dots, X_p\right).$$

We can use a measurable function which changes the position of entries on both sides (move the second entry to the last position) and use Lemma A.13 i) to get

$$\left(X_1, X_2, \ldots, X_p, \tilde{X}_1\right) \stackrel{d}{=} \left(\tilde{X}_1, X_2, \ldots, X_p, X_1\right).$$

So, the pair $\left(X_1, \tilde{X}_1\right)$ is exchangeable in the joint distribution of $\left(X_1, \ldots, X_p, \tilde{X}_1\right)$.

2) Inductive step - Assuming the induction hypothesis holds until $j-1$, we prove that it holds after j steps:

Define $\mathfrak{X}_j$ to be the set of all possible values of X_j and define $\tilde{\mathfrak{X}}_j$ to be the set of all possible values of $\tilde{X}_j$.

As announced in the comments after Definition 3.5, we will use the function swap also for truncated input-vectors like $\left(X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)$ to perform swapping of a pair $\left(X_k, \tilde{X}_k\right)$.

Prep1 wts $\left(X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)_{\text{swap}(\{k\})} \stackrel{d}{=} \left(X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)$ for all $k \in \{1, \ldots, j-1\}$:

Let $k \in \{1, \ldots, j-1\}$ be arbitrary but fixed.

From step $j-1$, we know that

$$\left(X_1, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)_{\text{swap}(\{k\})} \stackrel{d}{=} \left(X_1, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right) \tag{21}$$

for all $k \in \{1, \ldots, j-1\}$.
Define $X_i' = \begin{cases} \tilde{X}_i, & i = k \\ X_i, & i \neq k \end{cases}$ and $\tilde{X}_i' = \begin{cases} X_i, & i = k \\ \tilde{X}_i, & i \neq k \end{cases}$ for all $i \in \{1, \ldots, j-1\}$.
And note that X_j is not affected.
Then, (21) can be written as

$$\left(X_1', \ldots, X_{j-1}', X_j, X_{j+1}', \ldots, X_p', \tilde{X}_1', \ldots, \tilde{X}_{j-1}'\right) \stackrel{d}{=} \left(X_1, \ldots, X_{j-1}, X_j, X_{j+1}, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right).$$

More explicit (we use a sloppy notation),

$$\begin{aligned} &\mathbb{P}\left(X_1', \ldots, X_{j-1}', X_j = x_j, X_{j+1}', \ldots, X_p', \tilde{X}_1', \ldots, \tilde{X}_{j-1}'\right) \\ &= \mathbb{P}\left(X_1, X_2, \ldots, X_{j-1}, X_j = x_j, X_{j+1}, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right). \end{aligned}$$

And we can now marginalize out X_j on both sides. To do that, we first sum over all possible values of X_j.

$$\begin{aligned} &\sum_{x_j \in \mathfrak{X}_j} \mathbb{P}\left(X_1', \ldots, X_{j-1}', X_j = x_j, X_{j+1}', \ldots, X_p', \tilde{X}_1', \ldots, \tilde{X}_{j-1}'\right) \\ &= \sum_{x_j \in \mathfrak{X}_j} \mathbb{P}\left(X_1, X_2, \ldots, X_{j-1}, X_j = x_j, X_{j+1}, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right). \end{aligned}$$

We can write that as

$$\mathbb{P}\left(X_1', \ldots, X_{j-1}', X_{j+1}', \ldots, X_p', \tilde{X}_1', \ldots, \tilde{X}_{j-1}'\right) = \mathbb{P}\left(X_1, X_2, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right).$$

In other words,

$$\left(X'_1,\dots,X'_{j-1},X'_{j+1},\dots,X'_p,\tilde{X}'_1,\dots,\tilde{X}'_{j-1}\right) \overset{d}{=} \left(X_1,X_2,\dots,X_{j-1},X_{j+1},\dots,X_p,\tilde{X}_1,\dots,\tilde{X}_{j-1}\right).$$

By using the definition of X'_i and $\tilde{X}'_i$, this is

$$\left(X_1,\dots X_{j-1},X_{j+1},\dots,X_p,\tilde{X}_1,\dots,\tilde{X}_{j-1}\right)_{\text{swap}(\{k\})} \overset{d}{=} \left(X_1,\dots X_{j-1},X_{j+1},\dots,X_p,\tilde{X}_1,\dots,\tilde{X}_{j-1}\right).$$

Because $k \in \{1,\dots,j-1\}$ was arbitrary but fixed, we proved the claim for all $k \in \{1,\dots,j-1\}$.

Prep2 wts $X_j|\left(X_{-j},\tilde{X}_1,\dots,\tilde{X}_{j-1}\right)_{\text{swap}(\{k\})} \overset{d}{=} \tilde{X}_j|\left(X_{-j},\tilde{X}_1,\dots,\tilde{X}_{j-1}\right)_{\text{swap}(\{k\})}$ for all $k \in \{1,\dots,j-1\}$:

Let $k \in \{1,\dots,j-1\}$ be arbitrary but fixed.
By construction of Algorithm 1, we have $\mathcal{L}\left(X_j|X_{-j},\tilde{X}_1,\dots,\tilde{X}_{j-1}\right) = \mathcal{L}\left(\tilde{X}_j|X_{-j},\tilde{X}_1,\dots,\tilde{X}_{j-1}\right)$.
In other words (by using our sloppy notation again),

$$\mathbb{P}\left(X_j|X_{-j},\tilde{X}_1,\dots,\tilde{X}_{j-1}\right) = \mathbb{P}\left(\tilde{X}_j|X_{-j},\tilde{X}_1,\dots,\tilde{X}_{j-1}\right).$$

By definition of conditional probability (used on both sides), this is

$$\frac{\mathbb{P}\left(X_1,\dots,X_p,\tilde{X}_1,\dots,\tilde{X}_{j-1}\right)}{\mathbb{P}\left(X_{-j},\tilde{X}_1,\dots,\tilde{X}_{j-1}\right)} = \frac{\mathbb{P}\left(X_{-j},\tilde{X}_1,\dots,\tilde{X}_{j-1},\tilde{X}_j\right)}{\mathbb{P}\left(X_{-j},\tilde{X}_1,\dots,\tilde{X}_{j-1}\right)}.$$

Because both denominators are the same, we have

$$\mathbb{P}\left(X_1,\dots,X_p,\tilde{X}_1,\dots,\tilde{X}_{j-1}\right) = \mathbb{P}\left(X_{-j},\tilde{X}_1,\dots,\tilde{X}_{j-1},\tilde{X}_j\right).$$

By using Lemma A.13 i), we get

$$\mathbb{P}\left(\left(X_1,\dots,X_p,\tilde{X}_1,\dots,\tilde{X}_{j-1}\right)_{\text{swap}(\{k\})}\right) = \mathbb{P}\left(\left(X_{-j},\tilde{X}_1,\dots,\tilde{X}_{j-1},\tilde{X}_j\right)_{\text{swap}(\{k\})}\right).$$

Note that the swap-function is not using X_j and $\tilde{X}_j$. Therefore, we can write

$$\mathbb{P}\left(X_j,\left(X_{-j},\tilde{X}_1,\dots,\tilde{X}_{j-1}\right)_{\text{swap}(\{k\})}\right) = \mathbb{P}\left(\left(X_{-j},\tilde{X}_1,\dots,\tilde{X}_{j-1}\right)_{\text{swap}(\{k\})},\tilde{X}_j\right).$$

By definition of conditional probability (used on both sides),

$$\begin{aligned}&\mathbb{P}\left(X_j|\left(X_{-j},\tilde{X}_1,\dots,\tilde{X}_{j-1}\right)_{\text{swap}(\{k\})}\right)\cdot\mathbb{P}\left(\left(X_{-j},\tilde{X}_1,\dots,\tilde{X}_{j-1}\right)_{\text{swap}(\{k\})}\right)\\ &=\mathbb{P}\left(\tilde{X}_j|\left(X_{-j},\tilde{X}_1,\dots,\tilde{X}_{j-1}\right)_{\text{swap}(\{k\})}\right)\cdot\mathbb{P}\left(\left(X_{-j},\tilde{X}_1,\dots,\tilde{X}_{j-1}\right)_{\text{swap}(\{k\})}\right).\end{aligned}$$

Hence,

$$\mathbb{P}\left(X_j|\left(X_{-j},\tilde{X}_1,\dots,\tilde{X}_{j-1}\right)_{\text{swap}(\{k\})}\right) = \mathbb{P}\left(\tilde{X}_j|\left(X_{-j},\tilde{X}_1,\dots,\tilde{X}_{j-1}\right)_{\text{swap}(\{k\})}\right).$$

In other words,

$$X_j|\left(X_{-j},\tilde{X}_1,\dots,\tilde{X}_{j-1}\right)_{\text{swap}(\{k\})} \overset{d}{=} \tilde{X}_j|\left(X_{-j},\tilde{X}_1,\dots,\tilde{X}_{j-1}\right)_{\text{swap}(\{k\})}.$$

Because $k \in \{1,\dots,j-1\}$ was arbitrary but fixed, we proved the claim for all $k \in \{1,\dots,j-1\}$.

Prep3 wts $\left(X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}, \tilde{X}_j\right)_{\text{swap}(\{k\})} \stackrel{d}{=} \left(X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}, \tilde{X}_j\right)$ for all $k \in \{1, \ldots, j-1\}$:

Let $k \in \{1, \ldots, j-1\}$ be arbitrary but fixed.
Let $b \in \tilde{\mathfrak{X}}_j$ be arbitrary but fixed.

$$\mathbb{P}\left(\left(X_1, \ldots X_{j-1}, X_{j+1}, \ldots, X_p, \tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_{j-1}, \tilde{X}_j = b\right)_{\text{swap}(\{k\})}\right)$$

Define $X'_i = \begin{cases} \tilde{X}_i, & i = k \\ X_i, & i \neq k \end{cases}$ and $\tilde{X}'_i = \begin{cases} X_i, & i = k \\ \tilde{X}_i, & i \neq k \end{cases}$ for $i \in \{1, \ldots, j-1\}$.

Note that $\tilde{X}_j$ is not affected.

$$= \mathbb{P}\left(X'_1, \ldots, X'_{j-1}, X'_{j+1}, \ldots, X'_p, \tilde{X}'_1, \ldots, \tilde{X}'_{j-1}, \tilde{X}_j = b\right)$$

By definition of conditional probability.

$$= \mathbb{P}\left(\tilde{X}_j = b | X'_1, \ldots, X'_{j-1}, X'_{j+1}, \ldots, X'_p, \tilde{X}'_1, \ldots, \tilde{X}'_{j-1}\right) \cdot \mathbb{P}\left(X'_1, \ldots, X'_{j-1}, X'_{j+1}, \ldots, X'_p, \tilde{X}'_1, \ldots, \tilde{X}'_{j-1}\right)$$

By Prep2

$$= \mathbb{P}\left(X_j = b | X'_1, \ldots, X'_{j-1}, X'_{j+1}, \ldots, X'_p, \tilde{X}'_1, \ldots, \tilde{X}'_{j-1}\right) \cdot \mathbb{P}\left(X'_1, \ldots, X'_{j-1}, X'_{j+1}, \ldots, X'_p, \tilde{X}'_1, \ldots, \tilde{X}'_{j-1}\right)$$

By definition of conditional probability.

$$= \mathbb{P}\left(X'_1, \ldots, X'_{j-1}, X_j = b, X'_{j+1}, \ldots, X'_p, \tilde{X}'_1, \ldots, \tilde{X}'_{j-1}\right)$$

We know from step $j-1$ that the pair $\left(X_k, \tilde{X}_k\right)$ is exchangeable in the joint distribution of $\left(X_1, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)$ for all $k \in \{1, \ldots, j-1\}$.

$$= \mathbb{P}\left(X_1, \ldots, X_{j-1}, X_j = b, X_{j+1}, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)$$

By definition of conditional probability.

$$= \frac{\mathbb{P}\left(X_j = b | X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)}{\mathbb{P}\left(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)}$$

Algorithm 1 guarantees that $\mathcal{L}\left(X_j | X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right) = \mathcal{L}\left(\tilde{X}_j | X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)$.

$$= \frac{\mathbb{P}\left(\tilde{X}_j = b | X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)}{\mathbb{P}\left(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)}$$

By definition of conditional probability.

$$= \mathbb{P}\left(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}, \tilde{X}_j = b\right).$$

Because $k \in \{1, \ldots, j-1\}$ was arbitrary but fixed, we proved the claim for all $k \in \{1, \ldots, j-1\}$.

Let $a \in \mathfrak{X}_j, b \in \tilde{\mathfrak{X}}_j$ be arbitrary but fixed.

Note that the joint distribution of $\left(X_1, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}, \tilde{X}_j\right)$ is given by:

$$\begin{aligned}
&\mathbb{P}\left(X_1, \ldots, X_j = a, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}, \tilde{X}_j = b\right) \\
&= \frac{\mathbb{P}\left(X_1, \ldots, X_j = a, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}, \tilde{X}_j = b\right)}{\mathbb{P}\left(X_1, \ldots, X_j = a, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)} \cdot \mathbb{P}\left(X_1, \ldots, X_j = a, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right) \\
&\qquad\qquad \text{By definition of conditional probability} \\
&\qquad\qquad \text{(used in numerator and denominator} \\
&\qquad\qquad \text{of the first term)} \\
&= \frac{\mathbb{P}\left(X_j = a | \tilde{X}_j = b, X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right) \cdot \mathbb{P}\left(\tilde{X}_j = b, X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)}{\mathbb{P}\left(X_j = a | X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right) \cdot \mathbb{P}\left(X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)} \\
&\quad \cdot \mathbb{P}\left(X_1, \ldots, X_j = a, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right) \\
&\qquad\qquad \text{Algorithm 1 guarantees that } \left(\tilde{X}_j \perp\!\!\!\perp X_j\right) \mid \left(X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right). \\
&\qquad\qquad \text{Therefore, by definition of conditional independence} \\
&= \frac{\mathbb{P}\left(X_j = a | X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right) \cdot \mathbb{P}\left(\tilde{X}_j = b, X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)}{\mathbb{P}\left(X_j = a | X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right) \cdot \mathbb{P}\left(X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)} \\
&\quad \cdot \mathbb{P}\left(X_1, \ldots, X_j = a, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right) \\
&= \frac{\mathbb{P}\left(\tilde{X}_j = b, X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)}{\mathbb{P}\left(X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)} \cdot \mathbb{P}\left(X_1, \ldots, X_j = a, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right) \\
&\qquad\qquad \text{In other words} \\
&= \frac{\mathbb{P}\left(\tilde{X}_j = b, X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right) \cdot \mathbb{P}\left(X_j = a, X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)}{\mathbb{P}\left(X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)}. \qquad (22)
\end{aligned}$$

We now show that the pair $\left(X_k, \tilde{X}_k\right)$ is exchangeable in the joint distribution of $\left(X_1, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}, \tilde{X}_j\right)$ for all $k \in \{1, \ldots, j\}$.

Case1 $k = j$:

The denominator of (22) is not affected and remains unchanged.
The numerator changes to $\mathbb{P}\left(X_j = b, X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right) \cdot \mathbb{P}\left(\tilde{X}_j = a, X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)$, but this is the same as we have in (22), because:

$$\begin{aligned}
&\mathbb{P}\left(\tilde{X}_j = b, X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right) \cdot \mathbb{P}\left(X_j = a, X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right) \\
&\qquad\qquad \text{By definition of conditional probability}
\end{aligned}$$

$$= \mathbb{P}\left(\tilde{X}_j = b | X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right) \cdot \mathbb{P}\left(X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)$$
$$\cdot \mathbb{P}\left(X_j = a | X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right) \cdot \mathbb{P}\left(X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)$$

Algorithm 1 guarantees

that $\mathcal{L}\left(X_j | X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right) = \mathcal{L}\left(\tilde{X}_j | X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)$.

$$= \mathbb{P}\left(X_j = b | X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right) \cdot \mathbb{P}\left(X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)$$
$$\cdot \mathbb{P}\left(\tilde{X}_j = a | X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right) \cdot \mathbb{P}\left(X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)$$

By definition of conditional probability

$$= \mathbb{P}\left(X_j = b, X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right) \cdot \mathbb{P}\left(\tilde{X}_j = a, X_{-j}, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right).$$

Therefore, the pair $\left(X_j, \tilde{X}_j\right)$ is exchangeable in the joint distribution of $\left(X_1, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}, \tilde{X}_j\right)$.

Case2 $k < j$:

By Prep1, the denominator of (22) remains unchanged.
By Prep3, the first term of the numerator remains unchanged.
And we know from step $j-1$ that the pair $\left(X_k, \tilde{X}_k\right)$ is exchangeable in the joint distribution of $\left(X_1, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}\right)$ for all $k \in \{1, \ldots, j-1\}$. Therefore, the second term of the numerator in (22) remains unchanged.

So, the pair $\left(X_k, \tilde{X}_k\right)$ is exchangeable in the joint distribution of $\left(X_1, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_{j-1}, \tilde{X}_j\right)$ for all $k \in \{1, \ldots, j-1\}$.

□

Comments:
It is maybe not clear at the first sight, why this proof has to be so complicated and lengthy. It is clear that Algorithm 1, at step j, produces a valid MX-Knockoff $\tilde{X}_j$ for X_j, because $\tilde{X}_j$ and X_j are conditionally independent given all other variables. The problem is that we have to make sure that alle the previous ones are still valid MX-Knockoffs. This is why this proof is not completed by just using an i.i.d. argument as we did in the base case.

If $X = (X_1, \ldots, X_p)$ is a Markov chain, Sesia et al. (2019b) show that Algorithm 1 can be implemented very efficiently. The same paper introduces algorithms to construct MX-Knockoffs if $X = (X_1, \ldots, X_p)$ is a hidden Markov model. Further discussions about MX-Knockoffs for hidden Markov models in genome-wide association studies are presented in Sesia et al. (2019c), extensions and applications using genetic data are given in Sesia et al. (2019a).
Gimenez et al. (2018) provides procedures to efficiently sample valid MX-Knockoffs from Bayesian Networks.

Bates et al. (2019) prove a more general version of Algorithm 1 and introduce a Metropolis-Hastings formulation of an exact MX-Knockoff sampler.

3.4.1.2 Special case - Gaussian rows

The exact way of constructing MX-Knockoffs is to consider the whole joint distribution of $(X, \tilde{X})$ and asking that $\left(X, \tilde{X}\right)_{\text{swap}(S)} \stackrel{d}{=} \left(X, \tilde{X}\right)$ for any subset S. The next example shows that this is satisfied if X is multivariate Gaussian. The reason is that in the Gaussian case, the exchangeability property reduces to matching first and second moments, i.e., mean and covariance matrix. We can always subtract the mean to get centered random vectors on both sides which have mean zero. Therefore, it suffices to focus on the second moments. This is the reason why we used a random vector X with mean $\mathbf{0}$ in the next example.

Example 3.20. Let X be a random row vector with $X \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. Then
i) A joint distribution obeying (13) is

$$\left(X, \tilde{X}\right) \sim \mathcal{N}(\mathbf{0}, \mathbf{G}), \quad \text{where} \quad \mathbf{G} = \begin{bmatrix} \mathbf{\Sigma} & \mathbf{\Sigma} - \text{diag}(\boldsymbol{s}) \\ \mathbf{\Sigma} - \text{diag}(\boldsymbol{s}) & \mathbf{\Sigma} \end{bmatrix}$$

and where $\text{diag}(\boldsymbol{s})$ is any diagonal matrix selected in such a way that the joint covariance matrix $\mathbf{G}$ is positive semidefinite.
ii) We can use the same equi-correlated construction and SDP construction as in Example 2.15 to construct $\boldsymbol{s}$.
iii) We can sample MX-Knockoffs $\tilde{X}$ from the conditional distribution

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

where $\boldsymbol{\mu} = X - X \cdot \mathbf{\Sigma}^{-1} \cdot \text{diag}(\boldsymbol{s})$ and $\mathbf{V} = 2 \cdot \text{diag}(\boldsymbol{s}) - \text{diag}(\boldsymbol{s}) \cdot \mathbf{\Sigma}^{-1} \cdot \text{diag}(\boldsymbol{s})$.

Proof:

ad i) wts $\left(X, \tilde{X}\right) \sim \mathcal{N}(\mathbf{0}, \mathbf{G})$ is obeying (13):

We have to show that $\left(X, \tilde{X}\right)_{\text{swap}(S)} \sim \mathcal{N}(\mathbf{0}, \mathbf{G})$ for any $S \subseteq \{1, \ldots, p\}$.

Note that by definition of a covariance matrix:

$$\mathbf{G} = \begin{pmatrix} \mathbb{V}ar(X_1) & \mathbb{C}ov(X_1, X_2) & \cdots & \mathbb{C}ov(X_1, X_p) & \mathbb{C}ov(X_1, \tilde{X}_1) & \mathbb{C}ov(X_1, \tilde{X}_2) & \cdots & \mathbb{C}ov(X_1, \tilde{X}_p) \\ \mathbb{C}ov(X_2, X_1) & \mathbb{V}ar(X_2) & & \mathbb{C}ov(X_2, X_p) & \mathbb{C}ov(X_2, \tilde{X}_1) & \mathbb{C}ov(X_2, \tilde{X}_2) & & \mathbb{C}ov(X_2, \tilde{X}_p) \\ \vdots & \vdots & \ddots & \vdots & & & \ddots & \vdots \\ \mathbb{C}ov(X_p, X_1) & \mathbb{C}ov(X_p, X_2) & \cdots & \mathbb{V}ar(X_p) & \mathbb{C}ov(X_p, \tilde{X}_1) & \mathbb{C}ov(X_p, \tilde{X}_2) & & \mathbb{C}ov(X_p, \tilde{X}_p) \\ \mathbb{C}ov(\tilde{X}_1, X_1) & \mathbb{C}ov(\tilde{X}_1, X_2) & \cdots & \mathbb{C}ov(\tilde{X}_1, X_p) & \mathbb{V}ar(\tilde{X}_1) & \mathbb{C}ov(\tilde{X}_1, \tilde{X}_2) & \cdots & \mathbb{C}ov(\tilde{X}_1, \tilde{X}_p) \\ \mathbb{C}ov(\tilde{X}_2, X_1) & \mathbb{C}ov(\tilde{X}_2, X_2) & & \mathbb{C}ov(\tilde{X}_2, X_p) & \mathbb{C}ov(\tilde{X}_2, \tilde{X}_1) & \mathbb{V}ar(\tilde{X}_2) & & \mathbb{C}ov(\tilde{X}_2, \tilde{X}_p) \\ \vdots & & \ddots & & \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}ov(\tilde{X}_p, X_1) & \mathbb{C}ov(\tilde{X}_p, X_2) & \cdots & \mathbb{C}ov(\tilde{X}_p, X_p) & \mathbb{C}ov(\tilde{X}_p, \tilde{X}_1) & \mathbb{C}ov(\tilde{X}_p, \tilde{X}_2) & \cdots & \mathbb{V}ar(\tilde{X}_p) \end{pmatrix} \tag{23}$$

Swapping MX-Knockoff variables and original variables in the row-vector $\left(X, \tilde{X}\right)$ changes the corresponding positions of entries in the mean-vector (we can ignore that because we have mean $\mathbf{0}$) and the corresponding positions of the entries in the covariance matrix $\mathbf{G}$.
Changing these entries in $\mathbf{G}$ can be done by using a $2 \cdot p \times 2 \cdot p$ permutation matrix $\mathbf{P}$. The swapped version of $\mathbf{G}$ is then given by $\mathbf{P} \cdot \mathbf{G} \cdot \mathbf{P}$.

By comparing the representation of $\mathbf{G}$ in the statement with (23), we can observe:

a) $\tilde{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$
b) $\mathbb{C}ov\left((X, \tilde{X})\right) = \mathbb{C}ov\left((\tilde{X}, X)\right)$
c) $\mathbb{C}ov(X_i, \tilde{X}_j) = (\mathbf{\Sigma} - \text{diag}(\boldsymbol{s}))_{i,j}$
d) $\mathbb{C}ov(\tilde{X}_i, X_j) = (\mathbf{\Sigma} - \text{diag}(\boldsymbol{s}))_{i,j}$
where $(\cdot)_{i,j}$ denotes the entry in row i and column j of a matrix.

Part a) implies that

$$\mathbb{C}ov(X_i, X_j) = \mathbb{C}ov(\tilde{X}_i, \tilde{X}_j) \quad \text{for } i, j \in \{1, \ldots, p\}. \tag{24}$$

Part b) implies that

$$\mathbb{C}ov(\tilde{X}_i, X_j) = \mathbb{C}ov(X_i, \tilde{X}_j) \quad \text{for } i, j \in \{1, \ldots, p\}. \tag{25}$$

Part c) implies that

$$\mathbb{C}ov(X_i, \tilde{X}_j) = \mathbb{C}ov(X_i, X_j) \quad \text{for } i, j \in \{1, \ldots, p\},\ i \neq j. \tag{26}$$

Part d) implies that

$$\mathbb{C}ov(\tilde{X}_i, X_j) = \mathbb{C}ov(X_i, X_j) \quad \text{for } i, j \in \{1, \ldots, p\},\ i \neq j. \tag{27}$$

This means that we can "revert" all changes which have occured in $\mathbf{P} \cdot \mathbf{G} \cdot \mathbf{P}$ and can write $\mathbf{P} \cdot \mathbf{G} \cdot \mathbf{P} = \mathbf{G}$. Thus, since $\mathbf{P} \cdot \mathbf{G} \cdot \mathbf{P} = \mathbf{G}$ for any swapping operation, the distribution is invariant.
In other words, $\left(X, \tilde{X}\right)_{\text{swap}(S)} \sim \mathcal{N}(\mathbf{0}, \mathbf{G})$ for any $S \subseteq \{1, \ldots, p\}$.

To visualize above argument, let $p = 3$ and $S = \{1\}$.

We start with the joint distribution

$$\left(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3\right) \sim \mathcal{N}(\mathbf{0}, \mathbf{G}),$$

where

$$\mathbf{G} = \begin{pmatrix} \mathbb{V}ar(X_1) & \mathbb{C}ov(X_1, X_2) & \mathbb{C}ov(X_1, X_3) & \mathbb{C}ov(X_1, \tilde{X}_1) & \mathbb{C}ov(X_1, \tilde{X}_2) & \mathbb{C}ov(X_1, \tilde{X}_3) \\ \mathbb{C}ov(X_2, X_1) & \mathbb{V}ar(X_2) & \mathbb{C}ov(X_2, X_3) & \mathbb{C}ov(X_2, \tilde{X}_1) & \mathbb{C}ov(X_2, \tilde{X}_2) & \mathbb{C}ov(X_2, \tilde{X}_3) \\ \mathbb{C}ov(X_3, X_1) & \mathbb{C}ov(X_3, X_2) & \mathbb{V}ar(X_3) & \mathbb{C}ov(X_3, \tilde{X}_1) & \mathbb{C}ov(X_3, \tilde{X}_2) & \mathbb{C}ov(X_3, \tilde{X}_3) \\ \mathbb{C}ov(\tilde{X}_1, X_1) & \mathbb{C}ov(\tilde{X}_1, X_2) & \mathbb{C}ov(\tilde{X}_1, X_3) & \mathbb{V}ar(\tilde{X}_1) & \mathbb{C}ov(\tilde{X}_1, \tilde{X}_2) & \mathbb{C}ov(\tilde{X}_1, \tilde{X}_3) \\ \mathbb{C}ov(\tilde{X}_2, X_1) & \mathbb{C}ov(\tilde{X}_2, X_2) & \mathbb{C}ov(\tilde{X}_2, X_3) & \mathbb{C}ov(\tilde{X}_2, \tilde{X}_1) & \mathbb{V}ar(\tilde{X}_2) & \mathbb{C}ov(\tilde{X}_2, \tilde{X}_3) \\ \mathbb{C}ov(\tilde{X}_3, X_1) & \mathbb{C}ov(\tilde{X}_3, X_2) & \mathbb{C}ov(\tilde{X}_3, X_3) & \mathbb{C}ov(\tilde{X}_3, \tilde{X}_1) & \mathbb{C}ov(\tilde{X}_3, \tilde{X}_2) & \mathbb{V}ar(\tilde{X}_3) \end{pmatrix}.$$

Now, for $S = \{1\}$, we have

$$\left(\tilde{X}_1, X_2, X_3, X_1, \tilde{X}_2, \tilde{X}_3\right) \sim \mathcal{N}(\mathbf{0}, \mathbf{G}^*),$$

where

$$\mathbf{G}^* = \begin{pmatrix} \mathbb{V}ar(\tilde{X}_1) & \mathbb{C}ov(\tilde{X}_1, X_2) & \mathbb{C}ov(\tilde{X}_1, X_3) & \mathbb{C}ov(\tilde{X}_1, X_1) & \mathbb{C}ov(\tilde{X}_1, \tilde{X}_2) & \mathbb{C}ov(\tilde{X}_1, \tilde{X}_3) \\ \mathbb{C}ov(X_2, \tilde{X}_1) & \mathbb{V}ar(X_2) & \mathbb{C}ov(X_2, X_3) & \mathbb{C}ov(X_2, X_1) & \mathbb{C}ov(X_2, \tilde{X}_2) & \mathbb{C}ov(X_2, \tilde{X}_3) \\ \mathbb{C}ov(X_3, \tilde{X}_1) & \mathbb{C}ov(X_3, X_2) & \mathbb{V}ar(X_3) & \mathbb{C}ov(X_3, X_1) & \mathbb{C}ov(X_3, \tilde{X}_2) & \mathbb{C}ov(X_3, \tilde{X}_3) \\ \mathbb{C}ov(X_1, \tilde{X}_1) & \mathbb{C}ov(X_1, X_2) & \mathbb{C}ov(X_1, X_3) & \mathbb{V}ar(X_1) & \mathbb{C}ov(X_1, \tilde{X}_2) & \mathbb{C}ov(X_1, \tilde{X}_3) \\ \mathbb{C}ov(\tilde{X}_2, \tilde{X}_1) & \mathbb{C}ov(\tilde{X}_2, X_2) & \mathbb{C}ov(\tilde{X}_2, X_3) & \mathbb{C}ov(\tilde{X}_2, X_1) & \mathbb{V}ar(\tilde{X}_2) & \mathbb{C}ov(\tilde{X}_2, \tilde{X}_3) \\ \mathbb{C}ov(\tilde{X}_3, \tilde{X}_1) & \mathbb{C}ov(\tilde{X}_3, X_2) & \mathbb{C}ov(\tilde{X}_3, X_3) & \mathbb{C}ov(\tilde{X}_3, X_1) & \mathbb{C}ov(\tilde{X}_3, \tilde{X}_2) & \mathbb{V}ar(\tilde{X}_3) \end{pmatrix}.$$

Note that $\mathbf{G}^* = \mathbf{P} \cdot \mathbf{G} \cdot \mathbf{P}$ with

$$\mathbf{P} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Our claim above is that we can use (24), (25), (26) and (27) to turn $\mathbf{G}^*$ into $\mathbf{G}$ again.

E.g. the top left block

$$\begin{pmatrix} \mathbb{V}ar(\tilde{X}_1) & \mathbb{C}ov(\tilde{X}_1, X_2) & \mathbb{C}ov(\tilde{X}_1, X_3) \\ \mathbb{C}ov(X_2, \tilde{X}_1) & \mathbb{V}ar(X_2) & \mathbb{C}ov(X_2, X_3) \\ \mathbb{C}ov(X_3, \tilde{X}_1) & \mathbb{C}ov(X_3, X_2) & \mathbb{V}ar(X_3) \end{pmatrix}$$

should be equal to

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbb{V}ar(X_1) & \mathbb{C}ov(X_1, X_2) & \mathbb{C}ov(X_1, X_3) \\ \mathbb{C}ov(X_2, X_1) & \mathbb{V}ar(X_2) & \mathbb{C}ov(X_2, X_3) \\ \mathbb{C}ov(X_3, X_1) & \mathbb{C}ov(X_3, X_2) & \mathbb{V}ar(X_3) \end{pmatrix}.$$

By using (24) for $i = j = 1$, we can replace $\mathbb{V}ar(\tilde{X}_1)$ by $\mathbb{V}ar(X_1)$.
By using (26), we can replace $\mathbb{C}ov(X_2, \tilde{X}_1)$ and $\mathbb{C}ov(X_3, \tilde{X}_1)$ by $\mathbb{C}ov(X_2, X_1)$ and $\mathbb{C}ov(X_3, X_1)$.
By using (27), we can replace $\mathbb{C}ov(\tilde{X}_1, X_2)$ and $\mathbb{C}ov(\tilde{X}_1, X_3)$ by $\mathbb{C}ov(X_1, X_2)$ and $\mathbb{C}ov(X_1, X_3)$.

Entries of the other blocks can be changed analogously.

ad ii):

Note that

$$\mathbf{G} = \begin{bmatrix} \mathbf{\Sigma} & \mathbf{\Sigma} - \text{diag}(\boldsymbol{s}) \\ \mathbf{\Sigma} - \text{diag}(\boldsymbol{s}) & \mathbf{\Sigma} \end{bmatrix}$$

is equal to $\mathbf{G}$ in Lemma 2.14 (in the fixed-X setting) where we proved that this matrix is positive semidefinite if and only if $\text{diag}(\boldsymbol{s}) \succeq \mathbf{0}$ and $2 \cdot \mathbf{\Sigma} - \text{diag}(\boldsymbol{s}) \succeq \mathbf{0}$ (note that we did not use $\mathbf{\Sigma} = \mathbf{X}^T \cdot \mathbf{X}$ in the proof of Lemma 2.14). This led to two possible construction methods for $\boldsymbol{s}$, the equi-correlated construction and SDP construction (given in Example 2.15).

ad iii) wts $\tilde{X}|X \overset{d}{=} \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$:

By Lemma B.4, we know that if the joint distribution $\left(X, \tilde{X}\right)$ is Gaussian, the conditional distribution $\tilde{X}|X$ is Gaussian too.

Furthermore,

$$\mathbb{E}(\tilde{X}|X)$$

Note that X and $\tilde{X}$ are row vectors.
Therefore, by Lemma B.5 i)

$$
\begin{aligned}
&= \mathbf{0} + (X - \mathbf{0}) \cdot \boldsymbol{\Sigma}^{-1} \cdot (\boldsymbol{\Sigma} - \operatorname{diag}(\boldsymbol{s})) \\
&= X \cdot \underbrace{\boldsymbol{\Sigma}^{-1} \cdot \boldsymbol{\Sigma}}_{=\mathbf{I}} - X \cdot \boldsymbol{\Sigma}^{-1} \cdot \operatorname{diag}(\boldsymbol{s}) \\
&= X - X \cdot \boldsymbol{\Sigma}^{-1} \cdot \operatorname{diag}(\boldsymbol{s})
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{C}ov(\tilde{X}|X) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{Note that } X \text{ and } \tilde{X} \text{ are row vectors.} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{Therefore, by Lemma B.5 ii)} \\
&= \boldsymbol{\Sigma} - (\boldsymbol{\Sigma} - \operatorname{diag}(\boldsymbol{s})) \cdot \boldsymbol{\Sigma}^{-1} \cdot (\boldsymbol{\Sigma} - \operatorname{diag}(\boldsymbol{s})) \\
&= \boldsymbol{\Sigma} - \left(\underbrace{\boldsymbol{\Sigma} \cdot \boldsymbol{\Sigma}^{-1}}_{=\mathbf{I}} - \operatorname{diag}(\boldsymbol{s}) \cdot \boldsymbol{\Sigma}^{-1} \right) \cdot (\boldsymbol{\Sigma} - \operatorname{diag}(\boldsymbol{s})) \\
&= \boldsymbol{\Sigma} - \left(\mathbf{I} \cdot \boldsymbol{\Sigma} - \mathbf{I} \cdot \operatorname{diag}(\boldsymbol{s}) - \operatorname{diag}(\boldsymbol{s}) \cdot \underbrace{\boldsymbol{\Sigma}^{-1} \cdot \boldsymbol{\Sigma}}_{=\mathbf{I}} + \operatorname{diag}(\boldsymbol{s}) \cdot \boldsymbol{\Sigma}^{-1} \cdot \operatorname{diag}(\boldsymbol{s}) \right) \\
&= \boldsymbol{\Sigma} - \left(\boldsymbol{\Sigma} - 2 \cdot \operatorname{diag}(\boldsymbol{s}) + \operatorname{diag}(\boldsymbol{s}) \cdot \boldsymbol{\Sigma}^{-1} \cdot \operatorname{diag}(\boldsymbol{s}) \right) \\
&= \boldsymbol{\Sigma} - \boldsymbol{\Sigma} + 2 \cdot \operatorname{diag}(\boldsymbol{s}) - \operatorname{diag}(\boldsymbol{s}) \cdot \boldsymbol{\Sigma}^{-1} \cdot \operatorname{diag}(\boldsymbol{s}) \\
&= 2 \cdot \operatorname{diag}(\boldsymbol{s}) - \operatorname{diag}(\boldsymbol{s}) \cdot \boldsymbol{\Sigma}^{-1} \cdot \operatorname{diag}(\boldsymbol{s}).
\end{aligned}
$$

Therefore

$$
\tilde{X}|X \overset{d}{=} \mathcal{N}\left(X - X \cdot \boldsymbol{\Sigma}^{-1} \cdot \operatorname{diag}(\boldsymbol{s}),\, 2 \cdot \operatorname{diag}(\boldsymbol{s}) - \operatorname{diag}(\boldsymbol{s}) \cdot \boldsymbol{\Sigma}^{-1} \cdot \operatorname{diag}(\boldsymbol{s})\right).
$$

In other words, having observed X, we can sample $\tilde{X}$ by creating random numbers which satisfy

$$
\tilde{X} \sim \mathcal{N}\left(X - X \cdot \boldsymbol{\Sigma}^{-1} \cdot \operatorname{diag}(\boldsymbol{s}),\, 2 \cdot \operatorname{diag}(\boldsymbol{s}) - \operatorname{diag}(\boldsymbol{s}) \cdot \boldsymbol{\Sigma}^{-1} \cdot \operatorname{diag}(\boldsymbol{s})\right).
$$

□

Comments:
a) Note that Definition 3.6 ii) is satisfied automatically, because $\tilde{X}$ are constructed without looking at Y. Therefore, $\tilde{X}$ (as constructed in iii)) are valid MX-Knockoffs.
b) This example shows a connection (actually a difference) to the fixed-X setting, where $\mathbf{X}$ is non-random (fixed) and a FX-Knockoff matrix $\tilde{\mathbf{X}}$ is constructed such that

$$
\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right] = \begin{bmatrix} \mathbf{X}^T \cdot \mathbf{X} & \mathbf{X}^T \cdot \tilde{\mathbf{X}} \\ \tilde{\mathbf{X}}^T \cdot \mathbf{X} & \tilde{\mathbf{X}}^T \cdot \tilde{\mathbf{X}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \operatorname{diag}(\boldsymbol{s}) \\ \boldsymbol{\Sigma} - \operatorname{diag}(\boldsymbol{s}) & \boldsymbol{\Sigma} \end{bmatrix} =: \mathbf{G}^{\mathrm{FX}} \tag{28}
$$

for $\boldsymbol{\Sigma} = \mathbf{X}^T \cdot \mathbf{X}$.
Note that a sample covariance matrix $\mathbf{Q}$ is a $p \times p$ matrix with entries $q_{jk} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_{ij} - \bar{\mathbf{X}}_j) \cdot (X_{ik} - \bar{\mathbf{X}}_k)$. (Usually, lower case letters are used for realizations, but we continue with upper case letters as in the whole thesis). For centered columns $\mathbf{X}_j$ of $\mathbf{X}$ (which means that the sample mean $\bar{\mathbf{X}}_j = \mathbf{0}$), we have $q_{jk} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} X_{ij} \cdot X_{ik} = \frac{1}{n-1} \cdot \mathbf{X}_j^T \cdot \mathbf{X}_k$. This means that if we ignore the scaling factor, we can think of $\mathbf{X}^T \cdot \mathbf{X}$ as a sample covariance matrix.
For the same reason, we can think of

$$
\left[\mathbf{X}, \tilde{\mathbf{X}}\right]^T \cdot \left[\mathbf{X}, \tilde{\mathbf{X}}\right]
$$

as sample covariance matrix (of the joint set of variables). And in this fixed-X setting, we are asking (see Lemma 2.9) that swapping columns in $\left[\mathbf{X}, \tilde{\mathbf{X}}\right]$ does not change the *sample* covariance matrix $\mathbf{G}^{\mathrm{FX}}$.

Whereas in the comparable model-X setting (see Example 3.20 i)), we are asking that swapping columns in $\left(X, \tilde{X}\right)$ does not change the *population* covariance matrix $\mathbf{G}$.

Example 3.21. Let the design matrix $\mathbf{X}$ have rows X with $X \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ (all rows follow the same distribution). A way of calculating a MX-Knockoff matrix $\tilde{\mathbf{X}}$ is:
1) Choose a procedure (e.g. equi or SDP) to build $\boldsymbol{s}$ (by using the given matrix $\mathbf{\Sigma}$).
2) Calculate $\mathbf{\Sigma}^{-1}$.
3) For every row X of $\mathbf{X}$, sample $\tilde{X} \sim \mathcal{N}\left(X - X \cdot \mathbf{\Sigma}^{-1} \cdot \text{diag}(\boldsymbol{s}),\, 2 \cdot \text{diag}(\boldsymbol{s}) - \text{diag}(\boldsymbol{s}) \cdot \mathbf{\Sigma}^{-1} \cdot \text{diag}(\boldsymbol{s})\right)$. Row-wise concatenation of all such $\tilde{X}$ gives $\tilde{\mathbf{X}}$.

3.4.2 Approximate construction of MX Knockoffs

Barber et al. (2018) claim that the MX-Knockoff framework is robust to errors in the underlying assumptions on the distribution of $\mathbf{X}$, which means that MX-Knockoff procedures can be applied in situations, where the distribution of $\mathbf{X}$ is estimated accurately but not known exactly. And Huang and Janson (2019) show several examples of constructing MX-Knockoff matrices under weaker assumptions (where the distribution of $\mathbf{X}$ is only known up to a parameterized family of distributions). One very simple way of constructing approximate MX-Knockoffs is to just consider the first two moments:

3.4.2.1 Second-order MX-Knockoffs

We can use the idea of the Gaussian special case to create an approximation-procedure for producing MX-Knockoffs in settings where the distribution of X is arbitrary (but of course, known). Instead of considering the whole joint distribution of $(X, \tilde{X})$ and asking that $\left(X, \tilde{X}\right)_{\text{swap}(S)} \stackrel{d}{=} \left(X, \tilde{X}\right)$ for any subset S, we can ask that $\left(X, \tilde{X}\right)_{\text{swap}(S)}$ and $\left(X, \tilde{X}\right)$ have the same first two moments, i.e., the same mean and covariance matrix.

By a similar calculation as in Example 3.20 i) (we did not use any property of the normal distribution), we can show that $\left(X, \tilde{X}\right)_{\text{swap}(S)}$ and $\left(X, \tilde{X}\right)$ have the same mean and covariance matrix if and only if

$$\mathbb{C}ov\left(\left(X, \tilde{X}\right)\right) = \begin{bmatrix} \mathbf{\Sigma} & \mathbf{\Sigma} - \text{diag}(\boldsymbol{s}) \\ \mathbf{\Sigma} - \text{diag}(\boldsymbol{s}) & \mathbf{\Sigma} \end{bmatrix},$$

where $\mathbb{C}ov(X) = \mathbf{\Sigma}$.

Romano et al. (2018) are extending these ideas to higher orders and arbitrary distributions of $\mathbf{X}$ by using deep generative models.

Per definition, covariance matrices have to be positive semidefinite. So, we could use the same methods to derive $\boldsymbol{s}$ as in Example 3.20 ii). But for large p, Candes et al. (2018) discovered difficulties of these two construction methods. Equi-correlated method is computationally easy, but produces small entries of $\boldsymbol{s}$ for high p which leads to poor power. And SDP seems to be computationally expensive for high p. To address these difficulties, authors proposed a two-step procedure which they call approximate semidefinite program (ASDP).

3.4.3 Construction of MX-Knockoff feature statistics

As noted earlier, assumptions for MX-Knockoff feature statistics are the same as for FX-Knockoff feature statistics, except for the sufficiency property, which is only required for FX-Knockoff feature statistics. This means that the MX-Knockoff-framework allows for a wider variety of feature statistics W_j to be used than in the fixed-X framework (and of course, Examples given in subsection 2.4.3 can be used in the MX-Setting too). The flip-sign condition is the only property MX-Knockoff feature statistics have to satisfy (see Definition 3.8). If Z_j and Z_{j+p} are importance measures of $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$, respectively, then the flip-sign condition is satisfied by setting $W_j = f_j\left(Z_j, Z_{j+p}\right)$, where f_j is any anti-symmetric function (which means that $f(a,b) = -f(b,a)$).

Example 3.22. There are many possible anti-symmetric functions f_j for importance measures Z_j and Z_{j+p}:
i) $f_j(Z_j, Z_{j+p}) = |Z_j| - |Z_{j+p}|$,
ii) $f_j(Z_j, Z_{j+p}) = (|Z_j| \vee |Z_{j+p}|) \cdot \text{sign}\left(|Z_j| - |Z_{j+p}|\right)$,
iii) $f_j(Z_j, Z_{j+p}) = \log\left(|Z_j|\right) - \log\left(|Z_{j+p}|\right)$.

This means that we can create MX-Knockoff feature statistics in two steps. First, importance measures $(Z_1, \ldots, Z_p, Z_{1+p}, \ldots, Z_{p+p})$ are created for $\left(\mathbf{X}_1, \ldots, \mathbf{X}_p, \tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_p\right)$. Second, for every $j \in \{1, \ldots, p\}$, W_j are constructed by setting $W_j = f_j\left(Z_j, Z_{j+p}\right)$ for some anti-symmetric function f_j.

Example 3.23. For the linear model $\mathbf{Y} = \left[\mathbf{X}, \tilde{\mathbf{X}}\right] \cdot \boldsymbol{\beta} + \boldsymbol{z}$ (where $\boldsymbol{z} \sim \mathcal{N}\left(0, \sigma^2 \cdot \mathbf{I}_n\right)$) we use Lasso to estimate $\boldsymbol{\beta}$ and get the solution $\hat{\boldsymbol{\beta}}(\lambda) = \left(\hat{\beta}_1(\lambda), \ldots, \hat{\beta}_p(\lambda), \hat{\beta}_{1+p}(\lambda), \ldots, \hat{\beta}_{p+p}(\lambda)\right)$. We define $Z_j = |\hat{\beta}_j(\lambda)|$ for $j \in \{1, \ldots, 2 \cdot p\}$ and end up with a $2 \cdot p$-dimensional vector $(Z_1, \ldots, Z_p, Z_{1+p}, \ldots, Z_{p+p})$. Then the following statistics are valid MX-Knockoff feature statistics:
i) Lasso Signed Max (LSM) statistics $W_j = (Z_j \vee Z_{j+p}) \cdot \text{sign}(Z_j - Z_{j+p})$ for $j \in \{1, \ldots, p\}$.
ii) Lasso coefficient-difference (LCD) statistics $W_j = Z_j - Z_{j+p}$ for $j \in \{1, \ldots, p\}$.

Proof:

ad i):

If variables $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ are swapped,

$$\begin{aligned}
&w_j\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right]_{\text{swap}(\{j\})},\mathbf{Y}\right)\\
&=(Z_{j+p}\vee Z_j)\cdot\text{sign}(Z_{j+p}-Z_j)\\
&=\left(|\hat{\beta}_{j+p}(\lambda)|\vee|\hat{\beta}_j(\lambda)|\right)\cdot\text{sign}\left(|\hat{\beta}_{j+p}(\lambda)|-|\hat{\beta}_j(\lambda)|\right)\\
&=\left(|\hat{\beta}_j(\lambda)|\vee|\hat{\beta}_{j+p}(\lambda)|\right)\cdot\text{sign}\left((-1)\cdot\left(|\hat{\beta}_j(\lambda)|-|\hat{\beta}_{j+p}(\lambda)|\right)\right)\\
&=(-1)\cdot\left(|\hat{\beta}_j(\lambda)|\vee|\hat{\beta}_{j+p}(\lambda)|\right)\cdot\text{sign}\left(|\hat{\beta}_j(\lambda)|-|\hat{\beta}_{j+p}(\lambda)|\right)\\
&=(-1)\cdot(Z_j\vee Z_{j+p})\cdot\text{sign}\left(Z_j-Z_{j+p}\right)\\
&=-w_j\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right],\mathbf{Y}\right).
\end{aligned}$$

If other variables are swapped, w_j is unaffected since Lasso-solutions $\hat{\beta}_j(\lambda)$ and $\hat{\beta}_{j+p}(\lambda)$ remain the same as long as entries $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ are not swapped.

ad ii):

If variables $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ are swapped,

$$\begin{aligned}
&w_j\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right]_{\text{swap}(\{j\})},\mathbf{Y}\right)\\
&=Z_{j+p}-Z_j\\
&=|\hat{\beta}_{j+p}(\lambda)|-|\hat{\beta}_j(\lambda)|\\
&=-\left(|\hat{\beta}_j(\lambda)|-|\hat{\beta}_{j+p}(\lambda)|\right)\\
&=-\left(Z_j-Z_{j+p}\right)\\
&=-w_j\left(\left[\mathbf{X},\tilde{\mathbf{X}}\right],\mathbf{Y}\right).
\end{aligned}$$

If other variables are swapped, w_j is unaffected since Lasso-solutions $\hat{\beta}_j(\lambda)$ and $\hat{\beta}_{j+p}(\lambda)$ remain the same as long as entries $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ are not swapped.

□

Comments:
a) Note that the value of the penalty λ does not need to be fixed in advance (as in Example (2.20) in the fixed-X setting). It can be selected by cross-validation of $\mathbf{Y}$ on $\left[\mathbf{X},\tilde{\mathbf{X}}\right]$. Then, w_j still obeys the flip-sign property and valid MX-Knockoff feature statistics are produced.
b) In fact, essentially any data-fitting or prediction algorithm can be used to get importance measures Z_j as long as the procedures are applied to the design matrix $\left[\mathbf{X},\tilde{\mathbf{X}}\right]$. Separately using procedures on $\mathbf{X}$ and $\tilde{\mathbf{X}}$ to produce importance measures Z_j would probably violate the flip-sign property of w_j.
c) We could even run several algorithms and then choose importance measures derived from whichever has the lowest crossvalidation error.
d) The reason why we can use such complex procedures is that, in the model-X setting, w_j does not have to obey a sufficiency property. It only has to obey a flip-sign property which is easy to check.
e) Across a wide range of simulations, Candes et al. (2018) found the LCD statistic to be uniformly more powerful than the LSM statistic, particularly under covariate dependence. This leads to the question which MX-Knockoff features statistics should be used for which situation to get high power. Gimenez et al. (2018) compare the power of several MX-Knockoff feature statistics and propose a stable and more powerful importance statistic.

References

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

Barber, R. F. and Candes, E. J. (2018). On the construction of knockoffs in case-control studies. *arXiv preprint arXiv:1812.11433*.

Barber, R. F. and Candes, E. J. (2019). A knockoff filter for high-dimensional selective inference. *The Annals of Statistics*, 47(5):2504–2537.

Barber, R. F., Candes, E. J., and Samworth, R. J. (2018). Robust inference with knockoffs. *arXiv preprint arXiv:1801.03896*.

Bates, S., Candes, E. J., Janson, L., and Wang, W. (2019). Metropolized knockoff sampling. *arXiv preprint arXiv:1903.00434*.

Bekker, P. A. (1988). The positive semidefiniteness of partitioned matrices. *Linear Algebra and its Applications*, 111:261–278.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer Science+Business Media, LLC.

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

Dai, R. and Barber, R. F. (2016). The knockoff filter for fdr control in group-sparse and multitask regression. *arXiv preprint arXiv:1602.03589*.

Eaton, M. L. (2007). *Multivariate Statistics - A Vector Space Approach*. Institute of Mathematical Statistics.

Edwards, D. (2000). *Introduction to Graphical Modelling*. Springer-Verlag New York Inc., second edition.

Fan, Y., Demirkaya, E., Li, G., and Lv, J. (2017). Rank: Large-scale inference with graphical nonlinear knockoffs. *arXiv preprint arXiv:1709.00092*.

Gimenez, J. R., Ghorbani, A., and Zou, J. (2018). Knockoffs for the mass: new feature importance statistics with false discovery guarantees. *arXiv preprint arXiv:1807.06214*.

Gimenez, J. R. and Zou, J. (2018). Improving the stability of the knockoff procedure: Multiple simultaneous knockoffs and entropy maximization. *arXiv preprint arXiv:1810.11378*.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations*. The Johns Hopkins University Press, third edition.

Horn, R. A. and Johnson, C. R. (2013). *Matrix Analysis*. Cambridge University Press, second edition.

Huang, D. and Janson, L. (2019). Relaxing the assumptions of knockoffs by conditioning. *arXiv preprint arXiv:1903.02806*.

Introduction (2019, September 27). Retrieved from https://www.randomservices.org/random/martingales/Introduction.html.

Janson, L. and Su, W. (2016). Familywise error rate control via knockoffs. *Electronic Journal of Statistics*, 10(1):960–975.

Katsevich, E. and Sabatti, C. (2019). Multilayer knockoff filter: Controlled variable selection at multiple resolutions. *The Annals of Applied Statistics*, 13(1):1–33.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models - Principles and Techniques*. Massachusetts Institute of Technology.

Lehmann, E. L. and Romano, J. P. (2005). Generalizations of the familywise error rate. *The Annals of Statistics*, 33(3):1138–1154.

Neapolitan, R. E. and Jiang, X. (2007). *Probabilistic Methods for Financial and Marketing Informatics*. Elsevier Inc.

Romano, Y., Sesia, M., and Candes, E. J. (2018). Deep knockoffs. *arXiv preprint arXiv:1811.06687*.

Sesia, M., Barber, R. F., Candes, E. J., Janson, L., and Patterson, E. (2018). *knockoff: The Knockoff Filter for Controlled Variable Selection*. R package version 0.3.2.

Sesia, M., Katsevich, E., Bates, S., Candes, E. J., and Sabatti, C. (2019a). Multi-resolution localization of causal variants across the genome. *bioRxiv preprint 631390*.

Sesia, M., Sabatti, C., and Candes, E. J. (2019b). Gene hunting with knockoffs for hidden markov models. *Biometrika*, 106(1):1–18.

Sesia, M., Sabatti, C., and Candes, E. J. (2019c). Rejoinder: "gene hunting with hidden markov model knockoffs". *Biometrika*, 106(1):35–45.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tong, Y. L. (1990). *The Multivariate Normal Distribution*. Springer-Verlag New York Inc.

Weinstein, A., Barber, R. F., and Candes, E. J. (2017). A power and prediction analysis for knockoffs with lasso statistics. *arXiv preprint arXiv:1712.06465*.

Williams, D. (1991). *Probability with Martingales*. Cambridge University Press.

Zhang, F. (2005). *The Schur complement and its applications*. Springer Science+Business Media, Inc.

A Appendix

Definition A.1. Let $x = (x_1, \ldots, x_p)$ be a row vector containing real numbers. The *signum function* of that vector is defined as

$$\operatorname{sign}(x) = \begin{pmatrix} \operatorname{sign}(x_1) & \cdots & \operatorname{sign}(x_p) \end{pmatrix},$$

where

$$\operatorname{sign}(x_j) = \begin{cases} -1, & x_j < 0 \\ 0, & x_j = 0 \\ +1, & x_j > 0 \end{cases}$$

for all $j \in \{1, \ldots, p\}$.

Lemma A.2. *(Theorem 3.3.3 in Tong (1990)). Let $Z \sim \mathcal{N}_n(\mu, \Sigma)$. Let A be a non-random $m \times n$ matrix. Let a be a non-random $m \times 1$ vector. Then*

$$A \cdot Z + a \sim \mathcal{N}_m\left(A \cdot \mu + a,\ A \cdot \Sigma \cdot A^T\right).$$

Definition A.3. (Definitions 1 and 2 of Introduction to martingales). Let $\mathbf{X} = \{X_t : t \in T\}$ be a stochastic process on an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$, having state space $\mathbb{R}$, and where the index set $T \subseteq [0, \infty)$ is representing time. Let $\mathfrak{F} = \{\mathcal{F}_t : t \in T\}$ be a filtration, i.e., an increasing family of sub σ-algebras of $\mathcal{F}$, such that $\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F}$ for $s, t \in T$ with $s \leq t$. Let $\mathbf{X}$ be adapted to $\mathfrak{F}$, i.e., X_t is measurable with respect to $\mathcal{F}_t$ for $t \in T$. Let $\mathbb{E}(|X_t|) < \infty$ for all $t \in T$.
i) $\mathbf{X}$ is called a *martingale* with respect to $\mathfrak{F}$ if $\mathbb{E}(X_t|\mathcal{F}_s) = X_s$ for all $s, t \in T$ with $s \leq t$.
ii) $\mathbf{X}$ is called a *super-martingale* with respect to $\mathfrak{F}$ if $\mathbb{E}(X_t|\mathcal{F}_s) \leq X_s$ for all $s, t \in T$ with $s \leq t$.

For the discrete case, where $T = \mathbb{N}$,
- "$\mathbb{E}(X_t|\mathcal{F}_s) = X_s$ for all $s, t \in T$ with $s \leq t$" reduces to "$\mathbb{E}(X_{n+1}|\mathcal{F}_n) = X_n$ for all $n \in \mathbb{N}$",
- "$\mathbb{E}(X_t|\mathcal{F}_s) \leq X_s$ for all $s, t \in T$ with $s \leq t$" reduces to "$\mathbb{E}(X_{n+1}|\mathcal{F}_n) \leq X_n$ for all $n \in \mathbb{N}$".

Lemma A.4. *(Theorem 10.10 in Williams (1991)). Let T be a bounded stopping time.*
i) Let X be a martingale. Then X_T is integrable and

$$\mathbb{E}(X_T) = \mathbb{E}(X_0).$$

ii) Let X be a supermartingale. Then X_T is integrable and

$$\mathbb{E}(X_T) \leq \mathbb{E}(X_0).$$

Lemma A.5. *(Theorem A.9 in Zhang (2005)) Let*

$$M = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}$$

be a matrix, where A is a positive definite matrix and C is a symmetric matrix. The following are equivalent:
i) M is positive semidefinite,
ii) $C - B^T \cdot A^{-1} \cdot B$ is positive semidefinite.

Lemma A.6. *(Theorem 1 in Bekker (1988)). Let*

$$M = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix},$$

where A and C are symmetric. Then $M \succeq 0$ if and only if
i) $C \succeq 0$,
ii) $B^T = C \cdot C^- \cdot B^T$, and
iii) $A - B \cdot C^- \cdot B^T \succeq 0$, where C^- denotes the Moore-Penrose inverse of C.

Lemma A.7. *(Theorem 2.1.14 in Horn and Johnson (2013)). (QR-factorization). Let A be a $n \times p$ matrix with $n \geq p$. Then:*
i) There is a $n \times p$ matrix Q with orthonormal columns and an upper triangular $p \times p$ matrix R with nonnegative main diagonal entries such that $A = Q \cdot R$.
ii) If the rank of A is p, then the factors Q and R in i) are uniquely determined and the main diagonal entries of R are all positive.

Lemma A.8. *(Theorem 5.2.1 in Golub and Van Loan (1996)). Let A be a $n \times p$ matrix with full column rank and let $A = Q \cdot R$ be a QR-factorization. We write $A = [a_1, \ldots, a_p]$, $Q = [q_1, \ldots, q_n]$ and $R = [r_1, \ldots, r_p]$ to denote columns of these matrices. Then*
i) $\text{span}\{a_1, \ldots, a_k\} = \text{span}\{q_1, \ldots, q_k\}$ for $k = 1, \ldots, p$.
ii) In particular, if $Q_1 = [q_1, \ldots, q_p]$ and $Q_2 = [q_{p+1}, \ldots, q_n]$, then

$$\text{range}(A) = \text{range}(Q_1),$$

$$\text{range}(A)^{\perp} = \text{range}(Q_2)$$

and $A = Q_1 \cdot R_1$ with $R_1 = [r_1, \ldots, r_p]$.

Lemma A.9. *(Observation 7.1.4 in Horn and Johnson (2013)). Each eigenvalue of a positive semidefinite matrix is a nonnegative real number.*

Lemma A.10. *(Corollary 7.2.9 in Horn and Johnson (2013)). (Cholesky factorization). Let A be a real and symmetric $n \times n$ matrix. Then A is positive semidefinite if and only if there is a real and lower triangular $n \times n$ matrix L with nonnegative diagonal entries such that $A = L \cdot L^T$.*

Lemma A.11. *(Exercise after Definition 7.7.1 in Horn and Johnson (2013)). Let A be a real and symmetric $n \times n$ matrix with smallest and largest eigenvalues $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively. Then*

$$\lambda_{\max}(A) \cdot I \succeq A \succeq \lambda_{\min}(A) \cdot I.$$

Lemma A.12. *(Proposition 1.45 in Eaton (2007)). Let A be a real symmetric $n \times n$ matrix. Then there exists an $n \times n$ orthogonal matrix U and a $n \times n$ diagonal matrix Λ such that $A = U \cdot \Lambda \cdot U^T$. The columns of U are the eigenvectors of A and the diagonal elements of Λ, say $\lambda_1, \ldots, \lambda_n$, are eigenvalues of A.*

Lemma A.13. *Let $\mathbf{X}$ and $\mathbf{Y}$ be random elements (variables/vectors/matrices) and let f be a suitable measurable function with codomain $\mathbb{R}^p$ for some $p \geq 1$. Then*
i) $\mathbf{X} \stackrel{d}{=} \mathbf{Y} \implies f(\mathbf{X}) \stackrel{d}{=} f(\mathbf{Y})$
ii) for non-random $\mathbf{A}$*:* $\mathbf{X} \stackrel{d}{=} \mathbf{Y} \implies f(\mathbf{A}, \mathbf{X}) \stackrel{d}{=} f(\mathbf{A}, \mathbf{Y})$

Proof:

ad i) wts $\mathbf{X} \stackrel{d}{=} \mathbf{Y} \implies f(\mathbf{X}) \stackrel{d}{=} f(\mathbf{Y})$:

Let $\mathbf{v} = (v_1, \ldots, v_p)^T \in \mathbb{R}^p$ be arbitrary but fixed.

$$\begin{aligned}
&\mathbb{P}\left(f(\mathbf{X}) \leq \mathbf{v}\right) \\
&= \mathbb{P}\left(f(\mathbf{X}) \in (-\infty, v_1] \times \ldots \times (-\infty, v_p]\right) \\
&= \mathbb{P}\left(\mathbf{X} \in f^{-1}\left((-\infty, v_1] \times \ldots \times (-\infty, v_p]\right)\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{We use } \mathbf{X} \stackrel{d}{=} \mathbf{Y}. \\
&= \mathbb{P}\left(\mathbf{Y} \in f^{-1}\left((-\infty, v_1] \times \ldots \times (-\infty, v_p]\right)\right) \\
&= \mathbb{P}\left(f(\mathbf{Y}) \in (-\infty, v_1] \times \ldots \times (-\infty, v_p]\right) \\
&= \mathbb{P}\left(f(\mathbf{Y}) \leq \mathbf{v}\right).
\end{aligned}$$

ad ii) wts for non-random $\mathbf{A}$: $\mathbf{X} \stackrel{d}{=} \mathbf{Y} \quad \Longrightarrow \quad f(\mathbf{A}, \mathbf{X}) \stackrel{d}{=} f(\mathbf{A}, \mathbf{Y})$:

Define $g(\cdot) := f(\mathbf{A}, \cdot)$.
Since $\mathbf{X} \stackrel{d}{=} \mathbf{Y}$, we can use i) and get $g(\mathbf{X}) \stackrel{d}{=} g(\mathbf{Y})$ which is $f(\mathbf{A}, \mathbf{X}) \stackrel{d}{=} f(\mathbf{A}, \mathbf{Y})$.

□

B Appendix

Definition B.1. (Definitions 3.6 and 3.7 in Neapolitan and Jiang (2007)). Let V be a set of random variables and $X \in V$.
i) A *Markov blanket* S of X is any set of variables such that X is conditionally independent of all other variables given S.
ii) A *Markov boundary* of X is any Markov blanket such that none of its proper subsets is a Markov blanket of X.

Definition B.2. (Definition 2.3 and Proposition 2.2 in Koller and Friedman (2009)). Let A, B and C be three events and let $\mathcal{L}$ be a distribution. We say A is *conditionally independent* of B given C (denoted $(A \perp\!\!\!\perp B)|C$), if

$$\mathcal{L}(A|B \cap C) = \mathcal{L}(A|C)$$

or if

$$\mathcal{L}(B \cap C) = 0.$$

Alternatively: $(A \perp\!\!\!\perp B)|C$ iff $\mathcal{L}(A \cap B|C) = \mathcal{L}(A|C) \cdot \mathcal{L}(B|C)$.

Definition B.3. (Definition 2.3 and Proposition 2.2 in Koller and Friedman (2009), (8.20) and (8.21) in Bishop (2006)). Let X, Y and Z be random elements (variables/vectors) and let $\mathcal{L}$ be a distribution. We say that X is *conditionally independent* of Y given Z (short: $(X \perp\!\!\!\perp Y)|Z$) if

$$\mathcal{L}(X|Y, Z) = \mathcal{L}(X|Z) \text{ whenever } \mathcal{L}(Y, Z) > 0.$$

Alternatively: $(X \perp\!\!\!\perp Y)|Z$ iff $\mathcal{L}(X, Y|Z) = \mathcal{L}(X|Z) \cdot \mathcal{L}(Y|Z)$.

Lemma B.4. *(Proposition 3.13 in Eaton (2007)). Let X and Y be two random column vectors. Suppose that the joint distribution of X and Y is Gaussian with mean $(\mu_1, \mu_2)^T$ and covariance matrix*

$$\mathbb{C}ov\left(\begin{pmatrix} X \\ Y \end{pmatrix}\right) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}.$$

Let $\mathcal{L}(X|Y=y)$ denote the conditional distribution of X given $Y=y$. Then

$$\mathcal{L}(X|Y=y) = \mathcal{N}\left(\mu_1 + \Sigma_{12}\cdot\Sigma_{22}^{-}\cdot(Y-\mu_2),\ \Sigma_{11} - \Sigma_{12}\cdot\Sigma_{22}^{-}\cdot\Sigma_{12}^T\right),$$

where Σ_{22}^{-} denotes the generalized inverse of Σ_{22}.

Lemma B.5. *Let $X = (X_1, \ldots, X_p)$ and $\tilde{X} = \left(\tilde{X}_1, \ldots, \tilde{X}_p\right)$ be two random row-vectors. Let*

$$\left(X, \tilde{X}\right) = \left(X_1, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_p\right) \sim \mathcal{N}\left(\begin{pmatrix} \mathbb{E}(X) & \mathbb{E}\left(\tilde{X}\right) \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}\right). \tag{29}$$

Then
i) $\mathbb{E}(\tilde{X}|X) = \mathbb{E}\left(\tilde{X}\right) + (X - \mathbb{E}(X))\cdot\mathbf{\Sigma}_{11}^{-1}\cdot\mathbf{\Sigma}_{12}$.
ii) $\mathbb{C}ov(\tilde{X}|X) = \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}\cdot\mathbf{\Sigma}_{11}^{-1}\cdot\mathbf{\Sigma}_{12}$.

Proof:

Note that because of using row vectors instead of the usual column vectors, definition of covariance matrix changes too.
Instead of $\mathbb{C}ov(V) = \mathbb{E}\left[(V - \mathbb{E}(V))\cdot(V - \mathbb{E}(V))^T\right]$, we have to use $\mathbb{C}ov(V) = \mathbb{E}\left[(V - \mathbb{E}(V))^T\cdot(V - \mathbb{E}(V))\right]$ for a row vector V.

Regarding notation:
We will use $\mathbb{C}ov(V)$ for the covariance matrix of a random row vector V and $\mathbb{C}ov(U,V)$ for the cross-covariance matrix of two random row vectors U and V.

Define $\mathbf{A} = -\mathbf{\Sigma}_{11}^{-1}\cdot\mathbf{\Sigma}_{12}$ and $Z = \tilde{X} + X\cdot\mathbf{A}$.

Prep1 wts $\mathbb{C}ov(Z, X) = \mathbf{0}$:

$\mathbb{C}ov(X, Z)$

We know that $Z = \tilde{X} + X\cdot\mathbf{A}$.

$$\begin{aligned}
&= \mathbb{C}ov\left(X, \tilde{X} + X\cdot\mathbf{A}\right) \\
&= \mathbb{E}\left[(X - \mathbb{E}(X))^T\cdot\left(\tilde{X} + X\cdot\mathbf{A} - \mathbb{E}\left(\tilde{X} + X\cdot\mathbf{A}\right)\right)\right] \\
&= \mathbb{E}\left[\left(X^T - (\mathbb{E}(X))^T\right)\cdot\left(\tilde{X} + X\cdot\mathbf{A} - \mathbb{E}\left(\tilde{X} + X\cdot\mathbf{A}\right)\right)\right] \\
&= \mathbb{E}\left[\left(X^T - (\mathbb{E}(X))^T\right)\cdot\left(\tilde{X} + X\cdot\mathbf{A} - \mathbb{E}\left(\tilde{X}\right) - \mathbb{E}(X\cdot\mathbf{A})\right)\right] \\
&= \mathbb{E}\left[X^T\cdot\tilde{X} + X^T\cdot X\cdot\mathbf{A} - X^T\cdot\mathbb{E}\left(\tilde{X}\right) - X^T\cdot\mathbb{E}(X\cdot\mathbf{A})\right.
\end{aligned}$$

$$
\begin{aligned}
&\quad - (\mathbb{E}(X))^T \cdot \tilde{X} - (\mathbb{E}(X))^T \cdot X \cdot \mathbf{A} + (\mathbb{E}(X))^T \cdot \mathbb{E}\left(\tilde{X}\right) + (\mathbb{E}(X))^T \cdot \mathbb{E}\left(X \cdot \mathbf{A}\right)\Big] \\
&= \mathbb{E}\left[X^T \cdot \tilde{X} - X^T \cdot \mathbb{E}\left(\tilde{X}\right) - (\mathbb{E}(X))^T \cdot \tilde{X} + (\mathbb{E}(X))^T \cdot \mathbb{E}\left(\tilde{X}\right)\right] \\
&\quad + \mathbb{E}\left[X^T \cdot X \cdot \mathbf{A} - X^T \cdot \mathbb{E}\left(X \cdot \mathbf{A}\right) - (\mathbb{E}(X))^T \cdot X \cdot \mathbf{A} + (\mathbb{E}(X))^T \cdot \mathbb{E}\left(X \cdot \mathbf{A}\right)\right] \\
&= \mathbb{E}\left[\left(X^T - (\mathbb{E}(X))^T\right) \cdot \left(\tilde{X} - \mathbb{E}\left(\tilde{X}\right)\right)\right] + \mathbb{E}\left[\left(X^T - (\mathbb{E}(X))^T\right) \cdot \left(X \cdot \mathbf{A} - \mathbb{E}\left(X \cdot \mathbf{A}\right)\right)\right] \\
&= \mathbb{C}ov\left(X, \tilde{X}\right) + \mathbb{C}ov\left(X, X \cdot \mathbf{A}\right)
\end{aligned}
$$

We can see in (29), that the cross-covariance matrix of X and $\tilde{X}$ is $\mathbf{\Sigma}_{12}$.

$$
\begin{aligned}
&= \mathbf{\Sigma}_{12} + \mathbb{C}ov\left(X, X \cdot \mathbf{A}\right)
\end{aligned}
$$

By a similar calculation as in Prep5 (below)

$$
\begin{aligned}
&= \mathbf{\Sigma}_{12} + \mathbb{C}ov\left(X, X\right) \cdot \mathbf{A} \\
&= \mathbf{\Sigma}_{12} + \mathbb{C}ov\left(X\right) \cdot \mathbf{A}
\end{aligned}
$$

We can see from (29) that $\mathbb{C}ov\left(X\right) = \mathbf{\Sigma}_{11}$.

$$
\begin{aligned}
&= \mathbf{\Sigma}_{12} + \mathbf{\Sigma}_{11} \cdot \mathbf{A}
\end{aligned}
$$

We know that $\mathbf{A} = -\mathbf{\Sigma}_{11}^{-1} \cdot \mathbf{\Sigma}_{12}$.

$$
\begin{aligned}
&= \mathbf{\Sigma}_{12} + \mathbf{\Sigma}_{11} \cdot \left(-\mathbf{\Sigma}_{11}^{-1} \cdot \mathbf{\Sigma}_{12}\right) \\
&= \mathbf{\Sigma}_{12} - \underbrace{\mathbf{\Sigma}_{11} \cdot \mathbf{\Sigma}_{11}^{-1}}_{=\mathbf{I}} \cdot \mathbf{\Sigma}_{12} \\
&= \mathbf{\Sigma}_{12} - \mathbf{\Sigma}_{12} \\
&= \mathbf{0}.
\end{aligned}
$$

Note that this implies that X and Z are uncorrelated.
Furthermore, note that X and Z are jointly normal (see proof of Proposition 3.13 in Eaton (2007))
Therefore, X and Z are independent.

ad i) wts $\mathbb{E}(\tilde{X}|X) = \mathbb{E}\left(\tilde{X}\right) + (X - \mathbb{E}(X)) \cdot \mathbf{\Sigma}_{11}^{-1} \cdot \mathbf{\Sigma}_{12}$:

$$
\begin{aligned}
&\mathbb{E}(\tilde{X}|X) \\
&\qquad\qquad \text{We know that } Z = \tilde{X} + X \cdot \mathbf{A}. \\
&= \mathbb{E}\left(Z - X \cdot \mathbf{A}|X\right) \\
&= \mathbb{E}\left(Z|X\right) - \mathbb{E}\left(X \cdot \mathbf{A}|X\right) \\
&\qquad\qquad \text{By Prep1, we know that } X \text{ and } Z \text{ are independent.} \\
&= \mathbb{E}\left(Z\right) - \mathbb{E}\left(X \cdot \mathbf{A}|X\right) \\
&= \mathbb{E}\left(Z\right) - X \cdot \mathbb{E}\left(\mathbf{A}|X\right) \\
&= \mathbb{E}\left(Z\right) - X \cdot \mathbf{A} \\
&\qquad\qquad \text{We know that } Z = \tilde{X} + X \cdot \mathbf{A}. \\
&= \mathbb{E}\left(\tilde{X} + X \cdot \mathbf{A}\right) - X \cdot \mathbf{A} \\
&= \mathbb{E}\left(\tilde{X}\right) + \mathbb{E}\left(X \cdot \mathbf{A}\right) - X \cdot \mathbf{A} \\
&= \mathbb{E}\left(\tilde{X}\right) + \mathbb{E}\left(X\right) \cdot \mathbf{A} - X \cdot \mathbf{A} \\
&= \mathbb{E}\left(\tilde{X}\right) + \left(\mathbb{E}\left(X\right) - X\right) \cdot \mathbf{A} \\
&\qquad\qquad \text{We know that } \mathbf{A} = -\mathbf{\Sigma}_{11}^{-1} \cdot \mathbf{\Sigma}_{12}.
\end{aligned}
$$

$$= \mathbb{E}\left(\tilde{X}\right) + \left(\mathbb{E}\left(X\right) - X\right) \cdot \left(-\mathbf{\Sigma}_{11}^{-1} \cdot \mathbf{\Sigma}_{12}\right)$$
$$= \mathbb{E}\left(\tilde{X}\right) + \left(X - \mathbb{E}\left(X\right)\right) \cdot \mathbf{\Sigma}_{11}^{-1} \cdot \mathbf{\Sigma}_{12}.$$

ad ii) wts $\mathbb{C}ov\left(\tilde{X}|X\right) = \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21} \cdot \mathbf{\Sigma}_{11}^{-1} \cdot \mathbf{\Sigma}_{12}$:

Prep2 wts $\mathbb{C}ov\left(Z|X\right) = \mathbb{C}ov(Z)$:

$$\begin{aligned}
&\mathbb{C}ov\left(Z|X\right) \\
&= \mathbb{E}\left[\left(Z - \mathbb{E}\left(Z|X\right)\right)^T \cdot \left(Z - \mathbb{E}\left(Z|X\right)\right)\right] \\
&\qquad\qquad \text{By Prep1, we know} \\
&\qquad\qquad \text{that } X \text{ and } Z \text{ are independent.} \\
&= \mathbb{E}\left[\left(Z - \mathbb{E}\left(Z\right)\right)^T \cdot \left(Z - \mathbb{E}\left(Z\right)\right)\right] \\
&= \mathbb{C}ov(Z).
\end{aligned}$$

Prep3 wts $\mathbb{C}ov\left(-X \cdot \mathbf{A}|X\right) = \mathbf{0}$:

$$\begin{aligned}
&\mathbb{C}ov\left(-X \cdot \mathbf{A}|X\right) \\
&= \mathbb{E}\left[\left(-X \cdot \mathbf{A} - \mathbb{E}\left(-X \cdot \mathbf{A}|X\right)\right)^T \cdot \left(-X \cdot \mathbf{A} - \mathbb{E}\left(-X \cdot \mathbf{A}|X\right)\right)\right] \\
&= \mathbb{E}\left[\left(-X \cdot \mathbf{A} + \mathbb{E}\left(X \cdot \mathbf{A}|X\right)\right)^T \cdot \left(-X \cdot \mathbf{A} + \mathbb{E}\left(X \cdot \mathbf{A}|X\right)\right)\right] \\
&= \mathbb{E}\left[\left(-X \cdot \mathbf{A} + X \cdot \mathbb{E}\left(\mathbf{A}|X\right)\right)^T \cdot \left(-X \cdot \mathbf{A} + X \cdot \mathbb{E}\left(\mathbf{A}|X\right)\right)\right] \\
&= \mathbb{E}\left[\left(-X \cdot \mathbf{A} + X \cdot \mathbf{A}\right)^T \cdot \left(-X \cdot \mathbf{A} + X \cdot \mathbf{A}\right)\right] \\
&= \mathbf{0}.
\end{aligned}$$

Prep4 wts $\mathbb{C}ov\left(X \cdot \mathbf{A}, \tilde{X}\right) = \mathbf{A}^T \cdot \mathbb{C}ov\left(X, \tilde{X}\right)$:

$$\begin{aligned}
&\mathbb{C}ov\left(X \cdot \mathbf{A}, \tilde{X}\right) \\
&= \mathbb{E}\left[\left(X \cdot \mathbf{A} - \mathbb{E}\left(X \cdot \mathbf{A}\right)\right)^T \cdot \left(\tilde{X} - \mathbb{E}\left(\tilde{X}\right)\right)\right] \\
&= \mathbb{E}\left[\left(\left(X - \mathbb{E}\left(X\right)\right) \cdot \mathbf{A}\right)^T \cdot \left(\tilde{X} - \mathbb{E}\left(\tilde{X}\right)\right)\right] \\
&= \mathbb{E}\left[\mathbf{A}^T \cdot \left(X - \mathbb{E}\left(X\right)\right)^T \cdot \left(\tilde{X} - \mathbb{E}\left(\tilde{X}\right)\right)\right] \\
&= \mathbf{A}^T \cdot \mathbb{E}\left[\left(X - \mathbb{E}\left(X\right)\right)^T \cdot \left(\tilde{X} - \mathbb{E}\left(\tilde{X}\right)\right)\right] \\
&= \mathbf{A}^T \cdot \mathbb{C}ov\left(X, \tilde{X}\right).
\end{aligned}$$

Prep5 wts $\mathbb{C}ov\left(\tilde{X}, X \cdot \mathbf{A}\right) = \mathbb{C}ov\left(\tilde{X}, X\right) \cdot \mathbf{A}$:

$$\begin{aligned}
&\mathbb{C}ov\left(\tilde{X}, X \cdot \mathbf{A}\right) \\
&= \mathbb{E}\left[\left(\tilde{X} - \mathbb{E}\left(\tilde{X}\right)\right)^T \cdot \left(X \cdot \mathbf{A} - \mathbb{E}\left(X \cdot \mathbf{A}\right)\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left(\tilde{X} - \mathbb{E}\left(\tilde{X}\right)\right)^T \cdot (X \cdot \mathbf{A} - \mathbb{E}(X) \cdot \mathbf{A})\right] \\
&= \mathbb{E}\left[\left(\tilde{X} - \mathbb{E}\left(\tilde{X}\right)\right)^T \cdot (X - \mathbb{E}(X)) \cdot \mathbf{A}\right] \\
&= \mathbb{E}\left[\left(\tilde{X} - \mathbb{E}\left(\tilde{X}\right)\right)^T \cdot (X - \mathbb{E}(X))\right] \cdot \mathbf{A} \\
&= \mathbb{C}ov\left(\tilde{X}, X\right) \cdot \mathbf{A}.
\end{aligned}
$$

Prep6 wts $\mathbb{C}ov(X \cdot \mathbf{A}) = \mathbf{A}^T \cdot \mathbb{C}ov(X) \cdot \mathbf{A}$:

$$
\begin{aligned}
&\mathbb{C}ov(X \cdot \mathbf{A}) \\
&= \mathbb{E}\left[(X \cdot \mathbf{A} - \mathbb{E}(X \cdot \mathbf{A}))^T \cdot (X \cdot \mathbf{A} - \mathbb{E}(X \cdot \mathbf{A}))\right] \\
&= \mathbb{E}\left[(X \cdot \mathbf{A} - \mathbb{E}(X) \cdot \mathbf{A})^T \cdot (X \cdot \mathbf{A} - \mathbb{E}(X) \cdot \mathbf{A})\right] \\
&= \mathbb{E}\left[((X - \mathbb{E}(X)) \cdot \mathbf{A})^T \cdot (X - \mathbb{E}(X)) \cdot \mathbf{A}\right] \\
&= \mathbb{E}\left[\mathbf{A}^T \cdot (X - \mathbb{E}(X))^T \cdot ((X - \mathbb{E}(X)) \cdot \mathbf{A})\right] \\
&= \mathbf{A}^T \cdot \mathbb{E}\left[(X - \mathbb{E}(X))^T \cdot ((X - \mathbb{E}(X)))\right] \cdot \mathbf{A} \\
&= \mathbf{A}^T \cdot \mathbb{C}ov(X) \cdot \mathbf{A}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathbb{C}ov\left(\tilde{X}|X\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{We know that } Z = \tilde{X} + X \cdot \mathbf{A}. \\
&= \mathbb{C}ov(Z - X \cdot \mathbf{A}|X) \\
&= \mathbb{C}ov(Z + (-X \cdot \mathbf{A})|X) \\
&= \mathbb{C}ov(Z|X) + \mathbb{C}ov(-X \cdot \mathbf{A}|X) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{By Prep3} \\
&= \mathbb{C}ov(Z|X) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{By Prep2} \\
&= \mathbb{C}ov(Z) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{We know that } Z = \tilde{X} + X \cdot \mathbf{A}. \\
&= \mathbb{C}ov\left(\tilde{X} + X \cdot \mathbf{A}\right) \\
&= \mathbb{E}\left[\left(\tilde{X} + X \cdot \mathbf{A} - \mathbb{E}\left(\tilde{X} + X \cdot \mathbf{A}\right)\right)^T \cdot \left(\tilde{X} + X \cdot \mathbf{A} - \mathbb{E}\left(\tilde{X} + X \cdot \mathbf{A}\right)\right)\right] \\
&= \mathbb{E}\left[\left(\tilde{X} + X \cdot \mathbf{A} - \mathbb{E}\left(\tilde{X}\right) - \mathbb{E}(X \cdot \mathbf{A})\right)^T \cdot \left(\tilde{X} + X \cdot \mathbf{A} - \mathbb{E}\left(\tilde{X}\right) - \mathbb{E}(X \cdot \mathbf{A})\right)\right] \\
&= \mathbb{E}\left[\left(\tilde{X}^T + \mathbf{A}^T \cdot X^T - \left(\mathbb{E}\left(\tilde{X}\right)\right)^T - (\mathbb{E}(X \cdot \mathbf{A}))^T\right) \cdot \left(\tilde{X} + X \cdot \mathbf{A} - \mathbb{E}\left(\tilde{X}\right) - \mathbb{E}(X \cdot \mathbf{A})\right)\right] \\
&= \mathbb{E}\Big[\tilde{X}^T \cdot \tilde{X} + \tilde{X}^T \cdot X \cdot \mathbf{A} - \tilde{X}^T \cdot \mathbb{E}\left(\tilde{X}\right) - \tilde{X}^T \cdot \mathbb{E}(X \cdot \mathbf{A}) \\
&\qquad + \mathbf{A}^T \cdot X^T \cdot \tilde{X} + \mathbf{A}^T \cdot X^T \cdot X \cdot \mathbf{A} - \mathbf{A}^T \cdot X^T \cdot \mathbb{E}\left(\tilde{X}\right) - \mathbf{A}^T \cdot X^T \cdot \mathbb{E}(X \cdot \mathbf{A}) \\
&\qquad - \left(\mathbb{E}\left(\tilde{X}\right)\right)^T \cdot \tilde{X} - \left(\mathbb{E}\left(\tilde{X}\right)\right)^T \cdot X \cdot \mathbf{A} + \left(\mathbb{E}\left(\tilde{X}\right)\right)^T \cdot \mathbb{E}\left(\tilde{X}\right) + \left(\mathbb{E}\left(\tilde{X}\right)\right)^T \cdot \mathbb{E}(X \cdot \mathbf{A})
\end{aligned}
$$

$$
\begin{aligned}
&\quad - \left(\mathbb{E}\left(X \cdot \mathbf{A}\right)\right)^T \cdot \tilde{X} - \left(\mathbb{E}\left(X \cdot \mathbf{A}\right)\right)^T \cdot X \cdot \mathbf{A} + \left(\mathbb{E}\left(X \cdot \mathbf{A}\right)\right)^T \cdot \mathbb{E}\left(\tilde{X}\right) + \left(\mathbb{E}\left(X \cdot \mathbf{A}\right)\right)^T \cdot \mathbb{E}\left(X \cdot \mathbf{A}\right)\Big] \\
&= \mathbb{E}\left[\tilde{X}^T \cdot \tilde{X} - \tilde{X}^T \cdot \mathbb{E}\left(\tilde{X}\right) - \left(\mathbb{E}\left(\tilde{X}\right)\right)^T \cdot \tilde{X} + \left(\mathbb{E}\left(\tilde{X}\right)\right)^T \cdot \mathbb{E}\left(\tilde{X}\right)\right] \\
&\quad + \mathbb{E}\left[\mathbf{A}^T \cdot X^T \cdot X \cdot \mathbf{A} - \mathbf{A}^T \cdot X^T \cdot \mathbb{E}\left(X \cdot \mathbf{A}\right) - \left(\mathbb{E}\left(X \cdot \mathbf{A}\right)\right)^T \cdot X \cdot \mathbf{A} + \left(\mathbb{E}\left(X \cdot \mathbf{A}\right)\right)^T \cdot \mathbb{E}\left(X \cdot \mathbf{A}\right)\right] \\
&\quad + \mathbb{E}\left[\tilde{X}^T \cdot X \cdot \mathbf{A} - \tilde{X}^T \cdot \mathbb{E}\left(X \cdot \mathbf{A}\right) - \left(\mathbb{E}\left(\tilde{X}\right)\right)^T \cdot X \cdot \mathbf{A} + \left(\mathbb{E}\left(\tilde{X}\right)\right)^T \cdot \mathbb{E}\left(X \cdot \mathbf{A}\right)\right] \\
&\quad + \mathbb{E}\left[\mathbf{A}^T \cdot X^T \cdot \tilde{X} - \mathbf{A}^T \cdot X^T \cdot \mathbb{E}\left(\tilde{X}\right) - \left(\mathbb{E}\left(X \cdot \mathbf{A}\right)\right)^T \cdot \tilde{X} + \left(\mathbb{E}\left(X \cdot \mathbf{A}\right)\right)^T \cdot \mathbb{E}\left(\tilde{X}\right)\right] \\
&= \mathbb{E}\left[\left(\tilde{X} - \mathbb{E}(\tilde{X})\right)^T \cdot \left(\tilde{X} - \mathbb{E}(\tilde{X})\right)\right] \\
&\quad + \mathbb{E}\left[\left(X \cdot \mathbf{A} - \mathbb{E}\left(X \cdot \mathbf{A}\right)\right)^T \cdot \left(X \cdot \mathbf{A} - \mathbb{E}\left(X \cdot \mathbf{A}\right)\right)\right] \\
&\quad + \mathbb{E}\left[\left(\tilde{X} - \mathbb{E}\left(\tilde{X}\right)\right)^T \cdot \left(X \cdot \mathbf{A} - \mathbb{E}\left(X \cdot \mathbf{A}\right)\right)\right] \\
&\quad + \mathbb{E}\left[\left(X \cdot \mathbf{A} - \mathbb{E}\left(X \cdot \mathbf{A}\right)\right)^T \cdot \left(\tilde{X} - \mathbb{E}\left(\tilde{X}\right)\right)\right] \\
&= \mathbb{C}ov\left(\tilde{X}\right) + \mathbb{C}ov\left(X \cdot \mathbf{A}\right) + \mathbb{C}ov\left(\tilde{X}, X \cdot \mathbf{A}\right) + \mathbb{C}ov\left(X \cdot \mathbf{A}, \tilde{X}\right) \\
&\hspace{20em} \text{By Prep4, Prep5 and Prep6} \\
&= \mathbb{C}ov\left(\tilde{X}\right) + \mathbf{A}^T \cdot \mathbb{C}ov\left(X\right) \cdot \mathbf{A} + \mathbb{C}ov\left(\tilde{X}, X\right) \cdot \mathbf{A} + \mathbf{A}^T \cdot \mathbb{C}ov\left(X, \tilde{X}\right) \\
&\hspace{20em} \text{By (29)} \\
&= \mathbf{\Sigma}_{22} + \mathbf{A}^T \cdot \mathbf{\Sigma}_{11} \cdot \mathbf{A} + \mathbf{\Sigma}_{21} \cdot \mathbf{A} + \mathbf{A}^T \cdot \mathbf{\Sigma}_{12} \\
&\hspace{20em} \text{We know that } \mathbf{A} = -\mathbf{\Sigma}_{11}^{-1} \cdot \mathbf{\Sigma}_{12}. \\
&= \mathbf{\Sigma}_{22} + \left(-\mathbf{\Sigma}_{11}^{-1} \cdot \mathbf{\Sigma}_{12}\right)^T \cdot \mathbf{\Sigma}_{11} \cdot \left(-\mathbf{\Sigma}_{11}^{-1} \cdot \mathbf{\Sigma}_{12}\right) + \mathbf{\Sigma}_{21} \cdot \left(-\mathbf{\Sigma}_{11}^{-1} \cdot \mathbf{\Sigma}_{12}\right) + \left(-\mathbf{\Sigma}_{11}^{-1} \cdot \mathbf{\Sigma}_{12}\right)^T \cdot \mathbf{\Sigma}_{12} \\
&= \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{12}^T \cdot \left(\mathbf{\Sigma}_{11}^{-1}\right)^T \cdot \mathbf{\Sigma}_{11} \cdot \left(-\mathbf{\Sigma}_{11}^{-1} \cdot \mathbf{\Sigma}_{12}\right) - \mathbf{\Sigma}_{21} \cdot \mathbf{\Sigma}_{11}^{-1} \cdot \mathbf{\Sigma}_{12} - \mathbf{\Sigma}_{12}^T \cdot \left(\mathbf{\Sigma}_{11}^{-1}\right)^T \cdot \mathbf{\Sigma}_{12} \\
&= \mathbf{\Sigma}_{22} + \mathbf{\Sigma}_{12}^T \cdot \left(\mathbf{\Sigma}_{11}^{-1}\right)^T \cdot \underbrace{\mathbf{\Sigma}_{11} \cdot \mathbf{\Sigma}_{11}^{-1}}_{=\mathbf{I}} \cdot \mathbf{\Sigma}_{12} - \mathbf{\Sigma}_{21} \cdot \mathbf{\Sigma}_{11}^{-1} \cdot \mathbf{\Sigma}_{12} - \mathbf{\Sigma}_{12}^T \cdot \left(\mathbf{\Sigma}_{11}^{-1}\right)^T \cdot \mathbf{\Sigma}_{12} \\
&= \mathbf{\Sigma}_{22} + \mathbf{\Sigma}_{12}^T \cdot \left(\mathbf{\Sigma}_{11}^{-1}\right)^T \cdot \mathbf{\Sigma}_{12} - \mathbf{\Sigma}_{21} \cdot \mathbf{\Sigma}_{11}^{-1} \cdot \mathbf{\Sigma}_{12} - \mathbf{\Sigma}_{12}^T \cdot \left(\mathbf{\Sigma}_{11}^{-1}\right)^T \cdot \mathbf{\Sigma}_{12} \\
&= \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21} \cdot \mathbf{\Sigma}_{11}^{-1} \cdot \mathbf{\Sigma}_{12}.
\end{aligned}
$$

□

C Appendix - English abstract

In many fields of science, a response variable together with a large number of potential explanatory variables are observed and researchers would like to know which variables are truly associated with the response. Variable selection procedures are used to detect such important explanatory variables. FDR-controlling variable selection procedures accomplish this task and make sure that most of the important variables are selected while at the same time, the number of falsely selected variables is not too high. Barber and Candes (2015) introduced such a FDR-controlling variable selection method using so-called Knockoffs to tease apart important and unimportant variables. Using the same ideas, Candes et al. (2018) created a FDR-controlling variable selection method for a setting, where the distribution of the explanatory variables is known and where the number of explanatory variables can be much higher than the number of observations.
This thesis presents both techniques in a lecture-note style with detailed proofs.

D Appendix - German abstract

Oftmals wird eine abhängige Variable zusammen mit einer Vielzahl an Kovariaten beobachtet und Wissenschaftler sind an denjenigen Kovariaten interessiert, die einen echten Zusammenhang mit der abhängigen Variable aufweisen. Diese wichtigen Kovariaten werden mittels Variablenselektionsverfahren ausgewählt. Eine Gruppe solcher Verfahren hält dabei die FDR unter Kontrolle. Das heißt, es werden möglichst viele der wichtigen Kovariaten ausgewählt, während gleichzeitig die Anzahl an fälschlicherweise gewählten Kovariaten nicht zu groß wird. Barber and Candes (2015) haben eine solche FDR-einhaltende Variablenselektionsmethode vorgestellt, die sogenannte Knockoffs nutzt um wichtige von unwichtigen Kovariaten zu unterscheiden. Diese Ideen haben Candes et al. (2018) verwendet, um eine FDR-einhaltende Variablenselektionsmethode für eine Situation zu schaffen, in der die Verteilung der Kovariaten bekannt ist und in der die Anzahl der Kovariaten beliebig groß sein kann.
Diese Arbeit stellt beide Prozeduren mit ausführlichen Beweisen im Stil eines Vorlesungsskriptes vor.