ETH
Swiss Federal Institute of Technology Zurich

Seminar for Statistics

Department of Mathematics

Master Thesis Summer 2021

Chrysovalantis Karypidis

The (multiple) knockoff filter: A powerful variable selection with FDR control

Submission Date: September 16th 2021

Adviser: Dr. Lukas Meier

Abstract

Today's data sets often contain a large number of variables, and the researcher is interested in finding those few explaining the response of interest without too many false discoveries. The control of the false discovery rate (FDR) ensures that most findings are true effects and can be reproduced by follow-up research. This thesis deals with the knockoff filter, a modern variable selection technique with FDR control. We start by elaborating on fixed-X knockoffs that achieve FDR control in low-dimensional Gaussian linear models. We show in simulations that the knockoff filter is more powerful than existing popular selection rules while controlling the FDR. In a novel follow-up simulation, we conclude that the choice of a proper score function, a key ingredient of the knockoff filter to compare the importance of an original variable and its knockoff, does not affect the FDR control but has a great influence on the power. We suggest that the researcher selects one that embeds the most information. Continuing with the extension to model-X knockoffs, which can be applied to almost any model regardless of the dimensionality, we develop open research questions that should be examined by future research, and we aim to answer two of those. First, we show that model-X knockoffs are superior to fixed-X knockoffs in a linear model, even when the model is misspecified. Second, we study whether the aggregation of multiple knockoff runs leads to power improvements. We compare the three aggregation schemes union knockoffs, p-value knockoffs and ADAGES. While ADAGES and union knockoffs have more power than model-X knockoffs with one run, this is not always the case for p-value knockoffs. ADAGES leads to the largest power but suffers from empirical FDR values above the pre-specified nominal level in some settings. Union knockoffs provide empirical FDR values that are very close to model-X knockoffs but improve their power. Thus, multiple knockoffs can indeed be used to improve the power, but the researcher has to accept (slightly) larger FDR values. We also provide a user-friendly implementation of all three aggregation techniques with our package `multiknockoffs` in the statistical software `R`.

Keywords: Variable selection, false discovery rate, knockoff filter, Lasso

Contents

List of Figures

List of Tables

Acronyms

ADAGES adaptive aggregation with stability for distributed feature selection.

ASDP approximate semidefinite program.

BFDR Bayesian false discovery rate.

BH Benjamini-Hochberg.

BY Benjamini-Yekutieli.

BYK Benjamini-Yekutieli-Krieger.

CV cross-validation.

FDP false discovery proportion.

FDR false discovery rate.

FWER family-wise error rate.

FX fixed-X.

GLM generalized linear models.

i.i.d. independently and identically distributed.

KL Kullback-Leibler.

Lasso Least absolute shrinkage and selection operator.

LCD Lasso coefficient difference.

LLD Lasso lambda difference.

LLSM Lasso lambda signed max.

LS least squares.

MCorr marginal correlation.

MX model-X.

OMP orthogonal matching pursuit.

PEF pairwise exchangeability for the features.

PER pairwise exchangeability for the response.

pFDR positive false discovery rate.

pKO p-value knockoffs.

PRDS positive regression dependence on a subset.

SCIP Sequential Conditional Independent Pairs.

SDP semidefinite programming.

SNR signal-to-noise ratio.

uKO union knockoffs.

Chapter 1

Introduction

The technical developments in recent decades have given scientists and companies access to extremely large data sets. Nowadays, it is common that we can collect thousands of features for a single observation. This has also changed the philosophy of research on testing hypotheses to quite an extent. Especially before the 90s, the conventional approach of many statistical analyses was to formulate the desired hypotheses prior to the data collection (hypothesis-driven research). After collecting the data, researchers tested their hypothesis by investigating the influence of the variable of interest on the response, which mathematically encodes the hypothesis, while controlling for other predictors. Nowadays, however, many scientists conduct data-driven research by first collecting large data sets and then trying to find significant variables among the thousands through individual hypotheses tests. A variable is declared as significant if its p-value is below the "0.05 mark". Even though this approach does yield some insights, it bears the risk of overinterpreting these findings. Because of the type-one error cumulation, some variables will be significant purely by chance without a real effect. Hence, research that is solely based on individual (non-corrected) p-values is often not reproducible by follow-up experiments (Ioannidis 2005). The media refers to the difficulties of replicating scientific results as the reproducibility crisis (Baker 2016). In medicine, several studies have shown that the results of numerous well-cited papers could not be reproduced (Begley and Ellis 2012). This phenomenon is not only prominent in medicine. The irreproducibility of research results is also present in economics (Chang and Li 2015), psychology (Open Science Collaboration 2015) or environmental sciences (Stagge et al. 2019).
The statistical research community has developed several joint type-one error measures to account for conducting many individual hypothesis tests at once. In this work, we focus on the false discovery rate (FDR) formalized by Benjamini and Hochberg (1995), that is the expected proportion of false discoveries. By controlling the FDR, researchers have some security that most of their significant variables correspond to true effects and should be reproducible by subsequent experiments. The first and by far most popular method, the Benjamini-Hochberg (BH) procedure, has sparked a series of follow up research extending their algorithm (Benjamini and Yekutieli 2001; Benjamini et al. 2006) or proposing similar approaches (Storey 2002; Storey et al. 2004). Most of the classical FDR controlling approaches require the computation of p-values and can only be applied in the low-dimensional framework ($n \geq p$). In addition, the p-values have to be independent (Benjamini and Hochberg 1995) or satisfy some sort of positive dependence property (Benjamini and Yekutieli 2001). The Benjamini-Yekutieli (BY) method accounts

for general dependencies at a price of being very conservative. Many modern regression techniques involve penalization terms such as the Least absolute shrinkage and selection operator (Lasso) model (Tibshirani 1996), for which the p-value computation is already problematic in the low-dimensional setting, and even more non-trivial for high-dimensional applications ($n < p$). An approach that has gained acceptance is de-biasing the Lasso coefficients, which have an asymptotically normal distribution afterwards (Zhang and Zhang 2014). The researcher can use the asymptotic p-values and apply the BH procedure, but the p-values will still suffer from dependencies. Further research has focused on procedures that are more tailored to the FDR control for Lasso. G'Sell et al. (2015) achieve FDR control based on the λ-sequence at which new covariates enter the Lasso regularization path. Their approach relies on the assumption that signal variables enter the model before any null variable, although they are often interspersed along the path in practice. Javanmard and Javadi (2019) develop a procedure to control the (directional) FDR with the de-biased Lasso in a framework where the number of true variables among the p is bounded by $o(\sqrt{n}/\log(p)^2)$. In addition, their FDR control requires an asymptotic regime $(n, p) \to \infty$.

Barber and Candès (2015) introduced knockoffs as a completely new approach to control the FDR. The knockoff filter works by constructing fake features, that mimic certain correlation properties of the original variables. Since knockoffs behave similar to the original features but are known to be artificial null variables, they serve as a negative control group. The knockoff filter does not rely on p-values and controls the FDR without making assumptions on the design matrix, coefficient sizes or the noise level. The FDR control holds in finite samples, so it does not require asymptotic statements. Knockoffs are very flexible, and they can be applied with a wide range of variable selection techniques, such as Lasso. However, their application area is restricted to low-dimensional Gaussian linear models. Since the initial work by Barber and Candès (2015) treats the covariates as fixed, their procedure is also called fixed-X knockoffs. In a follow-up work, Candès et al. (2018) develop the extension model-X knockoffs that can be applied to almost any linear and non-linear estimation technique regardless of the dimensionality. The model-X knockoff filter treats the covariates as random, and their distribution must be known. Since model-X knockoffs do not assume anything on the relationship between the response and the covariates, they can be used together with almost any modern machine-learning model. Hence, the FDR control of model-X knockoffs provides some interpretability of black-box model approaches, for which a well-developed inference theory does not exist. An undesirable characteristic of the probabilistic nature of model-X knockoffs is that repeatedly running the procedure with the same data will lead to different knockoffs and thus (slightly) different selection sets of variables each time. The variability of the selection sets causes fluctuations of FDP and power values. The research community has developed multiple knockoff procedures that aim to aggregate the variable selection sets of multiple knockoff runs, leading to more stable results. They do not only claim to retain FDR control but also to boost the power compared to one knockoff run. The most promising aggregation schemes are union knockoffs (Xie and Lederer 2021), p-value knockoffs (Nguyen et al. 2020) and ADAGES (Gui 2020). Since all three methods are fairly new, there is a lack of detailed comparative studies on how they perform against each other. Although Gui (2020) shows that ADAGES achieves greater power than the other two methods while retaining FDR control, his comparison is not very representative. He does not choose the authors' recommended hyperparameters for union and p-value knockoffs, and he investigates only very few parameter variations. Nevertheless, his analysis is a first reference point.

This work seeks to contribute to the current research literature in three ways: i.) We

will begin by presenting fixed-X knockoffs in a detailed and illustrative manner. With a different simulation model than Barber and Candès (2015), we will provide additional evidence for or against the superiority of knockoffs compared to the BH and BY procedure. Subsequently, we will carry out a novel simulation where we compare the power of various score functions for knockoff filters. These score functions are an essential part of the procedure that quantify whether and to which extend an original variable is more important than its knockoff counterpart in explaining the response. ii.) We move on by introducing model-X knockoffs as a natural extension. We will develop open research questions that are in our opinion the most fundamental issues which have to be tackled by future research. By comparing fixed-X and model-X knockoffs in a linear model but with non-Gaussian features, we will answer one of those open research questions, namely whether fixed-X knockoffs are outdated or still useful in some specific parameter settings. iii.) Having gained the necessary knowledge on model-X knockoffs, we will dive deeper and pick up another open research question where we deal with the extension to multiple knockoffs. After we have described the three aggregation methods union knockoffs, p-value knockoffs and ADAGES, we conduct an extensive simulation study where we answer the question of whether the aggregation methods are superior to model-X knockoffs and whether one of them has uniformly higher power while controlling the FDR. To the best of our knowledge, we are the first who compare the three methods under fair conditions and various settings. Since none of the aggregation methods is easily accessible in the statistical software `R` so far, we develop our own package called `multiknockoffs`. It implements all three multiple knockoff procedures in a very user-friendly way and with great flexibility.
The work is structured as follows: In Chapter 2, we discuss the framework of multiple testing, present some error measures, and then focus on the FDR as well as procedures for its control. We continue by elaborating on fixed-X knockoffs in Chapter 3, followed by their extension model-X knockoffs in Chapter 4. In Chapter 5, we elaborate on multiple knockoff procedures before we finish with a conclusion of the whole work in Chapter 6.

Chapter 2

Large-scale hypothesis testing

The development of modern technologies facilitates firms and researchers to collect and analyze enormous data sets. Nowadays, a large number of different variables p for each individual are often obtained with ease. Especially in biology, the flood of high-throughput technologies have provided huge genome-wide data sets to perform hypotheses tests on thousands of different features. For example, DNA microarrays enable the investigation of thousands of genes simultaneously and thus to analyze which genes show different expression levels between two or more physiological states (e.g. cancer/no cancer). Hence, the researcher's goal could be to find a few interesting variables among a haystack (Storey and Tibshirani 2003). Of course, this does not only apply to biostatistics. Due to the increasing dimensionality of data sets, finding those features that truly affect the outcome variable has become an essential step in statistical analyses of any field.
Section 2.1 starts with a general and mathematical definition of the variable selection problem before Section 2.2 briefly elaborates on the multiple testing problem. We continue with the formalization of the FDR as an error measure and present two procedures to control for it in Section 2.3. In addition, we also discuss ways to improve their power and introduce an empirical Bayes view on the FDR theory in this last section.

2.1 Formalizing the variable selection

To introduce the variable selection problem in statistical terms, let the random variable Y be the response, which depends on some factors we want to find out. Further, let $X = (X_1, \dots, X_p)$ be a possibly large collection of p potential explanatory features. We denote $F_{Y|X}$ to be the conditional distribution of the response Y given the features $X = (X_1, \dots, X_p)$. This distribution characterizes how the response depends on the features, and it can be arbitrary. Depending on the statistical model applied, different distributional assumptions are made on $F_{Y|X}$. We denote $(X_{i,1}, \dots, X_{i,p}, Y_i)$ to be an observation, and we have a sample n of them. Moreover, we assume that conditionally on the features, the outcome observations are independently and identically distributed (i.i.d.), and their conditional distribution only depends on $(X_{i,1}, \dots, X_{i,p})$, that is

$$Y_i|(X_{i,1}, \dots, X_{i,p}) \overset{i.i.d.}{\sim} F_{Y|X}, \quad i = 1, \dots, n.$$

The fundamental belief of variable selection is that $F_{Y|X}$ only depends on a small subset of variables $\mathcal{S} \subset \{1, \dots, p\}$ which we want to identify. Conditionally on that set $\mathcal{S}$, Y is

independent of all other covariates

$$Y \perp\!\!\!\perp \{X_j\}_{j \in \mathcal{S}^c} | \{X_j\}_{j \in \mathcal{S}},$$

where $\mathcal{S}^c$ is the complement of $\mathcal{S}$. These other variables in $\mathcal{S}^c$ can be considered as null or unimportant features because they do not provide any additional information in explaining Y. In the area of graphical models, the smallest set $\mathcal{S}$ fulfilling this property is defined as the Markov boundary of Y (Pearl 1988, p. 97). To rule out ill-defined cases of this Markov boundary, we formally define null and signal variables as it was done by Candès et al. (2018):

Definition 2.1. *A variable X_j is defined as null varible if and only if the response is independent of that variable conditionally on all other variables $X_{-j} = \{X_1, \ldots, X_p\} \setminus \{X_j\}$, that is,*

$$j \in \mathcal{H}_0 \iff Y \perp\!\!\!\perp X_j | X_{-j},$$

where $\mathcal{H}_0 \subset \{1, \ldots, p\}$ represents the set of all null variables. Equivalently, we define a variable X_j as signal or relevant variable if $j \notin \mathcal{H}_0$, and we denote the set of all signal variables as $\mathcal{S}_0$.

Under some non-restrictive conditions on $F_{Y|X}$, the Markov boundary is usually unique and it coincides with the set of signal variables.[1] Hence, the variable selection problem is validly defined in mathematical terms. Ideally, our goal will be to find exactly the set $\mathcal{S}_0$ without making too many mistakes. In the next sections, we will explicitly define what we are going to consider as a mistake.

2.2 Primer: Classical multiple testing

After we have defined the necessary technicalities for variable selection, we focus on the setting of classic hypothesis testing and reconsider the gene expression level example. Assume the researcher has data about different gene expression levels for cancer patients and a control group. Examining the differences of thousands of genes between both groups translates statistically into performing a collection of individual hypothesis tests of the form

$$H_{0,j} = \text{gene } j \text{ is a null} \quad \text{vs.} \quad H_{A,j} = \text{gene } j \text{ is a signal.}$$

More specifically, a two-sample t-statistic can be computed for each gene j, testing if the expression levels of gene j follow the same distribution for normal and cancer patients under the null hypothesis. Equivalently, gene j is a signal if its expression levels follow different distributions for the two patient populations.
Having conducted all individual hypothesis tests, the question of which of them to reject arises. In linear models, the researcher might be tempted to judge her decision by evaluating if each p-value is smaller than the famous 0.05 bound. By doing so, she runs into the multiple testing problem. For example, performing 100 individual hypothesis tests at $\alpha = 0.05$ will lead to about five mistakenly significant results by chance even if all null hypotheses are true. If these tests are independent of each other, the probability of at least one false rejection is 99.4%. This global error rate will increase the more hypotheses the researcher tests. If too many of these variables − based on the "p-value ≤ 0.05 judgement"

[1] See Edwards (2000, pp. 7–8) for more details about these mild conditions.

– are wrongly claimed as signals but are purely significant by chance, research results will not be reproducible anymore. We would probably obtain different significant variables if we used new data from the same population. Hence, we need procedures to control for the overall error, and so early multiple testing theory has developed.
Table 2.1 illustrates the outcome of m hypothesis tests $\{H_j\}_{j=1}^m$, where m_0 of those are true. The number of true and wrong null hypotheses are unknown in practical applications. The unobserved random variables V and S denote the number of wrongly and correctly

Table 2.1: Multiple testing outcome

H_0	True	False	Total
Rejected	V	S	R
Accepted	U	T	$m-r$
Total	m_0	$m-m_0$	m

Random quantities: capital letters, non-random quantities: small letters.

rejected null hypotheses, whereas the unobserved random variables U and T define the number of correctly and mistakenly not rejected hypotheses. Moreover, the random variable R is the number of total discoveries made, and it is observable. Traditional multiple testing methods try to account for the family-wise error rate (FWER) as global error rate

$$\text{FWER} = \mathbb{P}(V \geq 1), \tag{2.1}$$

which is the probability of making at least one false rejection. It is controlled in a strong sense if FWER $\leq \alpha$, under all configurations of null hypotheses. Without going into the details, there exist many proposed methods that deal with the FWER, such as the Bonferroni correction (1936), Holm's correction (1979) and the Westall-Young permutation procedure (1989).
The FWER does not distinguish between making $V = 1$ or $V = 30$ errors among all m tests, which makes it a very strict measure. Both scenarios are judged equally by the FWER, and so it only accounts for if we make errors at all or not. Since before the 1990s, the data sets were rather small, the FWER was usually applied to cases with $m \leq 20$ tests (Efron and Hastie 2016, p. 275). If we compare only 10 different treatments, it seems reasonable to control if there will be any false rejection at all. However, we discussed above that from the 90s, scientists started to have access to thousands of variables that they could test at once. For these data sets, the FWER might be too conservative, and researchers often ended up with having no or too few significant variables after the correction. Hence, the power was considerably reduced when the number of hypotheses went up. The loss of power created the urge within the research community to develop a new more loose but still reasonable error measure. Instead of controlling making errors at all, we might allow for some false rejections among many significant ones.

2.3 False discovery rate

2.3.1 Defintion and properties

Benjamini and Hochberg (1995) introduced a new error measure, the false discovery rate (FDR), and a method to control for it. Its idea is based on the false discovery proportion (FDP)

$$\text{FDP} = \frac{V}{\max(R,1)} = \begin{cases} V/R & \text{if } R \geq 1 \\ 0, & \text{if } R = 0. \end{cases} \tag{2.2}$$

The FDP is the proportion of those rejections that are wrong, except for the scenario of no rejections, where the FDP is simply zero. Although we know how many rejections R we have made, the number of false ones V remains unknown, and thus the FDP is unobservable. However its expectation, the FDR, defined as

$$\text{FDR} = \mathbb{E}[\text{FDP}], \tag{2.3}$$

can be controlled.

Definition 2.2. *A procedure controls the FDR at a pre-specified nominal level* $q \in [0,1]$ *if* $\text{FDR} \leq q$.

In contrast to the FWER, the interpretation of the FDR does not directly relate to an individual experiment because of the expectation. It holds *on average* if we hypothetically repeat the underlying experiment multiple times. Thus, the FDP for an individual experiment might or might not exceed the pre-specified threshold. So, an FDR control of 20% in an example with 100 discoveries means that we can *expect* at most 20 false discoveries among those, but this only holds *on average* and not for the individual experiment. After the FDR correction, we have the security (in expectation) that most of the significant variables of our experiment are true effects and could be reproduced by other scientists with new data of the same population.
Before we introduce specific methods for the FDR control, we present two properties that relate the FWER to the FDR:[2]

i.) Under the global null $m_0 = m$, it holds that FWER = FDR.

ii.) Under the general case, FWER $\geq$ FDR.

Under the global null scenario, FDR control provides the so-called weak FWER control. The second property implies that any method that guarantees FWER control also yields FDR control but not the other way around. Hence, procedures that control the FDR are less strict and more powerful. The more (false) null hypotheses we have, the larger S tends to be, and the larger the difference in power between the strict FWER and the FDR (Benjamini and Hochberg 1995). Therefore, the FDR has gained its popularity with the rise of large-scale data sets, when the number of tests has probably increased to more than a thousand. In such a setting, researchers are willing to accept some more false positives among their many discoveries for a growth in power. Moreover, the FDR is a more flexible measure since it deals with the number of errors instead of the occurrence of errors at all. Depending on how many false rejections the researcher is willing to accept among her total rejections, she can vary the cut-off q to weigh power at the cost of reproducibility.

[2] See Appendix A.2.1 for a proof of both properties.

2.3.2 Benjamini-Hochberg procedure

The notion of the FDR already existed in Soric (1989). However, Benjamini and Hochberg (1995) were the first who formally introduced the FDR together with their Benjamini-Hochberg (BH) procedure, which was the first method with proven FDR control.
Let p_j be the p-value corresponding to the test of the null hypothesis $H_{0,j}$ and assume that we have done m tests. The BH procedure can be described by the following steps:

i.) Sort the p-values ascendingly, that is $p_{(1)} \leq \ldots \leq p_{(m)}$.

ii.) Find j_0 being the largest index of j for which $p_{(j)} \leq \frac{j}{m}q$, i.e.

$$j_0 = \max\left\{j : p_{(j)} \leq \frac{j}{m}q\right\}. \tag{2.4}$$

iii.) Reject all null hypotheses $H_{(j)}$ that belong to the ordererd p-values $p_{(j)}$ with indices $j = 1, 2, \ldots, j_0$.

The FDR control of the BH procedure relies on the assumption that the test statistics (p-values) are mutually independent. In a least squares (LS) regression setting, the estimated coefficients $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$ (and so individual tests and p-values) are independent if and only if $\mathbf{X}^\top\mathbf{X}$ is a diagonal matrix, i.e. under the orthogonal design of no pairwise feature correlations. We summarize the FDR control of the BH procedure in the following theorem:

Theorem 2.3 (**Benjamini-Hochberg FDR control**). *Let $q \in [0, 1]$. If the m test statistics (p-values) are independent of each other, then the Benjamini-Hochberg procedure yields FDR control at level*

$$\text{FDR} = \pi_0 q \leq q, \quad \pi_0 = \frac{m_0}{m},$$

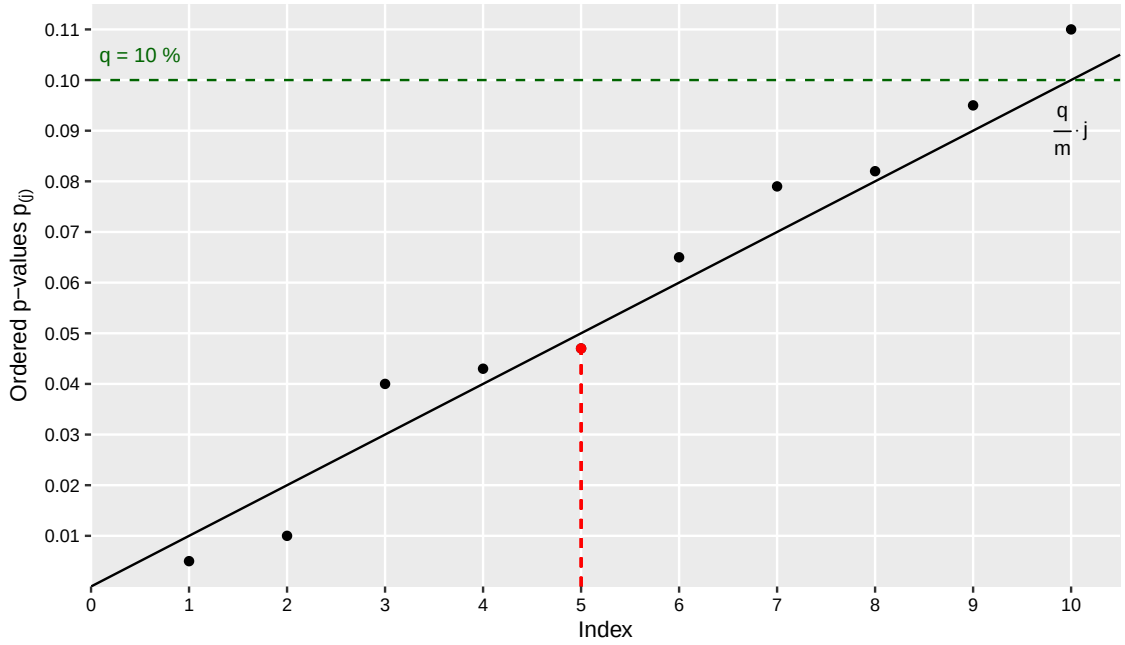
for any configuation of hypotheses. With $\pi_0 \leq 1$ being the unknown proportion of true null hypotheses, the FDR *is controlled at $\pi_0 q$.*

Note that the rejection threshold obtained in (2.4) is data-adaptive, i.e. two different sets of $\{p_{(1)}, \ldots, p_{(m)}\}$ but the same level q can result in very different thresholds. Figure 2.1 explains the determination of the index j_0 graphically by plotting the indices j against the ordered p-values $p_{(j)}$, together with a line that has the slope q/m. To find j_0, we search for the last point that is below or at the line. Then, we reject these hypotheses that belong to the points left of j_0 and j_0 itself. In the illustrated example with $m = 10$ hypotheses, the threshold index lies at $j_0 = 5$. Hence, we reject the hypotheses $H_{(1)}, \ldots, H_{(5)}$ to achieve FDR control according to Theorem 2.3.
The initial paper by Benjamini and Hochberg (1995) showed FDR control of the BH procedure under independent test statistics. In a subsequent work, Benjamini and Yekutieli (2001) proved that BH also guarantees FDR control at $\pi_0 q$ for test statistics whose joint distribution satisfies the positive regression dependence on a subset (PRDS) property, where the subset consists of the test statistics belonging to the true null hypotheses $\mathcal{H}_0$. Instead of going into the technical details of the PRDS property, we will give a simple example where the PRDS holds.[3]

[3] See Benjamini and Yekutieli (2001, p. 1168) for a formal description of the PRDS property.

Example 2.4. Let $X \sim \mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a random vector of multivariate normal test statistics each testing $H_{0,j} : \mu_j = 0$ vs. $H_{A,j} : \mu_j > 0$, with $j = 1, \ldots, m$. If we assume that for each $j \in \mathcal{H}_0$ and for all $j \neq i$ we have positive dependencies $\Sigma_{ji} \geq 0$, the joint distribution of X satisfies the PRDS property over $\mathcal{H}_0$. However, using the alternative $H_{A,j} : \mu_j \neq 0$, the PRDS property is no longer valid for the two-sided test statistics $|X_j|$.

An important property of the PRDS is that it does not rely on dependencies among the alternative test statistics. However, the correlations between the null hypotheses with other null *and* alternative hypotheses have to be known. Example 2.4 already illustrates that the PRDS property is not satisfied under the usual two-sided individual test and general correlations. In other words, the additional result by Benjamini and Yekutieli (2001) implies that only for some special correlation and testing structures, the PRDS property is satisfied, and hence the BH procedure provides FDR control under correlated test designs.[4] Even though the theoretical FDR control of the BH procedure does not hold under general correlation structures and two-sided tests, we still often observe empirical FDR control in simulations across many model settings (see upcoming simulation results in Section 3.9.)



The example consists of $m = 10$ hypotheses and a nominal level of $q = 0.1$. The red vertical line and point mark the position of the threshold index $j_0 = 5$.

Figure 2.1: Benjamini-Hochberg procedure

2.3.3 Benjamini-Yekutieli procedure

Benjamini and Yekutieli (2001) proposed a modification of the threshold (2.4) in the original BH procedure to account for arbitrary dependence structures. More precisely, the

[4] See Benjamini and Yekutieli (2001, pp. 1173–1177) for additional examples where the PRDS holds.

Benjamini-Yekutieli (BY) procedure selects the threshold index j_0 according to

$$j_0 = \max\left\{j : p_{(j)} \leq \frac{j}{m}\frac{q}{c(m)}\right\}, \quad c(m) = \sum_{j'=1}^{m}\frac{1}{j'}, \tag{2.5}$$

to control the FDR at no more than $\pi_0 q$. The following theorem summarizes the FDR control of the BY procedure:

Theorem 2.5 (**Benjamini-Yekutieli FDR control**). *Under general dependence, the BH procedure controls the FDR at level*

$$\text{FDR} \leq \pi_0 q \cdot c(m), \quad c(m) = \sum_{j'=1}^{m}\frac{1}{j'}.$$

In other words, if the BH procedure is conducted with $q/c(m)$ instead of q, it controls the FDR at $\pi_0 q$ under any dependence structure.

If the tests are independent or fulfil the PRDS property, then $c(m) = 1$. Otherwise, $c(m)$ is the m-th harmonic number, which can be approximated with a Taylor series expansion by $c(m) \approx \log(m) + 0.577$ for large m. The allowance for general dependencies comes at a price of a more conservative procedure and thus lower power compared to the BH procedure. The division by $c(m)$ decreases the threshold for which the ordered p-values are rejected. With respect to Figure 2.1, the BY method implies a less steep line with slope $\frac{q}{mc(m)}$, shifting the largest index whose ordered p-value lies below or at the line to the left. Hence, fewer null hypotheses will be rejected compared to the plain BH procedure given the same set of p-values.

2.3.4 Informal review: Improvement of power

Much research has been done after the publication of the novel paper by Benjamini and Hochberg (1995), mostly with the goal to improve the power of FDR controlling procedures. In Theorems 2.3 & 2.5, the fraction of true null hypotheses π_0 is unknown. Storey (2002) pointed out that additional information of the p-values' distribution can be used to estimate $\hat{\pi}_0$, leading to a possible power increase. Instead of fixing the error rate q and finding the threshold for which the p-values are rejected to control the FDR (as BH and BY), Storey's approach estimates the FDR for a fixed threshold region. His method controls the positive false discovery rate (pFDR) and provides q-values which are minimum estimated pFDR levels for which the hypothesis $H_{0,j}$ with p-value p_j is rejected. Storey's procedure was improved by the subsequent work of Storey et al. (2004), who linked the BH procedure to be a special case of their improved FDR controlling method. In particular, they showed that if m in the BH procedure is replaced by their estimate $\hat{\pi}_0 m$, the FDR control is asymptotically closer to q, resulting in larger power. The power of their method power depends on the quality of the estimator $\hat{\pi}_0$, which also relies on the choice of an additional tuning parameter. However, in the framework of large-scale hypothesis testing with probably many null variables, the true proportion will be close to $\pi_0 \approx 1$. Differences in power between BH and procedures with plugged-in $\hat{\pi}_0$ may not be considerably large (Efron and Hastie 2016, p. 276). Benjamini et al. (2006) proposed the two-stage Benjamini-Yekutieli-Krieger (BYK) procedure, where they first apply BH to find the number of rejected hypotheses in order to estimate $\hat{m}_0$, and then use BH again on an adjusted nominal level q^*. The BYK method is more powerful than BH, when a large number of null hypotheses are false. In simulations, it achieves good power under

positive dependencies compared to BH and the method of Storey et al. (2004). The latter had an estimated FDR that was almost double as large as desired under some correlated designs. Addressing different dependency structures for FDR controlling procedures is still a fruitful research area.

2.3.5 Empirical Bayes view

In this section, we will briefly cover the empirical Bayes perspective of the FDR to gain some additional insights on its notion. We can formulate the multiple testing problem in the Bayesian framework as a two groups model: Assume that our m hypotheses based on z-statistics are either null or non-null with prior and density

$$\text{null: } \pi_0 \text{ and } z \sim f_0 \quad \text{and} \quad \text{non-null: } \pi_1 = 1 - \pi_0 \text{ and } z \sim f_1.$$

For a region $\mathcal{Z}$ on the real line, F_0 and F_1 are the probability distributions for a feature in the null and non-null group

$$F_0(\mathcal{Z}) = \int_{\mathcal{Z}} f_0(z)dz \quad \text{and} \quad F_1(\mathcal{Z}) = \int_{\mathcal{Z}} f_1(z)dz.$$

The marginal density of $z \in \mathcal{Z}$ (mixture density) and the corresponding mixture probability distribution are defined as

$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z) \quad \text{and} \quad F(\mathcal{Z}) = \pi_0 F_0(\mathcal{Z}) + \pi_1 F_1(\mathcal{Z}).$$

Assume we observe $z \in \mathcal{Z}$ and want to know whether it belongs to the null or non-null group. This probability is obtained by applying Bayes' rule

$$\phi(\mathcal{Z}) = \mathbb{P}(\text{null}|z \in \mathcal{Z}) = \frac{\mathbb{P}(z \in \mathcal{Z}|\text{null})\mathbb{P}(\text{null})}{\mathbb{P}(z \in \mathcal{Z})} = \frac{\pi_0 F_0(\mathcal{Z})}{F(\mathcal{Z})}, \tag{2.6}$$

and we define this quantity as the Bayesian false discovery rate (BFDR). If we interpret $z \in \mathcal{Z}$ as non-null, this quantity is nothing else than the probability of a false discovery. In practice, we usually have a threshold z_c and the region $\mathcal{Z}$ will be $[z_c, \infty), (-\infty, z_c]$ or $(-\infty, z_c] \cup [z_c, \infty)$. To estimate the BFDR, we need knowledge of f_0, f_1 and π_0. Typically

- f_0 is known: Under the usual assumption $z \sim \mathcal{N}(0, 1)$, the null density is $f_0(z) = e^{-\frac{1}{2}z^2}/\sqrt{2\pi}$ and hence $F_0(\mathcal{Z}) = \Phi(\mathcal{Z})$.[5]
- π_0 is "almost" known: In large-scale hypotheses testing, the majority of the hypotheses are nulls, i.e. π_0 is near 1. It can be can either estimated $\hat{\pi}_0$ (Storey 2002; Storey et al. 2004; Benjamini et al. 2006) or simply set to $\pi_0 = 1$.
- f_1 is usually unknown a-priori.

We can make use of the properties of large-scale applications (m is large) and directly estimate the denominator in (2.6) by its empirical distribution of the m z-values

$$\hat{F}(\mathcal{Z}) = \frac{\#\{z_j \in \mathcal{Z}\}}{m},$$

which is the proportion of observed z_j values that lie inside the region $\mathcal{Z}$. This yields an estimated BFDR of

$$\widehat{\text{BFDR}}(\mathcal{Z}) = \frac{\pi_0 F_0(\mathcal{Z})}{\hat{F}(\mathcal{Z})}.$$

[5] Alternatively, the empirical null of f_0 can be used.

For large m, we expect that $\hat{F}(\mathcal{Z})$ is close to $F(\mathcal{Z})$ and $\widehat{\text{BFDR}}(\mathcal{Z})$ is close to $\text{BFDR}(\mathcal{Z})$. Lemma 2.1 & 2.2 in Efron (2010, pp. 22–24) concretise the approximation more formally: In essence, the estimated BFRD is always an upward biased estimate of BFDR, but the bias is negligible and vanishes for a large number m of z-values. In the case of independent z-values, $\widehat{\text{BFDR}}(\mathcal{Z})$ is reasonably unbiased and accurate if the expected number of z_j's falling into $\mathcal{Z}$, which is simply estimated by its count $\#\{j : z_j \in \mathcal{Z}\}$, is not too small (say larger 10). However, for correlated z-values, this does not hold anymore. While $\hat{F}(\mathcal{Z})$ remains nearly unbiased, its variability depends on the correlation of z-value pairs. More precisely, let the mean squared correlation of all z-value pairs be

$$\alpha^2 = \left[\sum_{j=1}^{m}\sum_{j\neq k}\text{cov}(z_j, z_k)^2\right] \Big/ m(m-1).$$

The variance of $\hat{F}$ can be then approximated by

$$\text{Var}(\hat{F}) \doteq b(\hat{F}(\mathcal{Z})) + \hat{\alpha}^2 c(f(z)),$$

with b and c being functions depending on the empirical distribution and the mixture density respectively. Hence, the larger the sample mean squared correlation is, the less accurate the estimate $\text{Var}(\hat{F})$ will be. A detailed discussion of the correlation effects and the estimation of $\text{Var}(\hat{F})$ is given in Chapters 7 and 8 of Efron (2010).
We can also use the empirical Bayes framework to review the BH procedure from a different angle. First, we note the relationship of z-values and p-values through $p_j = F_0(z_j)$. Moreover, assume for the sake of this example that we are interested in left-tailed p-values, i.e. we consider the region $\mathcal{Z} = (-\infty, z_c)$.[6] We can re-write the quantities that define the BH method by

$$\begin{aligned} p_{(j)} &= F_0(z_{(j)}) = F_0((-\infty, z_{(j)}]), \\ \frac{j}{m} &= \frac{\#\{z_j \leq z_{(j)}\}}{m} = \hat{F}((-\infty, z_{(j)}]). \end{aligned}$$

With this empirical Bayes notation, we can express the frequentist threshold condition of the p-values by

$$p_{(j)} \leq \frac{j}{m}q \Leftrightarrow \frac{p_{(j)}}{j/m} \leq q \Leftrightarrow \frac{F_0(z_{(j)})}{\hat{F}((-\infty, z_{(j)}])} \leq q \Leftrightarrow \widehat{\text{BFDR}}(-\infty, z_{(j)}]) \leq \pi_0 q.$$

If we set $\pi_0 = 1$ in the above equation, we achieve equivalence between the frequentist threshold and the empirical Bayes formulation of the BH. The choice of $\pi_0 = 1$, however, results also in the most conservative estimate of the BFDR. Rejecting hypotheses according to $p_{(j)} \leq (j/m)q$ in the frequentist view translates into rejecting those cases that have a small empirical Bayesian posterior probability of being in the null group in the Bayesian view. If $\widehat{\text{BFDR}}(-\infty, z_{(j)}])$ is a reasonable estimate for BFDR, we can formulate a possible Bayes' approach of the BH procedure by (see Section 2.3.2 for the frequentist version):

i.) Sort the z-values ascendingly, that is $z_{(1)} \leq \ldots \leq z_{(m)}$.

ii.) Find the critical value by

$$z_c = \sup_z \left\{\widehat{\text{BFDR}}(-\infty, z]) \leq q\right\},$$

[6] The modification to right-tailed or two-tailed p-values and regions is done analogously.

that is the largest rejection region under the constraint that the empiricial Bayes posterior of nullness is not greater than q.

iii.) Select all variables for which $z_j \leq z_c$.

The Bayesian view of BH provides a more natural and clearer picture of how the method works, whereas the frequentist version does not offer a direct interpretation of the procedure, such as one for the critical line in Figure 2.1 or why we reject all hypotheses until the last point lying below or at the line. And as already stated by Efron "It is always a good sign when a statistical procedure enjoys both a frequentist and Bayesian support, and the BH algorithm passes the test." (Efron 2010, p. 54).
We summarize the benefits and insights of viewing the FDR and BH in an empirical Bayesian framework:

- From the empirical Bayes view, the FDR theory is more an estimation problem, whereas, in the frequentist view, it is a testing problem with error control.
- From the empirical Bayes view, the BFDR and the bound q have a probability interpretation: The empirical Bayesian posterior probability that a hypothesis is falsely rejected.
- $\widehat{\text{BFDR}}(\mathcal{Z})$ is an upward biased (conservative) estimator of $\text{BFDR}(\mathcal{Z})$. Under mild conditions, the estimator is nearly unbiased in large-scale settings (large m).
- Correlation does not affect the near unbiasedness but the variance of $\hat{F}(\mathcal{Z})$ and $\widehat{\text{BFDR}}(\mathcal{Z})$.
- Since the BH algorithm controls the expected FDP, it depends on the variability of $\hat{F}(\mathcal{Z})$ to what extent we can trust our results to call them reproducible. In presence of high correlation, our confidence in $\hat{F}(\mathcal{Z}), \widehat{\text{BFDR}}(\mathcal{Z})$ and the reproducibility of the reported set of variables should be smaller. From the frequentist view, we might expect more drastic changes of the FDP if we repeat the experiment many times.

With the empirical Bayesian framework studied, we immediately recognize that the approaches by Storey (2002) and Storey et al. (2004) to estimate the FDR for a fixed threshold and their q-values have a Bayesian character.
Although this section barely scratches the surface of the extensive theory of the FDR from an (empirical) Bayesian perspective, it already led us to valuable insights about the FDR in that simple framework.

Chapter 3

Fixed-X knockoffs

Section 3.1 introduces the underlying setting and a running example on which the subsequent sections are based on. The entire procedure of the knockoff filter in detail is described in Sections 3.2 – 3.6. We continue with providing intuition on the proof of the knockoffs' FDR control in Section 3.7 and compare knockoffs with permuted features as a control group in Section 3.8. In Section 3.9, we conduct a variety of simulations before we end with a discussion about fixed-X knockoffs in Section 3.10.

3.1 Setting

A novel approach for the FDR control than the previously discussed ones is to construct dummy variables that imitate the correlation structure of the original variables. Barber and Candès (2015) propose knockoff variables as dummies that allow controlling the FDR without any assumptions on the covariates, the number of coefficients, their signal strength or information about the noise level. However, the knockoff filter works only in a low-dimensional ($n \geq p$) homoscedastic Gaussian linear model, that is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}), \tag{3.1}$$

where $\mathbf{y} \in \mathbb{R}^n$ is the response vector, $\mathbf{X} \in \mathbb{R}^{n \times p}$ the design matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ the coefficient vector and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ the noise. We further assume that the coefficient vector is sparse, only a few entries of the p are non-zero, such that only some features are truly associated with the response. We define

$$\begin{aligned} \mathcal{S}_0 &= \{j : \beta_j \neq 0\} \quad \text{as the set of the true signals and} \\ \mathcal{H}_0 &= \{j : \beta_j = 0\} \quad \text{as the set of true null variables.} \end{aligned}$$

Throughout this chapter, we will use a running example to illustrate the knockoff filter by figures. According to the linear model (3.1), we assume the following data generating process

$$\begin{aligned} &\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \quad \Sigma_{j,k} = 0.3^{|j-k|} \quad \forall j \neq k \in \{1, \dots, p\}, \quad \sigma^2 = 1, \\ &n = 600, \quad p = 200, \quad |\mathcal{S}_0| = 30, \quad |\mathcal{H}_0| = 170 \end{aligned}$$

where we randomly select 30 indices to be signals. For those, we generate their coefficients uniformly from $\beta_j \sim \text{Unif}(0.1, 0.3)$, whereas the remaining 170 ones are zero. For all computations including the simulations, we will use the `knockoff` package in the statistical software `R`, maintained by Patterson and Sesia (2020).

3.2 Step 1: Construct knockoffs

Knockoffs are constructed independently of the response. They provide no additional value in explaining $\mathbf{y}$ as long as the original variables are in the model. Unlike the true unknown null features, knockoffs are known as artifical null variables that preserve the correlation structure with the original variables. Hence, the number of selected knockoffs by the algorithm will be a valid estimate for the false discoveries among the final selected original variables. Knockoffs act as a negative control group for the real features. To investigate whether a variable j is a signal, we will compare the importance of $\mathbf{X}_j$ in explaining $\mathbf{y}$ with that of its knockoff $\tilde{\mathbf{X}}_j$.

3.2.1 Knockoff properties

We begin by creating a knockoff copy $\tilde{\mathbf{X}}_j$ for each predictor $\mathbf{X}_j$ in the model. First, we normalize each column of $\mathbf{X}$ such that $\|\mathbf{X}_j\|_2^2 = 1 \ \forall j \in \{1, \dots, p\}$, followed by the calculation of the Gram matrix $\boldsymbol{\Sigma} = \mathbf{X}^\top \mathbf{X}$. Throughout this section, we will assume that $\boldsymbol{\Sigma}$ is invertible. The knockoff variables are designed in such a way that they mimic the correlation of the original variables by satisfying

$$\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} = \boldsymbol{\Sigma}, \qquad \mathbf{X}^\top \tilde{\mathbf{X}} = \boldsymbol{\Sigma} - \operatorname{diag}\{\mathbf{s}\},$$

where $\mathbf{s} \in \mathbb{R}_+^p$. The two conditions imply the following correlation structure:

i.) $\tilde{\mathbf{X}}_j^\top \tilde{\mathbf{X}}_k = \mathbf{X}_j^\top \mathbf{X}_k \ \forall j \neq k$: two distinct knockoffs $(\tilde{\mathbf{X}}_j, \tilde{\mathbf{X}}_k)$ have the same correlation as their original counterparts $(\mathbf{X}_j, \mathbf{X}_k)$.

ii.) $\mathbf{X}_j^\top \tilde{\mathbf{X}}_k = \mathbf{X}_j^\top \mathbf{X}_k \ \forall j \neq k$: distinctly original and knockoff variables $(\mathbf{X}_j, \tilde{\mathbf{X}}_k)$ have the same correlation as the corresponding original variables $(\mathbf{X}_j, \mathbf{X}_k)$.

iii.) $\mathbf{X}_j^\top \tilde{\mathbf{X}}_j = 1 - s_j$: is the correlation between an original variable j and its knockoff.

Thereby, the construction of the knockoffs does not rely on the response $\mathbf{y}$. In summary, the knockoff matrix $\tilde{\mathbf{X}}$ differs from the data matrix $\mathbf{X}$, but it preserves the same correlation structure and also has the same correlation with the original variables.

3.2.2 Normal case: $n \geq 2p$

In the following, we will refer to $[\mathbf{X} \ \tilde{\mathbf{X}}] \in \mathbb{R}^{n \times 2p}$ as the augmented data matrix which concatenates $\mathbf{X}$ and $\tilde{\mathbf{X}}$ column-wise. The construction of the knockoff matrix $\tilde{\mathbf{X}}$ can be understood as a solution to satisfy the previously explained correlation structure. In formal terms, we choose $\tilde{\mathbf{X}}$ such that

$$[\mathbf{X} \ \tilde{\mathbf{X}}]^\top [\mathbf{X} \ \tilde{\mathbf{X}}] = \begin{bmatrix} \mathbf{X}^\top \mathbf{X} & \mathbf{X}^\top \tilde{\mathbf{X}} \\ \tilde{\mathbf{X}}^\top \mathbf{X} & \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \operatorname{diag}\{\mathbf{s}\} \\ \boldsymbol{\Sigma} - \operatorname{diag}\{\mathbf{s}\} & \boldsymbol{\Sigma} \end{bmatrix} =: \mathbf{G}, \tag{3.2}$$

where $\mathbf{G}$ is a block matrix consisting of two unique $p \times p$ correlation matrices. Before we discuss the knockoff construction further, we will briefly state why $n \geq 2p$ is considered to be the normal case. We assume that $\mathbf{X}$ has full rank, i.e. $\operatorname{rank}(\mathbf{X}) = p$. Let $\tilde{\mathbf{U}} \in \mathbb{R}^{n \times p}$ be an orthonormal matrix, consisting of p linearly independent vectors in the orthogonal complement of the column space of $\mathbf{X}$ that have been orthonormalized, i.e. $\tilde{\mathbf{U}}^\top \mathbf{X} = \mathbf{0}$. This also implies that the dimension of the column space of $\mathbf{U}$ is p. Since the column space of $\mathbf{U}$ is orthogonal to the column space of $\mathbf{X}$, it has to be a subspace of the orthogonal

complement of the column space of $\mathbf{X}$, which has dimension $n - p$. Putting all arguments together, we can conclude

$$\text{col}(\mathbf{U}) \subseteq \text{col}(\mathbf{X})^{\perp} \Rightarrow \dim(\text{col}(\mathbf{U})) \leq \dim(\text{col}(\mathbf{X})^{\perp}) \Rightarrow p \leq n - p \Rightarrow 2p \leq n.$$

The existence of $\tilde{\mathbf{U}}$ is necessary because the matrix is used in the knockoff formula presented below that guarantees the desired correlation structure.

Turning back to the knockoff construction, the knockoff matrix $\tilde{\mathbf{X}}$ does only exist – is a solution to the matrix equality (3.2) – if and only if the augmented Gramm matrix $\mathbf{G}$ is positive semidefinite ($\mathbf{G} \succeq 0$). Since $\mathbf{G}$ is a block matrix, positive semidefinitness of $\mathbf{G}$ holds if and only if its Schur complement

$$\begin{aligned}
\mathbf{S} &= \boldsymbol{\Sigma} - (\boldsymbol{\Sigma} - \text{diag}\{\mathbf{s}\})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \text{diag}\{\mathbf{s}\}) \\
&= \boldsymbol{\Sigma} - (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1} - \text{diag}\{\mathbf{s}\}\boldsymbol{\Sigma}^{-1})(\boldsymbol{\Sigma} - \text{diag}\{\mathbf{s}\}) \\
&= \boldsymbol{\Sigma} - (\boldsymbol{\Sigma} - \text{diag}\{\mathbf{s}\}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma} - \text{diag}\{\mathbf{s}\} + \text{diag}\{\mathbf{s}\}\boldsymbol{\Sigma}^{-1}\,\text{diag}\{\mathbf{s}\}) \\
&= \boldsymbol{\Sigma} - \boldsymbol{\Sigma} + \text{diag}\{\mathbf{s}\} + \text{diag}\{\mathbf{s}\} - \text{diag}\{\mathbf{s}\}\boldsymbol{\Sigma}^{-1}\,\text{diag}\{\mathbf{s}\} \\
&= 2\,\text{diag}\{\mathbf{s}\} - \text{diag}\{\mathbf{s}\}\boldsymbol{\Sigma}^{-1}\,\text{diag}\{\mathbf{s}\}
\end{aligned}$$

is positive semidefinite, i.e. $\mathbf{G} \succeq 0 \Leftrightarrow \mathbf{S} \succeq 0$ (see Thoerem A.2). Rewriting the Schur complement $\mathbf{S}$ into a different block matrix $\mathbf{G}^*$ and applying the Schur complement again but with respect to $\text{diag}\{\mathbf{s}\}$, we obtain the conditions on $\mathbf{s}$ for the positive semidefinitness of the augmented Gramm matrix $\mathbf{G}$

$$\begin{aligned}
\mathbf{G}^* := \begin{bmatrix} \boldsymbol{\Sigma} & \text{diag}\{\mathbf{s}\} \\ \text{diag}\{\mathbf{s}\} & 2\,\text{diag}\{\mathbf{s}\} \end{bmatrix} \succeq 0 \quad &\iff \quad \begin{aligned} &2\,\text{diag}\{\mathbf{s}\} \succeq \mathbf{0} \\ &\boldsymbol{\Sigma} - \text{diag}\{\mathbf{s}\}(2\,\text{diag}\{\mathbf{s}\})^{-1}\,\text{diag}\{\mathbf{s}\} \succeq \mathbf{0}. \end{aligned} \\
&\iff \quad \begin{aligned} &\text{diag}\{\mathbf{s}\} \succeq \mathbf{0} \\ &2\boldsymbol{\Sigma} - \text{diag}\{\mathbf{s}\} \succeq \mathbf{0}. \end{aligned}
\end{aligned} \tag{3.3}$$

The positive semidefinitess of $\mathbf{S}$ also allows a Cholesky decomposition to $\mathbf{S} = \mathbf{C}^\top\mathbf{C} = 2\,\text{diag}\{\mathbf{s}\} - \text{diag}\{\mathbf{s}\}\boldsymbol{\Sigma}^{-1}\,\text{diag}\{\mathbf{s}\}$. Putting it all together, the following theorem summarizes how the knockoff matrix can be calculated.

Theorem 3.1 (**Knockoff construction**). *Let $n \geq 2p$ and choose $\mathbf{s}$ satisfying $2\boldsymbol{\Sigma} \succeq \text{diag}\{\mathbf{s}\} \succeq 0$. Assume that the orthonormal matrix $\tilde{\mathbf{U}} \in \mathbb{R}^{n\times p}$ exists and is orthogonal to the span of $\mathbf{X}$ such that $\tilde{\mathbf{U}}^\top\mathbf{X} = \mathbf{0}$, and $\mathbf{C} \in \mathbb{R}^{p\times p}$ is obtained by the Cholesky decomposition of $\mathbf{S} = \mathbf{C}^\top\mathbf{C}$. Then, if we calculate the knockoff matrix by*

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

it will lead to the desired correlation structure (3.2).

(See Appendix A.2.2 for a proof.)
It remains unclear how to choose $\mathbf{s} \in \mathbb{R}^p_+$ among all those satisfying $2\boldsymbol{\Sigma} \succeq \text{diag}\{\mathbf{s}\}$. In other words, even if the conditions for $\mathbf{s}$ hold such that the desired correlation structure exists, the power of the knockoff filter will still depend on the specific choice of $\mathbf{s}$. As we will see later, the knockoff procedure will be more successful in detecting signals the larger the difference between the original signal variable and its knockoff copy is. Hence, we choose s_j such that their pairwise correlation $\tilde{\mathbf{X}}_j^\top\mathbf{X}_j = 1 - s_j$ is close to zero, which is achieved by setting the components $\{s_j\}$ to be as large as possible. Barber and Candès (2015) suggest two approaches to determine $\mathbf{s}$:

i.) *Equi-correlated knockoffs:* This approach is based on choosing the same entry $s = s_j$ such that *all* pairwise correlations $(\mathbf{X}_j, \tilde{\mathbf{X}}_j)$ are equal. Then, in the spirit of (3.3), pick $s_j = 2\lambda_{\min}(\boldsymbol{\Sigma}) \wedge 1, \ \forall \ j$.[1] This choice minimizes the correlation between $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ under the equi-correlation constraint.

ii.) *SDP knockoffs:* The semidefinite programming (SDP) approach is based on the convex optimization problem

$$\min \|\text{diag}\{\boldsymbol{\Sigma}\} - \mathbf{s}\|_1 \quad \text{s.t.} \quad \begin{array}{l} \text{diag}\{\mathbf{s}\} \succeq \mathbf{0} \\ 2\boldsymbol{\Sigma} - \text{diag}\{\mathbf{s}\} \succeq \mathbf{0}, \end{array}$$

which minimizes the sum of absolute values of the pairwise correlations $(\mathbf{X}_j, \tilde{\mathbf{X}}_j)$ under the constraint that positive semidefinitness of the Gram matrix $\mathbf{G}$ is ensured.[2]

Although SDP knockoffs are computationally more expensive than equi-correlated knockoffs, they are more flexible in their choice and result in larger power. SDP is also the default method of the `knockoff`'s package fixed-X knockoff construction (Patterson and Sesia 2020).

3.2.3 Special case: $p \leq n < 2p$

In the low-dimensional case $n < 2p$, it is not possible to find an $\tilde{\mathbf{U}}$ such that $\tilde{\mathbf{U}}^\top \mathbf{X} = \mathbf{0}$. Hence, we cannot construct $\tilde{\mathbf{X}}$ according to Theorem 3.1. Barber and Candès (2015) propose two ways to deal with this scenario.
First, we can use a data augmentation technique when σ^2 can be estimated precisely. For example, in the Gaussian model, we can use the residual sum of squares of the full model multiplied with a correction factor as an estimate for the error variance. Then, to increase the sample size, we augment the design matrix $\mathbf{X}$ by $n - p$ additional rows of zeros and generate $n - p$ responses $\mathbf{y}'$ independently from $\mathcal{N}(0, \hat{\sigma}^2)$. Our augmented linear model with p variables and $2p$ observations follows approximately

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}' \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{X} \\ \mathbf{0} \end{bmatrix} \boldsymbol{\beta}, \sigma^2 \mathbf{I} \right).$$

The new observations do not provide additional value for the estimation of $\boldsymbol{\beta}$ but enable us to be in the normal case now, and we can construct knockoff variables as described in the $n \leq 2p$ setting. This approach will yield valid results if we can accurately estimate σ^2, i.e. when the gap $n - p$ is large.
The second approach does not rely on an estimate of σ^2. It is based on the idea of testing only a subset of $n-p$ predictors. More precisely, we select an index set $J \subseteq \{1, \ldots, p\}$ with cardinality $|J| = 2p - n$. For these variables $\mathbf{X}_j$ with $j \in J$, we construct exact duplicates $\tilde{\mathbf{X}}_j = \mathbf{X}_j$. This is achieved by setting the corresponding entries of $\mathbf{s}$ to zero such that we have a perfect knockoff-variable correlation. For the remaining $n - p$ variables with $j \notin J$, the knockoffs are constructed as usual with non-zero entries in $\mathbf{s}$ for those. Hence, we can find $\tilde{\mathbf{U}} \in \mathbb{R}^{n \times (n-p)}$ such that $\tilde{\mathbf{U}}^\top \mathbf{X} = \mathbf{0}$ and construct $\tilde{\mathbf{X}}$ as in Theorem 3.1. Although the FDR is controlled, the method will have no power in finding signals among the duplicated variables.[3]

[1] $\lambda_{\min}(\boldsymbol{\Sigma})$ denotes the minimal eigenvalue of $\boldsymbol{\Sigma}$.
[2] See Boyd and Vandenberghe (2004, pp. 168 ff.) for more information about SDP.
[3] The knockoff filter detects signals better the more orthogonal original variables and their knockoffs are. In the duplicated cases, they have perfect correlation.

As a remedy for this problem, we can cycle through disjoint index sets of duplicates. We explain the intuition for $n = 3p/2$: Split the p variables into two disjoint sets $\{J_1, J_2\}$, with cardinality $p/2$ each. Select J_1 as duplicates and create knockoffs for variables in J_2 before applying the knockoff method as described above. Then, switch the roles of J_1 and J_2 and apply the knockoff filter again. In each round, we control the FDR at $q/2$, so the overall FDR is still under q, with good power to detect signal variables. Although this technique works, the `knockoff` package applies the data augmentation approach in the special case $p \leq n < 2p$.

3.3 Step 2: Measure and compare variable importance

After constructing the p knockoffs, we can fit a traditional variable selection method to the augmented data matrix $[\mathbf{X}\ \tilde{\mathbf{X}}] \in \mathbb{R}^{n \times 2p}$. We will use the Lasso model (Tibshirani 1996) as a popular example throughout this section. It extends the usual LS regression by an additional ℓ_1-norm penalty of the coefficient vector

$$\hat{\boldsymbol{\beta}}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\} \quad \lambda \geq 0. \tag{3.4}$$

Lasso has the property to shrink coefficients towards or to exactly zero, usually resulting in sparse models. Depending on the assumptions placed on $\mathbf{X}$, Lasso is known to be consistent for variable selection or screening.[4] It is "empirically evident" that Lasso selects the majority of the relevant variables with a large enough signal and some additional null variables (Bühlmann and van de Geer 2011, pp. 17–18, pp. 23–24). In the case of our augmented data matrix, regressing $\mathbf{y}$ on $[\mathbf{X}\ \tilde{\mathbf{X}}]$ will result in a Lasso coefficient vector of dimension $\hat{\boldsymbol{\beta}}(\lambda) \in \mathbb{R}^{2p}$.

Having chosen an estimation technique, the goal is to measure the importance of a variable $\mathbf{X}_j$ and its knockoff $\tilde{\mathbf{X}}_j$ in explaining the response with a so-called importance statistics Z_j and $\tilde{Z}_j$ respectively. The importance measure should be large for relevant variables and close to zero for null variables. Two possible importance statistics for Lasso are:

i.) The absolute coefficient sizes: $Z_j = |\hat{\beta}_j(\lambda)|$ and $\tilde{Z}_j = |\hat{\beta}_{j+p}(\lambda)|$, for a fixed λ and for $j = 1, \ldots, p$.

ii.) The largest tuning parameter λ on the regularization path for which variable (or knockoff) j enters the model first, i.e. having a non-zero coefficient:
$Z_j = \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}$ and $\tilde{Z}_j = \sup\{\lambda : \hat{\beta}_{j+p}(\lambda) \neq 0\}$ for $j = 1, \ldots, p$.

In general, by fitting a variable selection method on the data $([\mathbf{X}\ \tilde{\mathbf{X}}], \mathbf{y})$, we obtain a $2p$-dimensional vector with importance measures

$$(Z_1, \ldots, Z_p, \tilde{Z}_1, \ldots, \tilde{Z}_p) = z([\mathbf{X}\ \tilde{\mathbf{X}}], \mathbf{y}).$$

A variety of statistics can be used, and the choice will depend on the reader's conception of importance. The statistic z has only to satisfy the *fairness requirement*. Before we explain this property, we have to introduce the swap operator.

Let S be any index subset of the variables $S \subseteq \{1, \ldots, p\}$. Then, the matrix $[\mathbf{X}\ \tilde{\mathbf{X}}]_{\text{swap}(S)}$ is obtained by swapping the columns $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ of the augmented matrix $[\mathbf{X}\ \tilde{\mathbf{X}}]$, $\forall j \in S$.

[4] Selection: Lasso selects the true model. Screening: The Lasso model contains the true predictors but also additional superfluous ones.

For example, with $p = 4$ variables $[\mathbf{X}\ \tilde{\mathbf{X}}] = [\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4, \tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, \tilde{\mathbf{X}}_3, \tilde{\mathbf{X}}_4]$ and subset $S = \{2, 3\}$, we obtain

$$[\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4, \tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, \tilde{\mathbf{X}}_3, \tilde{\mathbf{X}}_4]_{\text{swap}(\{2,3\})} = [\mathbf{X}_1, \tilde{\mathbf{X}}_2, \tilde{\mathbf{X}}_3, \mathbf{X}_4, \tilde{\mathbf{X}}_1, \mathbf{X}_2, \mathbf{X}_3, \tilde{\mathbf{X}}_4].$$

Turning back to the *fairness requirement*, it states that swapping $\mathbf{X}_j$ with $\tilde{\mathbf{X}}_j$ before calculating $(Z_1, \ldots, Z_p, \tilde{Z}_1, \ldots, \tilde{Z}_p)$ has the only effect of swapping Z_j and $\tilde{Z}_j$. More general, for any subset $S \subseteq \{1, \ldots, p\}$, we require

$$\begin{aligned} z([\mathbf{X}\ \tilde{\mathbf{X}}]_{\text{swap}(S)}, \mathbf{y}) &= z([\mathbf{X}\ \tilde{\mathbf{X}}], \mathbf{y})_{\text{swap}(S)} \\ &= (Z_1, \ldots, Z_p, \tilde{Z}_1, \ldots, \tilde{Z}_p)_{\text{swap}(S)}. \end{aligned}$$

In the spirit of the example above, swapping variable $\{2, 3\}$ with their knockoffs, changes the computed vector of importance measures in the following manner

$$z([\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4, \tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, \tilde{\mathbf{X}}_3, \tilde{\mathbf{X}}_4]_{\text{swap}\{2,3\}}, \mathbf{y}) = (Z_1, \tilde{Z}_2, \tilde{Z}_3, Z_4, \tilde{Z}_1, Z_2, Z_3, \tilde{Z}_4).$$

After computing the importance measures, we want to combine them for each variable $(Z_j,\ \tilde{Z}_j)$ into a single score via a function w

$$W_j = f_j(Z_j, \tilde{Z}_j) =: w_j([\mathbf{X}\ \tilde{\mathbf{X}}], \mathbf{y}), \quad \forall j \in \{1, \ldots, p\}$$

to assess whether and to which degree a variable or its knockoff is more important for the explanation of the response $\mathbf{y}$.[5] Similar to the importance statistics, large positive values are an indication that variable j is a signal, and hence evidence against $H_0 : \beta_j = 0$. In other words, large positive values imply that $\mathbf{X}_j$ is more important than its knockoff counterpart. In contrast, for a null variable, W_j is symmetrically distributed and so equally likely to take positive or negative values.[6] Two simple examples to combine the importance measures into a score are

i.) $W_j = Z_j - \tilde{Z}_j$.

ii.) $W_j = Z_j \vee \tilde{Z}_j \cdot \text{sign}(Z_j - \tilde{Z}_j)$, with $W_j = 0$ if $Z_j = \tilde{Z}_j$.

In this section, we will use ii.) for Z_j and W_j as explanatory examples. The score W_j will be positive if the original variable $\mathbf{X}_j$ is selected before its knockoff $\tilde{\mathbf{X}}_j$ in the model, and negative if the knockoff enters the Lasso model first. Hence, a large value of W_j provides evidence that variable j is part of the underlying truth. Again, the reader has a large flexibility in choosing the score function. However, there are also some general requirements that the score statistic W_j has to satisfy so that the knockoff filter provides valid FDR control. In general, the statistics $\mathbf{W}([\mathbf{X}\ \tilde{\mathbf{X}}], \mathbf{y}) \in \mathbb{R}^p$ have to satisfy the following two properties:

i.) The statistics $\mathbf{W}$ satisfy the *sufficiency property* if they are only a function of the Gram matrix $\mathbf{G}$ and the inner products of the features with the response

$$\mathbf{W} = w_j(\mathbf{G}, [\mathbf{X}\ \tilde{\mathbf{X}}]^\top \mathbf{y}) \tag{3.5}$$

for some function w_j.

[5] We use here two different functions f_j and w_j to underline that the score can be either written as a function of $(Z_j, \tilde{Z}_j)$ or directly of $([\mathbf{X}\ \tilde{\mathbf{X}}], \mathbf{y})$.

[6] We will explain the behaviour of the scores corresponding to null features in Section 3.4.

ii.) The score functions $\mathbf{W}$ satisfy the *antisymmetry property* if changing the columns $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ only changes the sign of W_j. So, for any subset $S \subset \{1, \dots, p\}$

$$w_j([\mathbf{X}\ \tilde{\mathbf{X}}]_{\text{swap}(S)}, \mathbf{y}) = \begin{cases} w_j([\mathbf{X}\ \tilde{\mathbf{X}}], \mathbf{y}), & j \notin S \\ -w_j([\mathbf{X}\ \tilde{\mathbf{X}}], \mathbf{y}), & j \in S. \end{cases} \tag{3.6}$$

This is an immediate result following from the *fairness requirement* on z.
Likewise, if we express the score as a function of the importance statistics, then f_j must be *antisymmetric*, i.e. $f_j(Z_j, \tilde{Z}_j) = -f_j(\tilde{Z}_j, Z_j)$.

It should be noted that it is perfectly valid to directly define $W_j = w_j([\mathbf{X}\ \tilde{\mathbf{X}}], \mathbf{y})$ with properties i.) and ii.) without the intermediate step of defining $(Z_j, \tilde{Z}_j)$. Considering that this work aims to explain the knockoff filter in great detail, we will often elaborate on both, the feature importance statistics and the scores.
The two examples for Z_j and W_j given above obey the required conditions respectively. However, the researcher must take care if she chooses the Lasso coefficients to be part of the score. While $Z_j = |\hat{\beta}_j(\lambda)|$, the absolute Lasso coefficients for a *fixed* λ, are valid statistics, tuning λ by cross-validation (CV) would not satisfy the sufficiency requirement anymore. Too see this, consider the following toy example.

Example 3.2. Suppose we want to perform 2-fold CV to obtain the optimal λ for Lasso, and we take as score $W_j = |\hat{\beta}_j(\lambda_{\text{CV}})| - |\hat{\beta}_{j+p}(\lambda_{\text{CV}})|$. We would start by randomly splitting the augmented data into two subsamples $([\mathbf{X}\ \tilde{\mathbf{X}}]_{\mathbf{n_1}}, \mathbf{y}_{\mathbf{n_1}})$ and $([\mathbf{X}\ \tilde{\mathbf{X}}]_{\mathbf{n_2}}, \mathbf{y}_{\mathbf{n_2}})$. Then to run Lasso for each CV round, we need $[\mathbf{X}\ \tilde{\mathbf{X}}]_{\mathbf{n_1}}^\top \mathbf{y}_{\mathbf{n_1}}$ and $[\mathbf{X}\ \tilde{\mathbf{X}}]_{\mathbf{n_2}}^\top \mathbf{y}_{\mathbf{n_2}}$ respectively, both of which are $2p$-dimensional, and neither of which can be inferred from the $2p$-dimensional feature-response inner products $[\mathbf{X}\ \tilde{\mathbf{X}}]^\top \mathbf{y}$, which is what the sufficiency property requires.

3.4 Exchangeability properties

Before we continue with the description of the knockoff procedure, we will introduce the pairwise exchangeability lemma $(Z_j, \tilde{Z}_j)$ of the nulls and the coin-flipping lemma of the scores W_j. They are fundamental for the knockoff filter to work, that is to control the FDR. Both lemmas are direct consequences of two exchangeability properties of our constructed sample $([\mathbf{X}\ \tilde{\mathbf{X}}], \mathbf{y})$:

i.) *Pairwise exchangeability for the features (PEF):* The Gram matrix $\mathbf{G}$ remains unchanged under any columnwise exchange of the original variables with their knockoffs. That is, for any $S \subset \{1, \dots, p\}$,

$$[\mathbf{X}\ \tilde{\mathbf{X}}]_{\text{swap(S)}}^\top [\mathbf{X}\ \tilde{\mathbf{X}}]_{\text{swap(S)}} = [\mathbf{X}\ \tilde{\mathbf{X}}]^\top [\mathbf{X}\ \tilde{\mathbf{X}}]. \tag{3.7}$$

ii.) *Pairwise exchangeability for the response (PER):* The distribution of the inner product $[\mathbf{X}\ \tilde{\mathbf{X}}]^\top \mathbf{y}$ is invariant to any swap of original null variables and their knockoffs. That is, for any $S \subset \mathcal{H}_0$,

$$[\mathbf{X}\ \tilde{\mathbf{X}}]_{\text{swap(S)}}^\top \mathbf{y} \overset{d}{=} [\mathbf{X}\ \tilde{\mathbf{X}}]^\top \mathbf{y}. \tag{3.8}$$

In other words, the specific knockoff construction (3.2) leads to the PEF and PER property. Thinking more about the PER provides additional intuition: Assume we try to detect signals by their marginal correlation strength with the response $|\mathbf{X}_j^\top \mathbf{y}|$. Since knockoffs are

constructed independently of the response, they are artifical null variables. Moreover, the PER implies that the marginal correlation of a null variable $j \in \mathcal{H}_0$ is equally distributed as this of its knockoff

$$\mathbf{X}_j^\top \mathbf{y} \overset{d}{=} \tilde{\mathbf{X}}_j^\top \mathbf{y},$$

so we can use the knockoffs as negative controls. If we want to detect a signal with some certainty, we hope to observe $|\mathbf{X}_j^\top \mathbf{y}| > |\tilde{\mathbf{X}}_j^\top \mathbf{y}|$. The inner product $\tilde{\mathbf{X}}_j^\top \mathbf{y}$ is a valid comparison because it has the same distribution as the corresponding inner product for a original null feature $\mathbf{X}_j^\top \mathbf{y}$ (Barber and Candès 2019). To summarize this intuition more generally, from the PER and the PEF, the fundamental pairwise exchangeability of the importance statistics for the nulls can be derived:

Lemma 3.3 (**Pairwise exchangeability of the importance statistics**). *The importance statistics of the null variables are pairwise exchangeable. That is, for any*

$$\begin{aligned} j \in \mathcal{H}_0 &\implies (Z_j, \tilde{Z}_j) \overset{d}{=} (\tilde{Z}_j, Z_j), \\ \textit{more generally} \quad S \subset \mathcal{H}_0 &\implies (\mathbf{Z}, \tilde{\mathbf{Z}})_{\text{swap}(S)} \overset{d}{=} (\mathbf{Z}, \tilde{\mathbf{Z}}). \end{aligned}$$

So, we can change the Z_j and $\tilde{Z}_j$'s of the null variables and the joint distribution

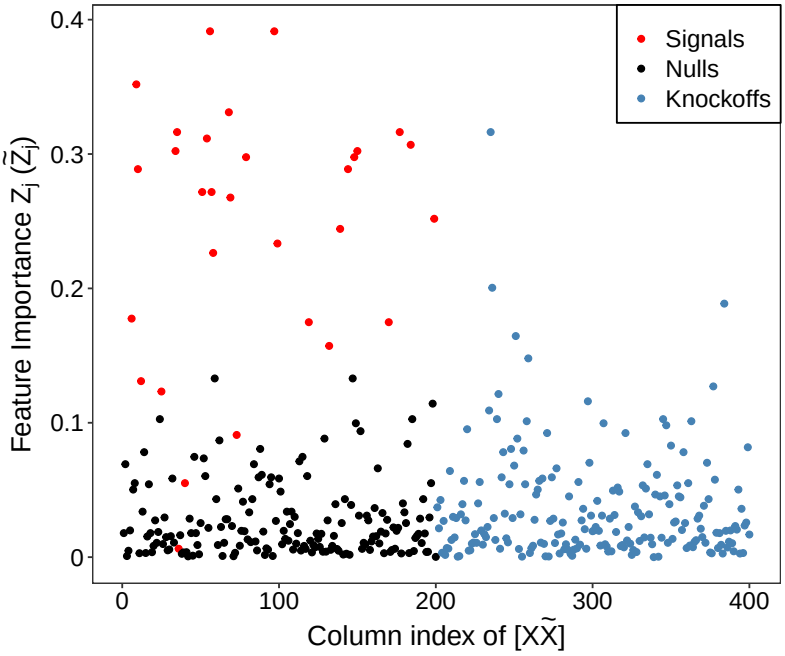
$$(Z_1, \ldots, Z_p, \tilde{Z}_1, \ldots, \tilde{Z}_p)$$

remains unchainged. Figure 3.1 displays the feature importance vector for our running example. We clearly see large values for the importance statistics of signal features (red), i.e. they enter the model even with a very large penalty. In comparison, the feature importance statistics of the null variables (black) are mostly near zero with a few notable large values.[7] The most important observation is that the $\tilde{Z}_j$'s of the knockoffs look very similar to those of the nulls. Although this is not a valid justification of their pairwise exchangeability, it visually shows that they approximately have the same behaviour. In simple terms, the similarity between the nulls and the knockoffs qualifies the $\tilde{Z}_j$'s to be a valid control group. The joint distribution of the feature importance statistics is very complicated and unknown for most modern selection procedures, which makes it rather difficult for the reader to directly work with it. For our running example, the joint distribution of the largest λ's at which variables enter the model is unknown. Even for a simpler importance measure, e.g. the absolute coefficient estimates of Lasso, the joint distribution is not well-known and struggles from difficulties such as point-mass at zero (Bühlmann et al. 2014). Fortunately, there is an attractive remedy for that. Using the exchangeability lemma of the nulls Z_j and $\tilde{Z}_j$ together with the antisymmetry property of W_j, the coin-flipping lemma can be derived.

Lemma 3.4 (**Coin-flipping lemma**). *Conditioned on* $(|W_1|, \ldots, |W_p|)$*, the signs of the null scores* W_j*,* $j \in \mathcal{H}_0$*, are i.i.d. coin flips. Hence,* $\forall j \in \mathcal{H}_0$*,* $\text{sign}(W_j) \overset{i.i.d.}{\sim} \text{Bern}(1/2)$.

Note that by null scores, we mean the scores corresponding to null variables in $\mathcal{H}_0$ and not that a score has a value of zero. The simple distribution of the sign of the null scores W_j as independent coin flips will allow us to estimate the FDP. We will use and elaborate on the coin-flipping lemma in the next section.

[7] We will explain the intuition behind the few large nulls in Section 3.8.



Importance statistic: $\sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}$.

Figure 3.1: Graphical representation of $(Z_1, \ldots, Z_p, \tilde{Z}_1, \ldots, \tilde{Z}_p)$

To summarize the gist of this section in simple words: By the specific construction of the knockoffs and their correlations, the importance statistics of the null variables are pairwise exchangeable. As a consequence, we can use knockoffs as valid control variables and compare their feature importance with that of an original variable. However, working with the joint distribution of the feature importances directly is not possible due to its complicated and unknown structure. But the coin-flipping lemma, a direct consequence of the pairwise exchangeability lemma, allows us to estimate the FDP in an elegant way by investigating the scores W_j.

3.5 Step 3: Calculate the data-driven threshold

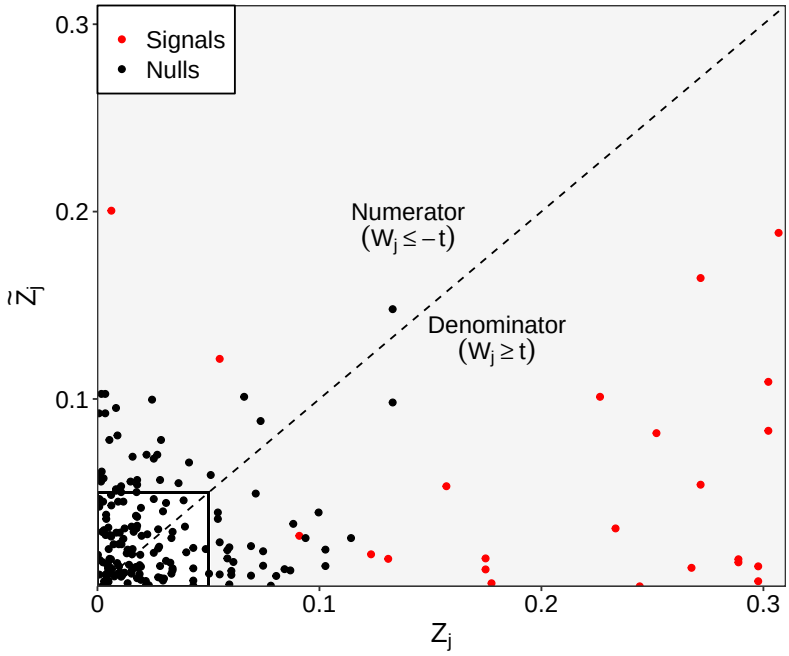
Once we have computed the score statistic for each variable, we want to select all features with large enough positive scores $W_j \geq t$. The value t defines the boundary of how more important a feature must be compared to its knockoff copy to be interpreted as a signal variable. This threshold will be calibrated by the given data such that our model selection procedure controls the FDR. Let $q \in [0, 1]$ be the desired FDR, the data-adaptive threshold T is then derived by

$$T = \min\left\{t \in \mathcal{W}_+ : \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q\right\}, \quad \mathcal{W}_+ = \{|W_j| : |W_j| > 0,\ j = 1, \ldots, p\},$$

with the convention $T = +\infty$ if this set is empty and $\mathcal{W}_+$ consisting of all unique non-zero absolute scores W_j. To ease the notation, we introduce two sets and re-write the threshold

$$\begin{aligned} \mathcal{S}^+(t) &\coloneqq \{j : W_j \geq t\} \\ \mathcal{S}^-(t) &\coloneqq \{j : W_j \leq -t\} \end{aligned} \quad \Rightarrow \quad T = \min\left\{t \in \mathcal{W}_+ : \frac{|\mathcal{S}^-(t)|}{|\mathcal{S}^+(t)| \vee 1} \leq q\right\}. \tag{3.9}$$

Before discussing the fraction in (3.9), we will introduce a graphical representation that will help us to understand this term further. Figure 3.2 depicts all importance measure pairs $(Z_j, \tilde{Z}_j)$ of our running example. Signal variables are visualized as red dots, whereas null features are pictured as black dots. As already mentioned, a positive score $W_j > 0$ implies that the original variable is more important than its knockoff ($Z_j > \tilde{Z}_j$). In our Lasso example, this means that the original variable is selected before its knockoff. All



The threshold $t = 0.05$ is illustraed by the square box. Not all signals are depticed because some of them had very large Z_j's.

Figure 3.2: Graphical representation of the pairs $(Z_j, \tilde{Z}_j)$

points below (above) the 45° line have positive (negative) scores W_j. Moreover, the null variables (black dots) are roughly symmetrically distributed around the 45° line. This symmetry of the null scores is a direct consequence of our coin flipping lemma. According to the lemma, it is equally likely to see positive or negative null scores, i.e. that either the original null variable or its knockoff enters the model first. The non-nulls, however, have the strong tendency to lie below the diagonal ($W_j > 0$), indicating a larger importance measure for Z_j than for its control variable $\tilde{Z}_j$. Fixing a threshold t, the numerator $\mathcal{S}^-(t)$ of (3.9) simply counts *all points* in the grey area above the 45° line, while the dominator $\mathcal{S}^+(t)$ is the number of *all* points in the grey area below the line.
If we select all variables $\{j : W_j \geq t\}$ at any given t, this fraction estimates the FDP because of

$$\begin{aligned} \mathrm{FDP}(t) = \frac{|\mathcal{S}^+(t) \cap \mathcal{H}_0|}{|\mathcal{S}^+(t)| \vee 1} &\approx \frac{|\mathcal{S}^-(t) \cap \mathcal{H}_0|}{|\mathcal{S}^+(t)| \vee 1} \\ &\leq \frac{|\mathcal{S}^-(t)|}{|\mathcal{S}^+(t)| \vee 1} =: \widehat{\mathrm{FDP}}(t). \end{aligned} \tag{3.10}$$

The first term defines the FDP since it is the number of selected null variables over all selections that we have made. However, in practical applications, the numerator will be unknown because we have no information about which of our discoveries are true signals

and nulls. We have made the visual observation in Figure 3.2 that the scores of the null variables are approximately symmetrically distributed around the diagonal line, which is a consequence of the coin-flipping lemma. In fact, it can be shown that, for any t, the coin-flipping lemma implies

$$|\mathcal{S}^+(t) \cap \mathcal{H}_0| \stackrel{d}{=} |\mathcal{S}^-(t) \cap \mathcal{H}_0|. \tag{3.11}$$

Hence, the number of black dots in the lower grey region $|\mathcal{S}^+(t) \cap \mathcal{H}_0|$ will be approximately equal to the number of null variables in the upper grey region $|\mathcal{S}^-(t) \cap \mathcal{H}_0|$. We use this approximation in the second step in (3.10). Since this term is still unknown to the reader in practice, we bound the numerator by the total number of points in the upper grey area. The last term is known and serves as an estimate for the FDP. The number of variables for which their knockoffs are notably more important, i.e. large negative value below $-t$, are an estimate of the false discoveries in the selected model.

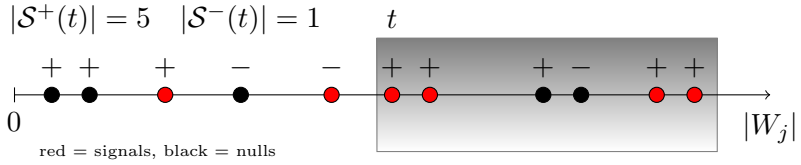


Figure 3.3: FDP estimation for a given t

Figure 3.3 shows a simplified FDP estimation by depicting the ordered absolute scores. The grey box includes all absolute scores with a value of at least t. Within that box, we count the number of points with a negative sign $|\mathcal{S}^-(t)|$ and positive sign $|\mathcal{S}^+(t)|$ respectively. Taking their ratio leads to an FDP estimate for a given t. In that example, the FDP would be 1 in 5. Note that the point with a negative sign is *not* part of the final model. But since we know that $\mathbb{P}(W_j \geq t) = \mathbb{P}(W_j \leq -t)$ for the nulls, it estimates that among the selected variables, there will be one null feature, which is in fact true. More generally, we can interpret $|\mathcal{S}^+(t)|$ as the selected set of variables for a given threshold t and $|\mathcal{S}^-(t)|$ as an estimate of the number of false discoveries within that model.
The upper bound in (3.10) and so the FDP estimate will usually be tight when the (true) coefficients are large enough in size. Then, most signals will have a clearly larger importance score than their knockoffs (i.e. large $W_j \geq 0$), and so most of the red dots in Figure 3.2 will be below the diagonal. Hence, the difference $|\mathcal{S}^-(t)| - |\mathcal{S}^-(t) \cap \mathcal{H}_0|$ will be small and the upper bound not too conservative.
We have broken down why the fraction in (3.9) is an estimate for the FDP, and we have also graphically explained its estimation for a given t. However, there will be more than one t with an estimated FDP below our specified level usually. Among all thresholds that control the estimated FDP, the knockoff procedure selects the smallest one, i.e. $T = \min\{t \in \mathcal{W}^+ : \widehat{\text{FDP}}(t) \leq q\}$. This guarantees the most signal detections and so the largest power while still controlling the estimated FDP.

3.6 Step 4: Selection rule and (modified) FDR control

After we have calculated the data-adaptive T from (3.9), we use the selection rule $\hat{\mathcal{S}} = \{j : W_j \geq T\}$ to obtain the final model. This selection procedure controls the modified FDR, which is summarized by the following theorem:

Theorem 3.5 (**Knockoff modified FDR control**). *Let $\hat{\mathcal{S}} = \{j : W_j \geq T\}$ be the selected model by the knockoff filter with knockoffs satisfying the structure in (3.2) and a score statistic W_j with sufficiency and antisymmetry property. Then, for any $q \in [0, 1]$, the knockoff method guarantees modified FDR control*

$$\mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}| + q^{-1}}\right] \leq q,$$

where $\mathbf{X}$ and $\tilde{\mathbf{X}}$ are fixed and the expectation refers to the noise.

We can also use the slightly more conservative knockoff+ method, which is based on the data-dependent threshold

$$T^+ = \min\left\{t \in \mathcal{W}_+ : \frac{1 + |\mathcal{S}^-(t)|}{|\mathcal{S}^+(t)| \vee 1} \leq q\right\}, \tag{3.12}$$

to control the FDR instead of its modified version. The difference arises by increasing the numerator by one, which behaves like adding one false discovery to the estimator, making T^+ always larger or equal to the threshold T. The following theorem summarizes the FDR control of the knockoff+ method:

Theorem 3.6 (**Knockoff+ FDR control**). *Let $\hat{\mathcal{S}} = \{j : W_j \geq T^+\}$ be the selected model by the knockoff+ filter with knockoffs satisfying the structure in (3.2) and a score statistic W_j with sufficiency and antisymmetry property. Then, for any $q \in [0, 1]$, the knockoff+ method guarantees FDR control*

$$\mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}| \vee \ 1}\right] \leq q,$$

where $\mathbf{X}$ and $\tilde{\mathbf{X}}$ are fixed and the expectation refers to the noise.

Usually, the difference between the modified FDR and FDR will be small when the set $\hat{\mathcal{S}}$ is large. So, in many applications, the knockoff filter will also yield an actual FDR below q, even though it is not theoretically justified. However, when we only make a few discoveries ($\hat{\mathcal{S}}$ is small), which can happen if the variables are highly correlated, the knockoff filter can result in an actual FDR above q. The modified FDR can notably differ from the FDR, and in such a scenario, the reader should use knockoff+ if FDR control is desired.
We summarize the key steps of the knockoff and knockoff+ procedure as a flowchart in Figure 3.4. The (modified) FDR control holds without specific assumptions on $\mathbf{X}$ or the unknown coefficients β and works without any knowledge of the noise level. Moreover, the (modified) FDR control is guaranteed to hold in finite sample settings since their proofs do not rely on asymptotic statements.

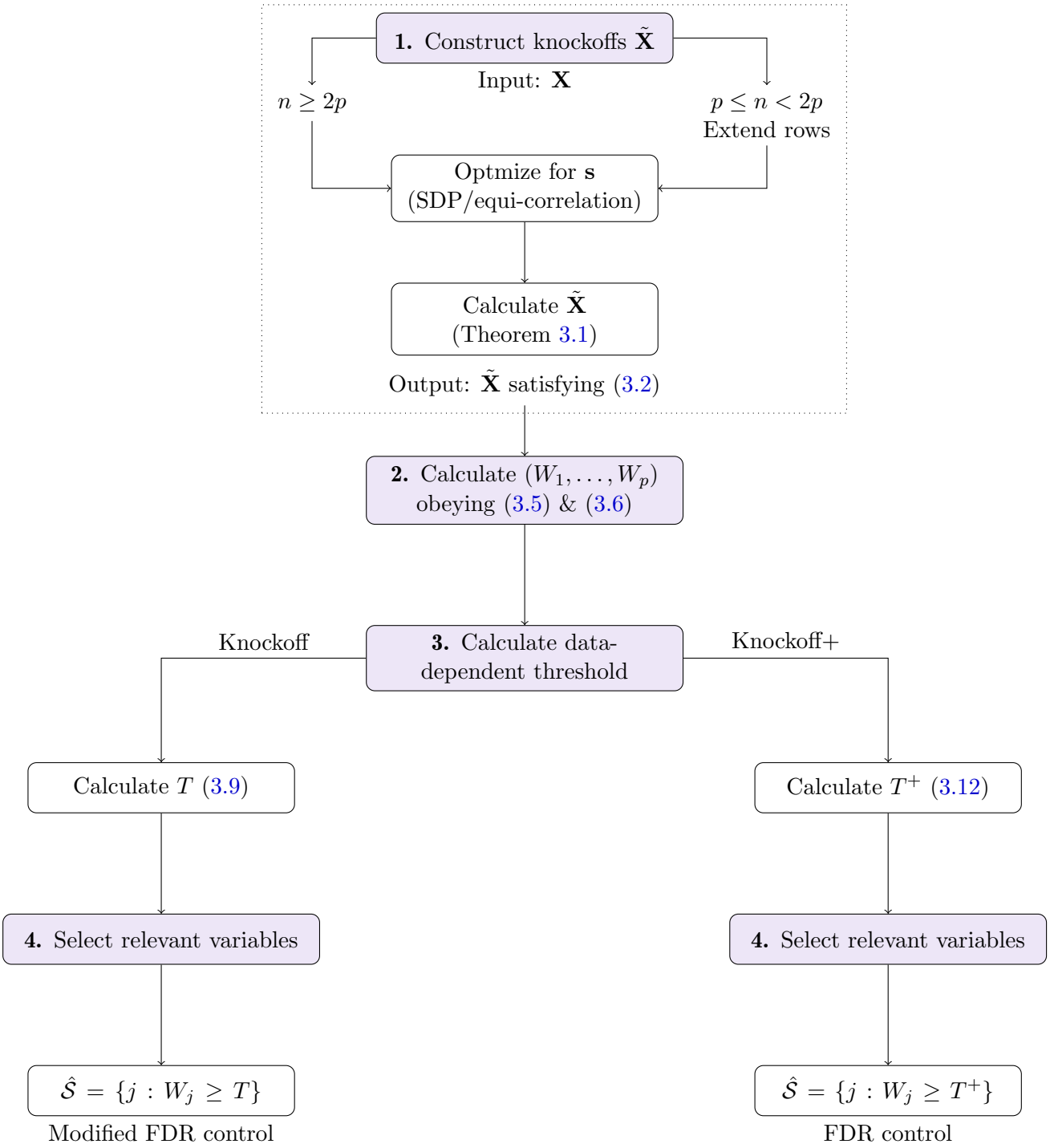


Figure 3.4: Flowchart: Fixed-X knockoffs

3.7 Theoretical guarantees

In this section, we present the fundamental components that are used to prove the FDR control of the knockoff+ method.[8] Since we will be rather explanatory, we refer to Sections 5 & 6 of Barber and Candès (2015) for a mathematically sound proof. Recall that the data-adaptive threshold for the knockoff+ is defined as

$$T^+ = \min\left\{t \in \mathcal{W}_+ : \widehat{\text{FDP}}(t) \coloneqq \frac{1 + |\mathcal{S}^-|}{|\mathcal{S}^+(t)| \vee 1} \leq q\right\}.$$

The goal is to show that the FDR is not larger than a pre-defined nominal level q. We begin by writing down the definition of the FDP for the selection set based on T^+. In the following, we will explain how to upper bound that FDP by

$$\begin{aligned}
\text{FDP}(T^+) &= \frac{|\mathcal{S}^+(T^+) \cap \mathcal{H}_0|}{|\mathcal{S}^+(T^+)| \vee 1} \\
&= \frac{|\mathcal{S}^+(T^+) \cap \mathcal{H}_0|}{|\mathcal{S}^+(T^+)| \vee 1} \cdot \frac{1 + |\mathcal{S}^-(T^+) \cap \mathcal{H}_0|}{1 + |\mathcal{S}^-(T^+) \cap \mathcal{H}_0|} \\
&= \underbrace{\frac{1 + |\mathcal{S}^-(T^+) \cap \mathcal{H}_0|}{|\mathcal{S}^+(T^+)| \vee 1}}_{\leq \widehat{\text{FDP}}(T^+) \leq q} \cdot \frac{\overbrace{|\mathcal{S}^+(T^+) \cap \mathcal{H}_0|}^{V^+(T^+)}}{1 + \underbrace{|\mathcal{S}^-(T^+) \cap \mathcal{H}_0|}_{V^-(T^+)}} \\
&\leq q \cdot \frac{V^+(T^+)}{1 + V^-(T^+)}.
\end{aligned}$$

In the second row, we extend the theoretical FDP by a quantity equal to one. After re-arranging the terms, we use the same argumentation as in (3.10) to upper bound the first term by the estimated FDP, which is by the definition of T^+ not larger than q. It remains to show that the ratio in the last term is not larger than one in expectation. By the coin-flipping lemma, both counts of nulls $V^+(T^+)$, $V^-(T^+)$ are equally distributed (see (3.11)), and hence the ratio will not be larger than one in expectation. However, we treated T^+ as fixed so far. We have to incorporate that the algorithm stops at a random time (random threshold) T^+. The following logic applies: Assume for simplicity that $|W_1| \geq \ldots \geq |W_p|$. The algorithm determines T^+ by starting with the smallest value $t = |W_p|$ and checks if this threshold satisfies $\widehat{\text{FDP}} \leq q$. If that is not the case, it continues with the next largest $t = |W_{p-1}|$, until it has found the first t for which $\widehat{\text{FDP}} \leq q$ (see Figure A.1). In the spirit of this approach, we can claim that the empirical process

$$\frac{V^+(t)}{1 + V^-(t)}$$

is a supermartingale with respect to a well defined filtration $\mathcal{F}_t = \{\sigma(V^{\pm}(u))\}_{u \leq t}$ and stopping time T^+. With this in mind, taking the expectation on both sides of our inequality above, we can show that the expected value of the supermartingale is not larger than one such that the FDR is controlled

$$\text{FDR}(T^+) \leq q \cdot \mathbb{E}\left[\frac{V^+(T^+)}{1 + V^-(T^+)}\right] \leq q \cdot \mathbb{E}\left[\frac{\overbrace{V^+(0)}^{\text{Bin}(|\mathcal{H}_0|,1/2)}}{1 + V^-(0)}\right] = q \cdot \underbrace{\mathbb{E}\left[\frac{V^+(0)}{1 + |\mathcal{H}_0| - V^+(0)}\right]}_{\leq 1} \leq q.$$

[8] The proof of the knockoff method works in a similar way.

By the introduction of the supermartingale, we can use the optional stopping time theorem (see Grimmett and Stirzaker 2001, pp. 491–495) to bound its expectation at a random time $t = T^+$ by its expectation at $t = 0$. At $t = 0$, calculations of the expected value become straightforward: The sets $V^+(0), V^-(0)$ are nothing else than the number of scores with a positive or negative sign for the nulls. Due to the coin-flipping lemma, the sign of each null score is a random coin flip, and, so $V^+(0)$ follows a binomial distribution $\text{Bin}(|\mathcal{H}_0|, 1/2)$. The last step applies the fact that $V^+(0) + V^-(0) = |\mathcal{H}_0|$ and a property of the binomial distribution (see Lemma A.5) to bound the expectation by one. We end up with an FDR not larger than q.

3.8 Advantage over permutation

There are also other techniques than knockoffs in the literature to create pseudovariables attempting to control the FDR (Wu et al. 2007). To provide some further intuition on the required correlation structure of knockoffs, we briefly want to compare them with permuted variables as controls. Assume that we create fake variables by permuting the rows of the centered and normalized data matrix $\mathbf{X}$, i.e.

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

where π is a random permutation of the indices $\{1, \dots, n\}$. Permutation preserves the correlation structure between predictors but breaks the feature-response correlation

$$\text{Corr}(\mathbf{X}_j, \mathbf{X}_k) = \text{Corr}(\mathbf{X}^{\pi}_j, \mathbf{X}^{\pi}_k), \quad \text{Corr}(\mathbf{X}^{\pi}_j, \mathbf{y}) = 0, \quad \forall j, k \in \{1, \dots, p\}.$$

The crucial difference between knockoffs and permutations is that the former are constructed such that a distinct knockoff and original variable have the same correlation as the original variables

$$\text{Corr}(\mathbf{X}_j, \tilde{\mathbf{X}}_k) = \text{Corr}(\mathbf{X}_j, \mathbf{X}_k), \quad \forall\ j \neq k,$$

whereas permuted variables do not capture by design cross-corelations between permuted and original features

$$\text{Corr}(\mathbf{X}^{\pi}_j, \mathbf{X}_k) \approx 0 \quad \forall j, k \in \{1, \dots, p\}.$$

Hence, the augmented Gram matrix corresponding to $[\mathbf{X}\ \mathbf{X}^{\pi}]$ can be written as

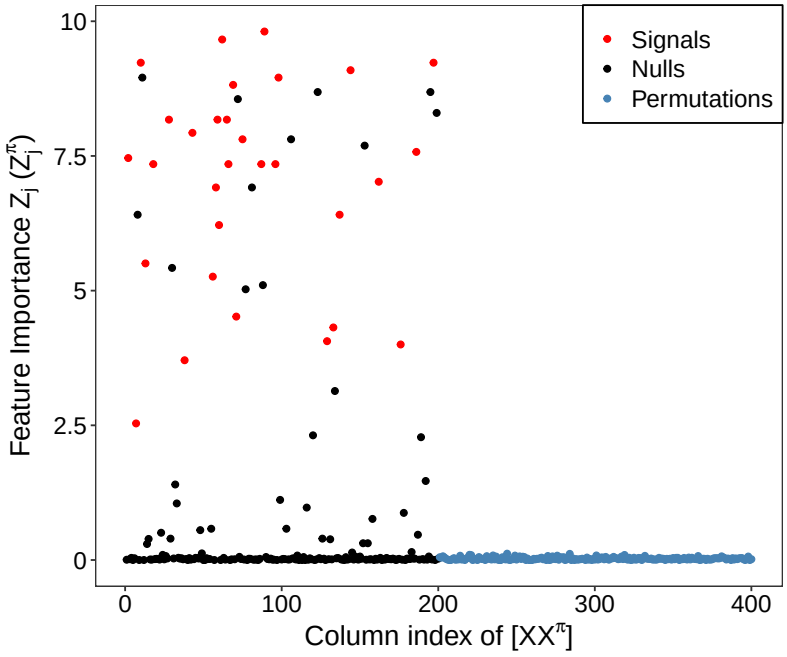
$$[\mathbf{X}\ \mathbf{X}^{\pi}]^{\top}[\mathbf{X}\ \mathbf{X}^{\pi}] \approx \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} \end{bmatrix}.$$

With this correlation structure, the coin-flipping lemma is no longer valid. In addition, the importance statistics of null variables are no longer pairwise exchangeable with their knockoff counterpart because they behave differently. The reason for that is, due to the missing permutation-original variable correlation, the PEF property, which is essential to prove Lemmas 3.3 & 3.4, does not hold anymore. To see this, consider a similar example as in Section 3.4, where we looked at the marginal correlation strength as feature importance statistics:

Example 3.7. Assume we have two standardized variables $(\mathbf{X}_1, \mathbf{X}_2)$ with pairwise correlation $\text{Corr}(\mathbf{X}_1, \mathbf{X}_2) = 0.5$, but only $\mathbf{X}_1$ is a signal, i.e. $\mathbf{y} = \mathbf{X}_1 + \boldsymbol{\varepsilon}$. Since both are correlated with each other, $\mathbf{X}_2$ will also have a non-zero marginal correlation with $\mathbf{y}$. However, by

construction, the marginal correlation of $\mathbf{y}$ with the permuted features is zero. Hence, the importance statistics of the original null and permuted variable are not distributed in the same way, that is $\mathbf{X}_2^\top \mathbf{y} \overset{d}{\neq} \mathbf{X}_2^{\pi\top} \mathbf{y}$.

Due to the missing cross-correlation between original variables and permutations, the knockoff procedure applied on $[\mathbf{X}\ \mathbf{X}^\pi]$ will result in a poor FDR control.
We will underline this argumentation by a small simulation. We generate data similar to the running example from Section 3.1 but with a covariance matrix of the form $\Sigma_{jj} = 1\ \forall j$ and $\Sigma_{ij} = 0.3\ \forall i \neq j$, and the signal coefficients are uniformly drawn from $\beta_j \sim \text{Unif}(0.5, 1.5)$.[9] We construct knockoff and permutation dummies respectively, and as importance measure, we compute the largest penalty parameter λ for which a variable enters the Lasso path first. The threshold and the final model are chosen according to knockoff+ (see Theorem 3.6). We compute the FDP after the selection and repeat the whole procedure $M = 1000$ times to obtain an estimate of the FDR at a target rate $q = 20\%$.



Importance statistic: $\sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}$.

Figure 3.5: Graphical representation of $(Z_1, \ldots, Z_p, Z_1^\pi, \ldots, Z_p^\pi)$

Figure 3.5 shows the misbehaviour of the permutations by displaying the importance measures for one iteration of the simulation. First, there is a notable number of original null variables with non-zero importance measures. Some of the nulls are probably correlated with signals, resulting in non-zero correlations between null variables and the response, and hence non-zero importance measures. In contrast, the corresponding importance statistics Z_j^π are almost all near zero due to the permutation. The pattern of the Z_j^π's does not really match with that of the null importance measures Z_j of the nulls. This is because only original null features are correlated with signals, whereas permuted controls do not display correlations with original variables at all. This shows the graphical intuition why

[9] We use stronger correlations and larger coefficients than in the running example to demonstrate the misbehaviour of permutations better.

the importance measures of the nulls are no longer pairwise exchangeable with their corresponding dummies.
Turning to the simulations results, Table 3.1 displays the FDR, power and number of cases where the dummy variable was more important than its original counterpart for both approaches. Knockoff controls provide an empirical FDR under the desired level of $q = 20\%$,

Table 3.1: Knockoffs vs. Permutations

	FDR	Power	$\overline{\#\{j : W_j < 0\}}$
Knockoff controls	17.35 %	84.20 %	81
Permutation controls	61.37 %	100 %	70

Nominal level $q = 0.2$. Values are averages over $M = 1000$ iterations. The large bar over $\#\{j : W_j < 0\}$ refers to the average.

whereas the permutation approach considerably exceeds the target FDR, being more than three times higher. Moreover, the knockoff method leads to a power of over 80%, which can be considered as large. The permutation controls lead to a suspiciously large power of 100%. One possible explanation could be: In Figure 3.5, some nulls have large Z_j's but near-zero Z_j^π's, resulting in large scores $W_j > 0$ for those, and more positive scores than the knockoff method in general. Table 3.1 confirms this by displaying that there are eleven more negative scores on average, and thus more positive scores, than in the knockoff approach. If we think of the scores in the fashion of Figure 3.3, the permutation approach will have the tendency to choose a lower threshold T^+. As a consequence, the permutation method mistakenly selects too many variables into its final model, resulting in a large power but also a poor FDR value.
In summary, for having valid dummy variables, it is crucial to preserve the cross-correlations as in the case of knockoffs (see Section 3.2.1). The permutation approach is not valid due to the missing correlation between pairs of permuted and original variables, resulting in a different behaviour of their importance statistics for null variables. Since they will be not pairwise exchangeable anymore, permutations cannot serve as valid controls.

3.9 Simulations

In this section, we conduct various simulations of the knockoff methods and the BH/BY procedures. After replicating the baseline simulation of Barber and Candès (2015), we turn to our own setting and vary the sample size n, the sparsity level $|\mathcal{S}_0|$ and the correlation structure. We finish this section by a novel simulation where we compare the robustness of the knockoff filter with different score functions W_j.

3.9.1 Baseline Simulation

We start by replicating Barber and Candès's (2015) base setting, which is Table 1 in their work. An exact copy of their results can be found in the appendix (see Table A.1), which should simplify it for the reader to compare our simulation results with the original ones. We apply the following methods for FDR control in the simulation: knockoff and knockoff+ with equi-variant and SDP knockoff construction respectively, the BH and the more conservative BY procedure. As a score function, we use $W_j = Z_j \vee \tilde{Z}_j \cdot \text{sign}(Z_j - \tilde{Z}_j)$, where the importance measure Z_j ($\tilde{Z}_j$) is the largest λ for which variable (knockoff) j enters the model first. We create a model containing $n = 3000$ observations, $p = 1000$ variables and $|\mathcal{S}_0| = 30$ signals in the following manner: We start by generating the design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ as i.i.d. entries from $\mathcal{N}(0, 1)$, followed by a normalization of each column. Then, we randomly select $|\mathcal{S}_0| = 30$ indices and randomly choose their coefficients to be $\{\pm 3.5\}$, where the remaining $p - |\mathcal{S}_0| = 970$ ones are zero.[10] The design matrix and the coefficient vector are treated as fixed over all iterations.[11] In the last step, the response is drawn according to $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{I})$, which is repeated in every iteration. Throughout all simulations of this work, we will use the following definitions for FDR and power

$$\text{FDR} = \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}| \vee 1}\right], \quad \text{Power} = \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{S}_0|}{|\mathcal{S}_0|}\right],$$

which will be estimated as averages over the iterations.
Table 3.2 presents the FDR and power of each method as averages over $M = 600$ iterations. The last column indicates whether the corresponding method yields FDR control at q in theory. All six methods provide the desired result of an empirical FDR under 20%. Among those, the BY procedure is the most conservative one with an FDR near 3%, which is accompanied by a great loss of power. Furthermore, in terms of power, all four knockoff constructions are clearly preferable over the BH procedure. Comparing the two construction methods, the SDP construction results in a slightly larger power than the equi-variant optimization, although the differences seem negligible under this generated setting. On the contrary, the power of the more conservative knockoff+ filter is by approximately six percentage points lower than the knockoff's power for both construction techniques.
If we compare our baseline results with the ones from Barber and Candès (2015) listed in Table A.1, they differ by no more than one percentage point on average, which is due to the random nature of simulations. Hence, we can support the findings of Barber and Candès.

[10] The absolute signal strength of 3.5 is chosen because it approximately equals the maximal noise level $\max_j |\mathbf{X}_j^\top \boldsymbol{\varepsilon}|$. This creates a framework "where it is possible, but not trivial, to distinguish signal from noise" (Barber and Candès 2015, p. 2072).

[11] This is not entirely clear from their paper but was confirmed by Candès in a written personal exchange.

Table 3.2: Baseline simulation

Method	FDR	Power	Theoretical FDR control
Knockoff+ (equi-variant)	13.63 %	60.04 %	Yes
Knockoff (equi-variant)	17.43 %	66.52 %	No
Knockoff+ (SDP)	14.19 %	61.11 %	Yes
Knockoff (SDP)	17.93 %	67.16 %	No
BH	19.89 %	48.87 %	No
BY	2.96 %	18.13 %	Yes

The simulated model contains $n = 3000$ observations, $p = 1000$ variables and $|\mathcal{S}_0| = 30$ signals with strength 3.5. Nominal level: $q = 0.2$. FDR and power are averages over $M = 600$ iterations.

3.9.2 Simulation: Parameter variation

Barber and Candès (2015) continue their simulation studies with their baseline setting but varying either the sparsity level, signal amplitude or feature correlation. By comparing the knockoff, knockoff+ and BH method, they find that knockoffs have a uniformly larger power and almost always a lower FDR than the BH procedure. Hence, the knockoff filters detect more signals while having a lower fraction of false discoveries on average. The power of the three methods has the tendency to converge when increasing the investigated parameter in the simulation.
In our following simulations, we will use a slightly different model than Barber and Candès to provide additional intuition and evidence for or against the knockoffs' superiority. In contrast to them, we will generate a new design matrix $\mathbf{X}$ in every iteration to cover more numerical examples and a larger space of variation. We will also use a different correlation and coefficient structure. More precisely, let our model be

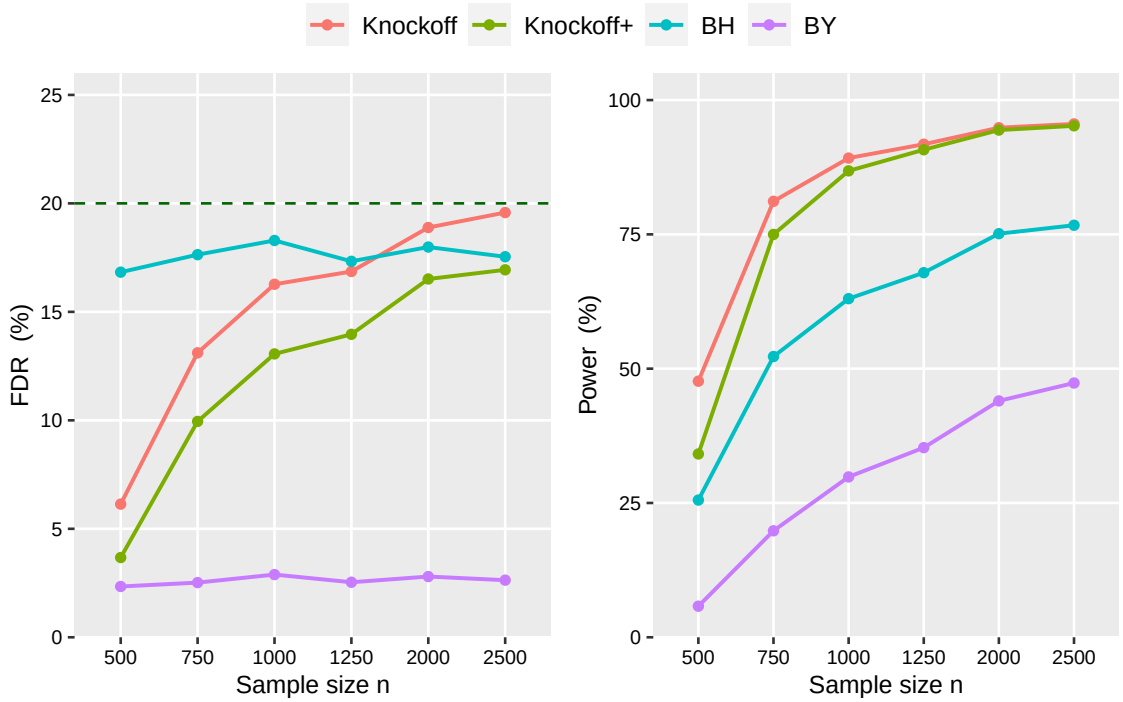
$$\begin{aligned} &\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \quad \Sigma_{j,k} = 0.25^{|j-k|} \quad \forall j \neq k \in \{1, \dots, p\}, \quad \sigma^2 = 1, \\ &n = 1000, \quad p = 300, \quad |\mathcal{S}_0| = 30, \quad |\mathcal{H}_0| = 270, \\ &\beta_j = \begin{cases} 3.5, & j \in \{1, \dots, 30\} \\ 0, & j \in \{31, \dots, 300\}. \end{cases} \end{aligned} \tag{3.13}$$

In each iteration, we start by drawing the design matrix $\mathbf{X}$ with a Toeplitz covariance structure and normalize its columns. The coefficient vector consists of two blocks, where the first $|\mathcal{S}_0|$ entries correspond to signals and the others being zero. Finally, we draw the response according to $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{I})$. We will compare the empirical FDR control and power of the knockoff and knockoff+ filter with SDP construction, the popular BH method and the more conservative BY procedure. Furthermore, we use the same score function for knockoff and knockoff+ as in the previous baseline simulation.

Variation sample size n

In contrast to Barber and Candès (2015), we will also study the effect of different sample sizes. We investigate the values $n \in \{500, 750, 1000, 1250, 2000, 2500\}$ and fix all other parameters as described in (3.13). Figure 3.6 displays the FDR and power as averages over $M = 1000$ iterations for the four methods. While all procedures control the FDR at a nominal level of $q = 0.2$, the BY method is considerably more conservative, with an FDR of around 2.5% for all sample sizes. Also, knockoff+ results in a uniformly

lower FDR than knockoff and BH. While the difference between the two knockoffs filters and the BH procedure is large for small sample sizes, their gap closes for growing n. Turning to the power, all methods show an increasing pattern for a growing sample size n Across the sample sizes investigated, the two knockoff methods have a significantly higher power than BH and BY. Moreover, the difference between knockoff and knockoff+ becomes indistinguishable with increasing sample size n.



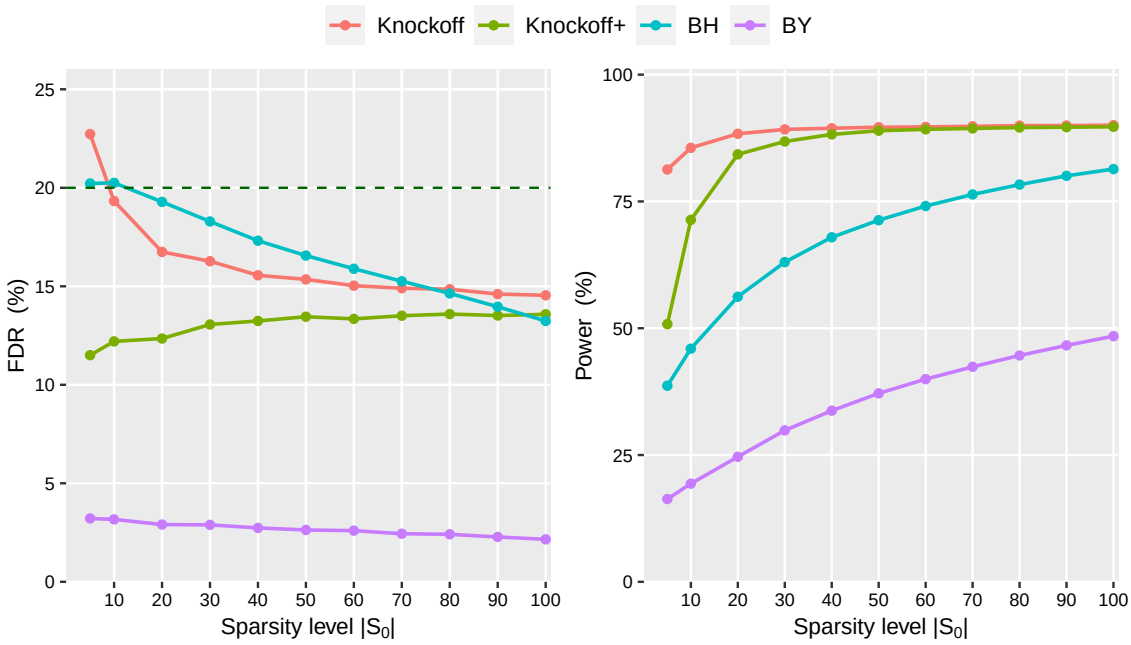
Nominal level $q = 0.2$. The model is generated with varying n, $p = 300$, $|\mathcal{S}_0| = 30$ and $\rho = 0.25$ in each of the $M = 1000$ trials.

Figure 3.6: Simulation: Varying sample size n

Variation sparsity level $|\mathcal{S}_0|$

Next, we turn our attention to different sparsity levels. We generate the model according to (3.13) but with varying $|\mathcal{S}_0| \in \{5, 10, 20, 30, \ldots, 80, 90, 100\}$. Figure 3.7 shows that for sparsity levels $|\mathcal{S}_0| > 10$, all methods successfully control the FDR, with the BY being extremely conservative again. However, if the generated models contain only 5 or 10 signals, the knockoff and BH method do not necessarily have an empirical FDR below $q = 0.2$. The same observation was also made by Barber and Candès (2015, Figure 3). While the knockoff+ filter achieves lower empirical FDR for smaller models, it catches up with the knockoff and BH method for larger $|\mathcal{S}_0|$ values. This justifies the explanation of Section 3.6, that the differences between the two knockoff methods will be larger if we only make a few discoveries, which is clearly the case if $|\mathcal{S}_0|$ is small. The addition of "+1" in the numerator of the knockoff+ threshold (3.12) will make a major difference there.
The power plot illustrates large differences between the knockoff filters and BH/BY for low and moderate sparsity levels. At $|\mathcal{S}_0| = 10$ for example, the knockoff filter has an average power of 85.57% compared to 45.97% for BH, even though both have a similar empirical FDRs. Since both knockoff filters are based on the Lasso regression, their estimation and

variable selection naturally accounts for the underlying sparsity structure. This is not the case for the BH and BY that simply involve an LS regression. Also, the noteworthy differences between both knockoff methods in very sparse models vanish quickly for $|\mathcal{S}_0| \geq 20$. In general, the power gap of BH closes with larger $|\mathcal{S}_0|$, when the true model does not have such a sparse structure anymore. Although not depicted, the power of BH is almost equal to the knockoff filters for around $|\mathcal{S}_0| = 150$ signals. The BY procedure pays its price for its conservativeness by a very flat increase in power with still a difference of more than 25% compared to the other methods, even for large $|\mathcal{S}_0| = 100$.



Nominal level $q = 0.2$. The model is generated with $n = 1000$, $p = 300$, varying $|\mathcal{S}_0|$ and $\rho = 0.25$ in each of the $M = 1000$ trials.

Figure 3.7: Simulation: Varying sparsity level $|\mathcal{S}_0|$

Variation correlation structure

Finally, we examine the effect of two different feature correlation structures

- Toeplitz structure: $\Sigma_{j,k} = \rho^{|j-k|}, \ \forall j \neq k,$
- Equi-variant correlation: $\Sigma_{j,k} = \rho, \ \forall j \neq k,$

and we also vary the correlation strength $\rho \in \{0.3, 0.5, 0.7, 0.9\}$. The remaining parameters are set as in (3.13). We can interpret the Toeplitz structure as the desirable case and the equi-variant correlation as the undesirable case because all signals and nulls have the same correlation. This structure will probably hamper the algorithm to distinguish between them. Our setting differs from Barber and Candès (2015) in two ways: First, we additionally investigate the "undesirable" equi-correlation structure. Second, we do not select the signal indices at random but as the first 30 entries. This will have implications on the correlation strength between signal and null variables in the Toeplitz structure.

Table 3.3 illustrates that all methods provide empirical FDR control in all cases of Toeplitz structures. Both knockoff methods become more conservative for increasing correlation

Table 3.3: Simulation: FDR for varying correlation structure

	ρ	Knockoff	Knockoff+	BH	BY
$\rho^{\lvert j-k\rvert}$	0.3	14.84 %	11.85 %	18.18 %	2.97 %
	0.5	10.70 %	7.93 %	17.50 %	2.83 %
	0.7	4.86 %	2.96 %	17.42 %	3.49 %
	0.9	0.07 %	0.03 %	18.18 %	2.12 %
$\rho_{jk} = \rho$	0.3	20.36 %	12.32 %	17.97 %	3.16 %
	0.5	21.81 %	10.16 %	18.64 %	2.44 %
	0.7	22.95 %	7.43 %	16.42 %	2.89 %
	0.9	25.90 %	4.39 %	19.67 %	3.10 %

Nominal level $q = 0.2$. The model is generated with $n = 1000$, $p = 300$, $|\mathcal{S}_0| = 30$ and varying ρ in each of the $M = 1000$ trials.

Table 3.4: Simulation: Power for varying correlation structure

	ρ	Knockoff	Knockoff+	BH	BY
$\rho^{\lvert j-k\rvert}$	0.3	88.72 %	86.46 %	59.79 %	26.27 %
	0.5	80.15 %	74.47 %	37.89 %	11.03 %
	0.7	57.66 %	42.76 %	12.74 %	2.52 %
	0.9	22.55 %	10.11 %	1.27 %	0.18 %
$\rho_{jk} = \rho$	0.3	35.07 %	22.17 %	47.85 %	17.22 %
	0.5	20.21 %	10.80 %	27.18 %	6.46 %
	0.7	10.43 %	4.54 %	9.14 %	1.67 %
	0.9	3.91 %	1.05 %	1.26 %	0.02 %

Nominal level $q = 0.2$. The model is generated with $n = 1000$, $p = 300$, $|\mathcal{S}_0| = 30$ and varying ρ in each of the $M = 1000$ trials.

strength, while BH and BY almost remain on the same FDR level. Barber and Candès (2015) observe a different behaviour of the normal knockoff filter for large ρ, which does not yield empirical FDR control for $\rho = 0.9$ anymore. This difference originates from treating the signal indices randomly. To replicate their findings and validate our point, we additionally conducted the same simulation in the appendix and chose $|\mathcal{S}_0| = 30$ indices at random to be non-zero. Similar to Barber and Candès, the normal knockoff leads to an empirical FDR above the nominal level 0.2 for $\rho = 0.9$ (see bold value in Table A.2).[12] The higher FDR values and lower power for the random assignment of signal indices could stem from the following reason: the correlation between signal and null variables will probably be stronger than in our simulation setting with block coefficients (3.13). As a consequence, the estimation techniques will have more difficulties to distinguish between neighbouring null and signal features.

Turning back to Table 3.3, we can observe the same issue for the knockoff method under the equi-variant structure. It does yield empirical FDR values above 0.2 for all correlation strengths in contrast to the other three methods. This is consistent with the theory of Section 3.6. When the final selection set $\hat{\mathcal{S}}$ is small, the modified FDR can noticeably differ from the FDR. Although not displayed here, in our simulations, the selection sets

[12] We also performed the two previous simulations with a random signal assignment (not shown here). All previously made arguments remain the same. Only the level of the curves slightly shifts.

of all four methods were smaller for larger correlations. Additionally, the final model includes relatively more false positives due to its difficulties of distinguishing between null and signal variables that have a large pairwise correlation.
Looking at the power in Table 3.4, we observe a strong decay for increasing ρ, which comes again from the difficulty of distinguishing between null and signal variables due to large correlations. The knockoff and knockoff+ filter lead to a larger power than the other two methods for the Toeplitz structure. However, in the equi-variant case for $\rho \in \{0.3, 0.5\}$, BH achieves higher power than both knockoff filters. Since the BH procedure relies on an LS estimation with data ($\mathbf{X}$,$\mathbf{y}$), "only" p variables suffer from the undesirable equi-correlation. The two knockoff procedures, however, augment the data set, and the Lasso regression has to deal with $2p$ regressors featuring an undesirable correlation. The nature of the knockoff procedure (the augmentation) seems to leverage the problem of strong correlations, making it even harder for the algorithm to distinguish between signal, null and knockoff variables. For larger $\rho \in \{0.7, 0.9\}$, the LS regression starts to suffer from the large correlation problem more and more, and the differences in power become smaller compared to the knockoff methods.

3.9.3 Simulation: Score functions

Barber and Candès (2015, pp. 2064 f.) propose different score statistics that obey the sufficiency and asymmetry properties, but without comparing them further. In this last part of our simulation section, we want to compare some of them regarding their FDR control and power. We will investigate the following statistics:

i.) Marginal correlations (MCorr): $W_j = |\mathbf{X}_j^\top \mathbf{y}| - |\tilde{\mathbf{X}}_j^\top \mathbf{y}|$.

ii.) LS coefficient difference: $W_j = |\hat{\beta}_j^{LS}| - |\hat{\beta}_{j+p}^{LS}|$,
where the estimated coefficient vector $\boldsymbol{\beta}$ is obtained by regressing the response $\mathbf{y}$ on the augmented data matrix $[\mathbf{X}\ \tilde{\mathbf{X}}]$.

iii.) Lasso coefficient difference (LCD): $W_j = |\hat{\beta}_j(\lambda)| - |\hat{\beta}_{j+p}(\lambda)|$ at a **fixed** λ.

iv.) Lasso lambda signed max (LLSM): $W_j = Z_j \vee \tilde{Z}_j \cdot \operatorname{sign}(Z_j - \tilde{Z}_j)$,
with $Z_j = \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}, \quad \tilde{Z}_j = \sup\{\lambda : \hat{\beta}_{j+p}(\lambda) \neq 0\}$.

v.) Lasso lambda difference (LLD): $W_j = |Z_j| - |\tilde{Z}_j|$,
with $Z_j = \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}, \quad \tilde{Z}_j = \sup\{\lambda : \hat{\beta}_{j+p}(\lambda) \neq 0\}$.

vi.) Orthogonal matching pursuit (OMP): $W_j = Z_j \vee \tilde{Z}_j \cdot \operatorname{sign}(Z_j - \tilde{Z}_j)$,
where $(Z_1, \ldots, Z_p, \tilde{Z}_1, \ldots, \tilde{Z}_p)$ defines the reverse order of the $2p$ variables entering a forward selection procedure. In an iterative manner, we choose the variable that maximizes the inner product

$$j_t = \arg\max_j |\langle \mathbf{X}_j, \mathbf{r}_{t-1} \rangle|.$$

The updated residual $\mathbf{r}_t$ for the next step is then the remainder of an LS regression on all previously selected features $\mathbf{X}_{j_1}, \ldots, \mathbf{X}_{j_t}$ (Pati et al. 1993).

Note that LLSM was the score function that we have used in the running example and in the previous simulations. It is important to underline that the LCD statistic would not satisfy the sufficiency requirement anymore if λ was tuned by cross-validation. Since LCD requires a fixed λ, we have decided to choose an arbitrary value for the tuning parameter, in our case the median value of the λ-sequence, in order to investigate how

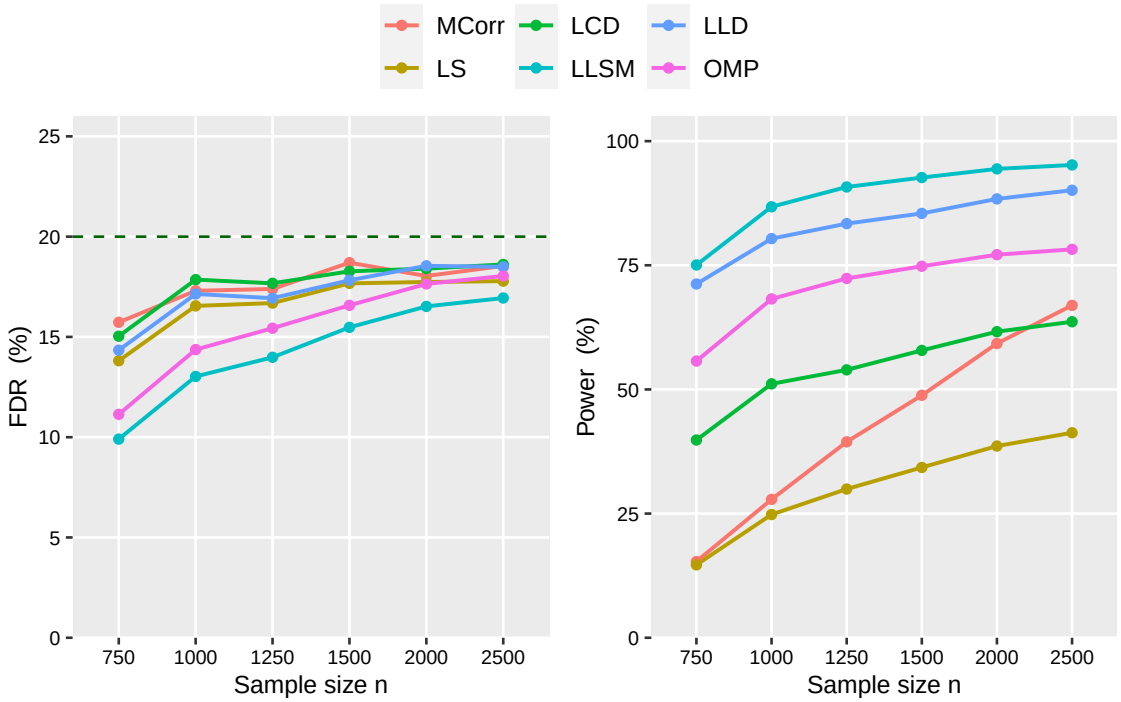
different degrees of information influence the FDR and power. Since all statistics are valid in the sense of satisfying the antisymmetry and sufficiency property, we expect that all of them yield FDR control. However, we can also anticipate significant differences in their power since the statistics are based on different estimation techniques and degrees of information. Marginal correlations ignore any type of interactions with other features. They do not view the importance of variable j conditionally on all other features. Next, although the LS difference takes conditional dependencies into account, it suffers from doubling the number of variables by augmenting the data matrix. The LS regression is not a variable selection technique by itself and its model is non-sparse. The large number of non-important variables (the original null and knockoff features) can therefore impair the estimation quality. Turning to the Lasso score functions, even though LCD accounts for the sparsity assumption by its nature, it may suffer from the lack of information by choosing an arbitrary tuning parameter. The Lasso regression does not perform equally well over the whole regularization path. Another issue is the coefficient bias introduced by Lasso. This could also have an unintended (negative) effect if we use the coefficient differences as a comparison for a feature's importance compared to its knockoff. The LLSM and LLD statistics are not based on biased coefficients and include an additional source of information by accounting for when each variable enters the model. Last but not least, the OMP is a completely different technique based on a forward selection approach. It probably performs better than i.) – iii.) since it has a variable selection character, and it is less arbitrary.

With all these different score statistics equipped, we run the same simulations as in Section 3.9.2. We use the knockoff+ method since its FDR control is theoretically guaranteed. Figure 3.8 shows the results for a varying sample size n.[13] As already expected, all six score functions achieve empirical FDR control at $q = 20\%$. There are also (large) differences in their ability to detect signals. LLD and LLSM, the scores that use the largest amount of information, perform best, whereas the LS statistic has the lowest power. Although the power of all scores increases with growing sample size, their differences remain almost the same. Moreover, MCorr is the only score whose power seems to grow linearly with n, whereas the others have a diminishing increase. The reason could be that only MCorr is not based on a regression technique and does not account for conditional dependencies. Except for MCorr, we also see a clear ranking of the scores regarding their power, which holds across all examined sample sizes.

Turning to the sparsity level $|\mathcal{S}_0|$, Figure A.2 shows that all scores lead to an empirical FDR control at q. The power curves depict the same ranking as in the sample size simulation, with LLSM being the score with the largest power closely followed by LLD. In contrast to the previous simulation, the differences in power become smaller for growing $|\mathcal{S}_0|$. For example, in the sparse case of $|\mathcal{S}_0| = 10$, LCD achieves a power of 23%, whereas LLSM has a power of almost 70%. In the less sparse case of $|\mathcal{S}_0| = 100$, LCD catches up to a power of roughly 75%, being only 15 percentage points lower than LLSM. Despite the converging tendency of the scores to each other, the differences still remain significantly.

Tables A.3 & A.4 present the results for varying correlation structures. All empirical FDR values are below the 0.2 bound for both correlation structures. The power shows a more interesting pattern. If we generate a Toeplitz structure, we see that each of the Lasso scores perform best in terms of power for a certain correlation strength. For example, while LLSM performs best at $\rho = 0.3$, LLD has the highest power at $\rho = \{0.5, 0.7\}$ and LCD at $\rho = 0.9$. For equi-correlations, LCD's power is surprisingly larger than LLSM

[13] We do not use the sample size $n = 500$ because statistic ii.) cannot be computed there. Instead, we additionally examine the setting $n = 1500$.



Nominal level $q = 0.2$. The model is generated with varying n, $p = 300$, $|\mathcal{S}_0| = 30$ and $\rho = 0.25$ in each of the $M = 1000$ trials.

Figure 3.8: Simulation score function: Varying sample size n

and LLD. OMP seems to outperform the two latter scores as well. For example, with an equi-correlation of $\rho = 0.5$, LCD achieves a power of 44.24%, almost 20 percentage points higher than OMP. The LLSM and LLD scores, however, lead only to a power of 10.60% and 12.30% respectively. The reason why OMP performs better than LLSM and LLD could be that the forward selective procedure does not include all variables at the same time. Hence the problem of the equi-correlation is less present than in the Lasso regressions with $2p$ equi-correlated regressors. However, it remains unclear why LCD performs better than all other methods then.

3.10 Discussion

This chapter introduced the knockoff filters that yield theoretical FDR control in a homoscedastic Gaussian linear model. This holds under arbitrary covariate matrices and regardless of the number of signal variables or their signal strength. Knockoffs even provide FDR control for modern penalization techniques such as Lasso, where inference is still not a trivial topic. Through simulations, we showed the superiority of knockoffs over the popular BH procedure under various settings. Attention should be paid if the data matrix is highly correlated as in the model with equi-variant correlations. Although the FDR is still controlled, the data augmentation seems to leverage the correlation problem by having $2p$ undesirable correlated features. This can probably result in a lower power than the BH procedure.
The flexibility of defining different score functions implies that its choice noticeably influences the resulting power. LLSM and LLD performed significantly better than score functions that were not based on selection techniques, such as marginal correlations or the LS coefficient differences. Again, we saw an unexpected behaviour in the case of equi-correlations, where LCD and OMP performed better. Mentioning the LCD statistic, the requirement of the sufficiency property (3.5) prevents the use of an LCD score based on a cross-validated λ, which could be an even more powerful score statistic due to its optimized character. From our novel simulation comparing different score functions, we conclude the following: In practice, it seems reasonable to choose LLSM because it is often the most powerful score. LLSM is also the default method in the `knockoff.filter` function in `R` when applying fixed-X knockoffs. However, it would be interesting to study power curves in a more theoretical framework to understand the behaviour and superiority of different scores better. For example, we can not fully explain by simulations the surprising behaviour of the scores when our model is either based on an undesired correlation structure such as equi-correlations or features strong correlations in general. First research in this direction has been conducted by Weinstein et al. (2017) and Ke et al. (2020). Due to the endless possibilities of score functions, it would be also interesting to investigate new score statistics that are tailored to certain model scenarios.
The fixed-X knockoff filter also has some general shortcomings. It works only in a low-dimensional ($n \geq p$) homoscedastic linear Gaussian design as described in (3.1). Many contemporary statistical applications are based on non-linear or non-Gaussian relationships such as random forests or generalized linear models (GLM). Moreover, many modern data sets are high-dimensional $p \gg n$, where the problem of selecting only a few relevant variables without too many false positives might be even more important. Fixed-X knockoffs cannot be used in both of these cases. In the next chapter, we will introduce model-X knockoffs, which extend the knockoffs framework to the high-dimensional case and non-linear models.[14]

[14] Actually, Barber and Candès (2019) extend fixed-X knockoffs to the high-dimensional framework, but we will not discuss their approach further. In practice, it is more common to apply model-X knockoffs in high dimensions. Model-X knockoffs have further advantages compared to fixed-X knockoffs that we will also explain in detail.

Chapter 4

Model-X knockoffs

Fixed-X (FX) knockoffs have proven FDR control, and they achieve usually greater power than classical FDR controlling methods. Their application, however, is restricted to the homoscedastic and low-dimensional Gaussian linear model. Candès et al. (2018) developed model-X (MX) knockoffs in a subsequent work, where they extend the knockoff framework beyond the low-dimensional Gaussian linear model.
Section 4.1 describes the underlying setting of MX knockoffs, followed by their general definition and construction in Section 4.2. In Section 4.3, we have a look at their (exchangeability) properties, after which we discuss their implications for the choice of score statistics W_j in Section 4.4. We continue by contrasting the properties of MX and FX knockoffs, and more importantly, discuss some open research questions and new developments in Section 4.5. Finally, Section 4.6 provides a simulation comparison of MX and FX knockoffs in a linear model but with model misspecifications.

4.1 Setting

Recall the formal variable selection framework of Section 2.1. Assume the covariates $X = (X_1, \dots, X_p) \in \mathbb{R}^p$ are i.i.d. with joint distribution F_X. With $Y \in \mathbb{R}$ being the response, we have

$$(X_{i,1}, \dots, X_{i,p}, Y_i) \overset{i.i.d.}{\sim} F_{XY}, \quad i = 1, \dots, n,$$

which is a $(p+1)$-dimensional joint distribution of the form $F_{XY} = F_{Y|X} \circ F_X$. To apply MX knockoffs,

- the feature joint distribution F_X has to be **known**.
- the conditional distribution of the response given the features $F_{Y|X}$ does **not** have to be **known** and can be **arbitrary**.

Placing distributional assumptions on the origin of the data matrix $\mathbf{X}$ allows a probabilistic construction of the knockoffs $\tilde{\mathbf{X}}$. Remember that in the FX knockoff method, the conditional distribution $F_{Y|X}$ describes a homoescedastic Gaussian linear model (3.1), whereas nothing is assumed about the probabilistic origin of $\mathbf{X}$, treating it as fixed.
Most regression models and variable selection methods place strong parametric assumptions on how the response is linked with the covariates ($F_{Y|X}$) but usually treat the covariates as fixed. However, MX knockoffs follow a different modelling approach: they make no

assumptions on the relationship between Y and X at a price of knowing the distribution of X. It shifts the burden of placing assumptions or having knowledge of a distributional origin. Therefore, the methodological philosophy of Candès et al. (2018) is modelling X instead of Y given X. This might or might not be appropriate, depending on the data.[1]
The newly placed assumptions enable us to run the MX knockoff filter with almost any regression or classification setting, including non-linearities, which are often the foundation of state-of-the-art machine learning models such as random forests or boosting methods. Since the MX approach does not place any restrictions on the dimensionality, it can also be applied to high-dimensional data sets $p > n$.
Not relying on p-values (as BH and BY) can be a huge advantage since their valid computation is not always guaranteed. Going beyond an LS regression with $n \geq p$, already a low-dimensional GLM requires the use of asymptotic p-values, which are not necessarily uniformly distributed under the null anymore in practice. In high-dimensional linear models, there is a post-selection inference theory, but it often relies on restrictive assumptions such as super sparsity to prove asymptotic statements. And for most modern machine learning models, we do not have p-values at all. MX knockoffs do not rely on p-values or asymptotics, and provide finite sample FDR control with almost any modelling method (Candès et al. 2018, p. 555–557). The usefulness of MX knockoffs is clearly their wide range of applications without any dimensionality restrictions. Combined with a machine learning model, MX knockoffs yield some interpretability for an otherwise black-box shaped model.

The MX knockoffs intuitions are very similar to their predecessor, the FX design. We will mainly focus on their differences: their definition, construction and the availability of a broader range of feature importance statistics (steps 1–2). Steps 3–4, namely the calculation of the threshold T (T^+), the selection rules and (modified) FDR Theorems 3.5 & 3.6, as well as the theoretical sketch of proof why knockoffs work, are essentially the same, and we refer to Sections 3.5–3.7 for a review. If we do not mention any further comments about a certain intuition, then the reader can assume that the same intuition as for the FX design holds.

4.2 Step 1: Construct knockoffs

4.2.1 Definition

We start with the general definition of valid knockoffs in the MX framework:

Definition 4.1 (**Model-X knockoffs**). *Let $X = (X_1, \ldots, X_p)$ be the random vector of the covariates. Then, the corresponding knockoffs $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)$ are constructed such that they obey the following properties:*

i.) Pairwise exchangeability: For any subset $S \subset \{1, \ldots, p\}$,

$$(X, \tilde{X})_{\text{swap}(S)} \stackrel{d}{=} (X, \tilde{X}). \tag{4.1}$$

ii.) Conditional independence: If there is a reponse Y in the model, then $\tilde{X} \perp\!\!\!\perp Y | X$.

[1] See Candès et al. (2018, p. 554–555) for examples, where the distributional modelling approach of the MX knockoffs is reasonable.

The first condition implies that original and knockoff features are pairwise exchangeable. We can take any subset of original variables and change their entries in the vector $(X_1, \dots, X_p, \tilde{X}_1, \dots, \tilde{X}_p)$ with their corresponding knockoffs without affecting their $2p$-dimensional joint distribution.[2] Taking the set $S = \{1, \dots, p\}$ and marginalizing the joint distributions in (4.1), we conclude that $X \overset{d}{=} \tilde{X}$, that is knockoffs have the same marginal joint distribution as the originals variables. The second condition ensures that knockoffs are fake null features. They do not provide any additional information about Y given that we know X. The condition is automatically guaranteed if we construct knockoffs without taking the response into account. Defintion 4.1 encodes the correlation properties between knockoffs and original variables described in Section 3.2.1 but as distributional and not geometric requirements. In fact, the pairwise exchangeability is even stricter because the distributional equality also involves higher moments than only correlations. The fulfillment of Definition 4.1 ensures that knockoffs mimic the behaviour of the original features but with no influence on the response, qualifying them as valid control variables.
In terms of the data $\mathbf{X}$ and $\mathbf{y}$ containing the i.i.d. samples row-wise, the construction of the knockoff matrix $\tilde{\mathbf{X}}$ satisfies the pairwise exchangeability property (4.1) with respect to the joint vector $(X_{i1}, \dots, X_{ip}, \tilde{X}_{i1}, \dots, \tilde{X}_{ip})$ for each i.[3]

4.2.2 Exact construction

The first step of the MX knockoff filter is to find a "knockoff sampler": Given the exact distribution F_X, we want to find a conditional distribution $F_{\tilde{X}|X}$ to generate $\tilde{X}$ given X such that the joint distribution of $(X, \tilde{X})$ satisfies the exchangeability property.
We start with a slightly different and sequential characterization of Definition 4.1 by Bates et al. (2020). The sequential characterization establishes what a knockoff sampler has to fulfil to create valid knockoffs.

Proposition 4.2 (Theorem 1, Bates et al. (2020))**.** *The random variables $(\tilde{X}_1, \dots, \tilde{X}_p)$ are valid knockoffs for $(X_1, \dots, X_p)$ if and only if the following conditions hold:*

i.) Pairwise conditional exchangeability: For any $j \in \{1, \dots, p\}$,

$$(X_j, \tilde{X}_j)|X_{-j}, \tilde{X}_{1:(j-1)} \overset{d}{=} (\tilde{X}_j, X_j)|X_{-j}, \tilde{X}_{1:(j-1)}.$$

The variable-knockoff pair $(X_j, \tilde{X}_j)$ is conditionally exchangeable given all other variables X_{-j} and already generated knockoffs $\tilde{X}_{1:(j-1)}$.

ii.) Knockoff symmetry: For any $j \in \{1, \dots, p\}$,

$$\mathbb{P}((X_j, \tilde{X}_j) \in \mathcal{B}|X_{-j}, \tilde{X}_{1:(j-1)})$$

is $\sigma(X_{(j+1):p}, \{X_1, \tilde{X}_1\}, \dots, \{X_{j-1}, \tilde{X}_{j-1}\})$-measureable for any Borel set $\mathcal{B}$, with $\{\cdot, \cdot\}$ being an unordered pair. This means that the conditional distribution remains the same if we change the previously sampled knockoffs with the original variables.

iii.) Conditional independence: If there is a reponse Y in the model, then $\tilde{X} \perp\!\!\!\perp Y|X$.

Candès et al. (2018) propose the Sequential Conditional Independent Pairs (SCIP) algorithm based on the sequential notion of Proposition 4.2, which creates valid knockoffs for

[2] We refer to Section 3.4 for a reminder of the swap notation.
[3] The i.i.d. requirement of the rows ensures that Definition 4.1 also holds if formulated in terms of the data $(\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{y})$.

any known covariate distribution F_X. The SCIP algorithm steps sequentially through all variables $j = 1, \dots, p$ and samples knockoffs from a conditional distribution in each step (Algorithm 1). In particular, $\mathcal{P}(X_j|X_{-j}, \tilde{X}_{1:j-1})$ denotes the conditional distribution of X_j

Algorithm 1 Sequential Conditional Independent Pairs

Input: X and F_X
  for $j \in \{1, \dots, p\}$ **do**
    Sample $\tilde{X}_j$ from $\mathcal{P}(X_j|X_{-j}, \tilde{X}_{1:j-1})$ conditionally independently of X_j
  end for
Output: $\tilde{X}$

given the remaining original variables and already computed knockoffs $(X_j, \tilde{X}_{1:j-1})$. We illustrate the procedure for a simple example with three variables: We start by sampling $\tilde{X}_1$ from the conditional distribution $\mathcal{P}(X_1|X_2, X_3)$. Since $\tilde{X}_1$ is conditionally independent of X_1 but with the same marginal distribution, it is exchangeable with X_1 given all other variables (X_1, X_2). Once this is computed, the joint distribution $\mathcal{P}(X_{1:3}, \tilde{X}_1)$ is known. From the joint distribution, we can derive the conditional distribution $\mathcal{P}(X_2|, X_1, X_3, \tilde{X}_1)$ and sample $\tilde{X}_2$ from that. After having access to the joint distribution $\mathcal{P}(X_{1:3}, \tilde{X}_{1:2})$, we compute the conditional distribution $\mathcal{P}(X_3|X_{1:2}, \tilde{X}_{1:2})$ and sample $\tilde{X}_3$ from that. Having cycled through all three variables, we finally end up with a joint distribution $\mathcal{P}(X_{1:3}, \tilde{X}_{1:3})$ obeying the pairwise exchangeability property (4.1).[4] A proof that shows why Algorithm 1 generates knockoffs with the exchangeability property is given in Appendix B of Candès et al. (2018).

The SCIP algorithm should be interpreted as a conceptual proof for the existence of valid knockoffs for any covariate distribution F_X rather than a computationally efficient algorithm for practical use. Its computational burden and complexity come from calculating joint and conditional distributions and sample knockoffs from them in each step, which will be time-consuming for large p. One exception where the SCIP algorithm works efficiently (linear in p) is when X is a Markov chain (Sesia et al. 2018). Moreover, although the SCIP algorithm produces valid knockoffs, it is also not guaranteed that these have the largest power. That issue is further investigated in Bates et al. (2020). Therefore, it remains to find suitable knockoffs samplers that we can use for the practical application of MX knockoff filters.

4.2.3 Example: Gaussian knockoffs

When X follows a multivariate normal distribution, then an explicit and efficient solution to generate exact knockoffs exists. For simplicity, we assume that the original features follow $F_X = \mathcal{N}(\mathbf{0}_p, \mathbf{\Sigma})$. Since Defintion 4.1 implies marginal distributional equality, the knockoffs do need to have the same distribution $F_{\tilde{X}} = \mathcal{N}(\mathbf{0}_p, \mathbf{\Sigma})$. Logically, the augmented joint distribution of $(X, \tilde{X})$ does have to be multivariate Gaussian as well but ensuring pairwise exchangeability (4.1). More precisely, let

$$(X, \tilde{X}) \sim \mathcal{N}(\mathbf{0}_{2p}, \mathbf{G}), \quad \text{with } \mathbf{G} = \begin{bmatrix} \mathbf{\Sigma} & \mathbf{\Sigma} - \text{diag}\{\mathbf{s}\} \\ \mathbf{\Sigma} - \text{diag}\{\mathbf{s}\} & \mathbf{\Sigma} \end{bmatrix}, \tag{4.2}$$

be the augmented joint distribution, with $2\mathbf{\Sigma} \succeq \text{diag}\{\mathbf{s}\} \succeq 0$ to guarantee positive semidefiniteness of $\mathbf{G}$ (compare Section 3.2.2). The off-diagonals of $\mathbf{\Sigma} - \text{diag}\{\mathbf{s}\}$ imply that for

[4] The conditinal independence property of Definition 4.1 is satisfied since Algorithm 1 does not include the response in the knockoff construction.

each pair (j,k) we have the same covariances $\text{Cov}(X_j, \tilde{X}_k) = \text{Cov}(X_j, X_k)$, or to put in the spirit of the more general Definition 4.1, $(X_j, X_k) \overset{d}{=} (X_j, \tilde{X}_k)$. We immediately see the resemblance between the above covariance matrix and the augmented Gram matrix (3.2) of the FX design. In the latter, we treated $\mathbf{X}$ as fixed and set $\boldsymbol{\Sigma} = \mathbf{X}^\top\mathbf{X}$, which is the sample covariance matrix in case of a column-wise centred design matrix. The FX framework requires that the *sample* covariance matrix satisfies the "deterministic" exchangeability property: swapping columns leaves the covariance unchanged, which was ensured by the PEF property (3.7). In contrast, the MX design requires invariance of the *population* covariance matrix. In particular, the MX knockoffs will be far from satisfying equivalent sample covariance matrices $\mathbf{X}^\top\mathbf{X} = \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}$ due to the probabilistic nature of the setting, whereas this equivalence holds for the FX design by construction of (3.2).
Turning back to the Gaussian joint vector $(X, \tilde{X})$, if we swap variables with their knockoffs, the resulting distribution will be also Gaussian with covariance matrix $\mathbf{PGP}$, where the permutation matrix $\mathbf{P}$ defines the swap. Since multiplication with $\mathbf{P}$ from both sides does not affect the covariance matrix itself $\mathbf{PGP} = \mathbf{G}$, the joint distribution remains the same, and so it obeys the exchangeability property (4.1).
Equivalent to Section 3.2.2, we choose $\mathbf{s}$ such that X_j and $\tilde{X}_j$ are as orthogonal as possible, while ensuring positive semidefinitness of $\mathbf{G}$. Having obtained $\mathbf{s}$, we calculate the conditional distribution from the joint distribution (4.2) and sample the knockoffs $\tilde{X}$ given X, that is $\tilde{X} \sim F_{\tilde{X}|X}$. More specifically, if we treat the covariates of the i-th observation $\mathbf{X}_{i,*}$ as $p \times 1$ vector, we can draw the rows $\tilde{\mathbf{X}}_{i,*}$ of the knockoff matrix $\tilde{\mathbf{X}}$ from

$$\begin{aligned} F_{\tilde{X}|X}(\cdot|\mathbf{X}_{i,*}) &= \mathcal{N}(\boldsymbol{\mu'}, \mathbf{V}) \\ \boldsymbol{\mu'} &= (\mathbf{I}_p - \text{diag}\{\mathbf{s}\}\boldsymbol{\Sigma}^{-1})\mathbf{X}_{i,*} \\ \mathbf{V} &= 2\,\text{diag}\{\mathbf{s}\} - \text{diag}\{\mathbf{s}\}\boldsymbol{\Sigma}^{-1}\,\text{diag}\{\mathbf{s}\} \end{aligned}$$

independently for each i. To see where the conditional mean and variance comes from, we apply the general formulas for a multivariate normal distribution (Definition A.3),

$$\begin{aligned} \mathbb{E}[\tilde{\mathbf{X}}_{i,*}|\mathbf{X}_{i,*}] &= \boldsymbol{\mu}_{\tilde{X}} + \boldsymbol{\Sigma}_{\tilde{\mathbf{X}}\mathbf{X}}\boldsymbol{\Sigma}_{\mathbf{X}}^{-\mathbf{1}}(\mathbf{X}_{i,*} - \boldsymbol{\mu}_X) = \mathbf{0} + (\boldsymbol{\Sigma} - \text{diag}\{\mathbf{s}\})\boldsymbol{\Sigma}^{-1}(\mathbf{X}_{i,*} - \mathbf{0}) \\ &= (\mathbf{I}_p - \text{diag}\{\mathbf{s}\}\boldsymbol{\Sigma}^{-1})\mathbf{X}_{i,*}, \end{aligned}$$

$$\begin{aligned} \text{Var}[\tilde{\mathbf{X}}_{i,*}|\mathbf{X}_{i,*}] &= \boldsymbol{\Sigma}_{\tilde{\mathbf{X}}} - \boldsymbol{\Sigma}_{\tilde{\mathbf{X}}\mathbf{X}}\boldsymbol{\Sigma}_{\mathbf{X}}^{-\mathbf{1}}\boldsymbol{\Sigma}_{\mathbf{X}\tilde{\mathbf{X}}} = \boldsymbol{\Sigma} - (\boldsymbol{\Sigma} - \text{diag}\{\mathbf{s}\})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \text{diag}\{\mathbf{s}\}) \\ &= \boldsymbol{\Sigma} - (\mathbf{I} - \text{diag}\{\mathbf{s}\}\boldsymbol{\Sigma}^{-1})(\boldsymbol{\Sigma} - \text{diag}\{\mathbf{s}\}) \\ &= \boldsymbol{\Sigma} - \boldsymbol{\Sigma} + \text{diag}\{\mathbf{s}\} + \text{diag}\{\mathbf{s}\}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma} - \text{diag}\{\mathbf{s}\}\boldsymbol{\Sigma}^{-1}\,\text{diag}\{\mathbf{s}\} \\ &= 2\,\text{diag}\{\mathbf{s}\} - \text{diag}\{\mathbf{s}\}\boldsymbol{\Sigma}^{-1}\,\text{diag}\{\mathbf{s}\}. \end{aligned}$$

The resulting knockoffs and the original variables have then the joint distribution (4.2) which satisfies the exchangeability condition.
In summary, the Gaussian knockoff case offers an explicit and efficient procedure for the knockoff construction without relying on the SCIP algorithm.

4.2.4 Approximate construction: second-order MX knockoffs

Definition 4.1 requires equality in distribution of $(X, \tilde{X})_{\text{swap}(S)}$ and $(X, \tilde{X})$, i.e. all moments have to be equal. Only in the case of a multivariate normal distribution, where an explicit knockoff construction procedure exists, matching the first two moments already ensures equality in distribution. For a general distribution F_X, we usually do not have such

an explicit procedure. A way of circumventing the SCIP algorithm is to weaken the exchangeability (4.1) by asking that only the first two moments of $(X, \tilde{X})_{\text{swap}(S)}$ and $(X, \tilde{X})$ are the same.

Definition 4.3 (**Approximate model-X knockoffs**). *Let $X = (X_1, \ldots, X_p)$ and $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)$ be the random vectors of the true covariates and knockoffs. Their joint distribution satisfies the pairwise exchangeability in a loose sense if $(X, \tilde{X})_{\text{swap}(S)}$ and $(X, \tilde{X})$ have the same mean and covariance. For the latter, equality is ensured by*

$$\text{Cov}(X, \tilde{X}) = \mathbf{G}, \quad \textit{with} \quad \mathbf{G} = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \text{diag}\{\mathbf{s}\} \\ \boldsymbol{\Sigma} - \text{diag}\{\mathbf{s}\} & \boldsymbol{\Sigma} \end{bmatrix},$$

where $\mathbf{s} \in \mathbb{R}_+^p$ has to ensure again positive semidefinitness of $\mathbf{G}$.

As a byproduct of Definition 4.3, we recognize the dependence requirements of Section 3.2.1 but not formulated as geometrical conditions:

i.) $\text{Cov}(\tilde{X}_k, \tilde{X}_j) = \text{Cov}(X_j, X_k)\ \forall j \neq k$: Two distinct knockoffs $(\tilde{X}_j, \tilde{X}_k)$ have the same covariance as their original counterparts (X_j, X_k).

ii.) $\text{Cov}(\tilde{X}_j, \tilde{X}_k) = \text{Cov}(X_j, X_k)\ \forall j \neq k$: Two distinctly original and knockoff variables $(X_j, \tilde{X}_k)$ have the same covariance as the corresponding original variables (X_j, X_k).

This weaker form of exchangeability is practically ensured by approximating the distribution of X (and $\tilde{X}$) as multivariate normal. By default, the `R knockoff` package creates second-order MX knockoffs by estimating the mean vector $\hat{\boldsymbol{\mu}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$ from the data before optimizing for $\mathbf{s}$. After finding the optimal $\mathbf{s}$, approximate Gaussian knockoffs are created with the estimated model equivalent to Section 4.2.3, that is obtaining the first two conditional moments $(\hat{\boldsymbol{\mu}}', \hat{\mathbf{V}})$ and sampling knockoffs from the conditional distribution $F_{\tilde{X}|X} = \mathcal{N}(\hat{\boldsymbol{\mu}}', \hat{\mathbf{V}})$.
We can use the same optimization strategies for $\mathbf{s}$ as in the FX knockoff design: either applying the equi-correlated or the SDP construction. Although both approaches are reasonable for small and moderate p, they cause new difficulties when p is large.

i.) Equi-correlated knockoffs problem: The minimal eigenvalue $\lambda_{\min}(\boldsymbol{\Sigma})$ that is used for the construction will usually be very small when p is large. Even though the approach is fast in high dimensions, it results in low power because the original and knockoffs variables will be very alike for a given index j.

ii.) SDP knockoffs problem: Although it is more powerful than the equi-correlated construction in low to moderate p settings, the algorithm becomes computationally intractable for large p.

Candès et al. (2018) generalized the two approaches above by the approximate semidefinite program (ASDP), consisting of two steps:

i.) Let $\boldsymbol{\Sigma}_{\text{approx}}$ be an approximation of $\boldsymbol{\Sigma}$, then solve the same SDP convex problem

$$\min \|\text{diag}\{\boldsymbol{\Sigma}_{\text{approx}}\} - \mathbf{s}\|_1 \quad \text{s.t.} \quad \begin{array}{l} \text{diag}\{\mathbf{s}\} \succeq \mathbf{0} \\ 2\boldsymbol{\Sigma}_{\text{approx}} - \text{diag}\{\mathbf{s}\} \succeq \mathbf{0}. \end{array}$$

ii.) Solve the problem by bisection search over $\gamma \in [0, 1]$

$$\max \gamma \quad \text{s.t.} \quad \text{diag}\{\gamma \hat{\mathbf{s}}\} \preceq 2\boldsymbol{\Sigma},$$

and select $\mathbf{s}^{\text{ASDP}} = \gamma \hat{\mathbf{s}}$.

Choosing $\mathbf{\Sigma}_{\text{approx}} = \mathbf{I}$ trivially results in the equi-correlated construction, with $\hat{s}_j = 1$ and $\gamma = 2\lambda_{\min}(\mathbf{\Sigma}) \wedge 1$, while $\mathbf{\Sigma}_{\text{approx}} = \mathbf{\Sigma}$ equals the SDP construction, with $\hat{\mathbf{s}} = \mathbf{s}^{\text{SDP}}$ and $\gamma = 1$. Without going too much into the technical details, the ASDP optimization profits from choosing $\mathbf{\Sigma}_{\text{approx}}$ to be an m-block-diagonal approximation of $\mathbf{\Sigma}$ which splits step 1 into m smaller SDP problems that can be parallelized. The `R` implementation runs by default SDP for $p \leq 500$ and ASDP for $p > 500$ (Patterson and Sesia 2020).
To summarize, the current implemented knockoff sampler generates exact knockoffs if X is multivariate normal distributed with a known mean vector and covariance matrix. Otherwise, in case of an unknown distribution, the knockoff sampler estimates the first two moments from data and uses a multivariate normal as approximated distribution. The exchangeability holds only in a loose sense, and the created knockoffs are only approximately valid. However, with an approximated exchangeability property, theoretical FDR control at q is not exactly guaranteed anymore. The approximative behaviour is only superficially investigated in Candès et al. (2018).

4.2.5 Robustness

Barber et al. (2020) investigate in a follow-up work the previously mentioned violation. They examine the effect on the FDR control of MX knockoffs when the researcher only has approximate knowledge of F_X rather than exact. Since this section works with swaps of a single variable instead of a subset, we re-formulate the pairwise exchangeability condition.

Proposition 4.4. *The pairwise exchangeability condition* (4.1) *holds for a set* $S \subset \{1, \dots, p\}$ *if and only if*

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

holds for all j *of* S.

We ask whether a single variable and its knockoff are exchangeable inside the joint distribution, including all other variables X_{-j} and $\tilde{X}_{-j}$ as well. We will denote the true feature joint distribution as F_X and the assumed or estimated feature joint distribution as $\hat{F}_X$. With observed data, $F_j(\cdot|\mathbf{X}_{-j})$ is the true conditional distribution of variable j and $\hat{F}_j(\cdot|\mathbf{X}_{-j})$ the assumed or estimated equivalent.[5] If $F_j \approx \hat{F}_j$, the pairwise exchangeability for j does only hold approximately, that is X_j and $\tilde{X}_j$ are only approximately exchangeable with respect to the true covariate distribution. The more similar the conditionals F_j and $\hat{F}_j$ are, the better the knockoff distribution mimics the feature distribution, and so the more reasonable they are as negative controls. The discrepancy between F_j and $\hat{F}_j$ can be measured by

$$\widehat{\text{KL}}_j := \sum_i \log\left(\frac{F_j(\mathbf{X}_{ij}|\mathbf{X}_{i,-j}) \cdot \hat{F}_j(\tilde{\mathbf{X}}_{ij}|\mathbf{X}_{i,-j})}{\hat{F}_j(\mathbf{X}_{ij}|\mathbf{X}_{i,-j}) \cdot F_j(\tilde{\mathbf{X}}_{ij}|\mathbf{X}_{i,-j})}\right)$$

the observed Kullback-Leibler (KL) divergence of $(\mathbf{X}_j, \tilde{\mathbf{X}}_j, \mathbf{X}_{-j}, \tilde{\mathbf{X}}_{-j})$ and $(\tilde{\mathbf{X}}_j, \mathbf{X}_j, \mathbf{X}_{-j}, \tilde{\mathbf{X}}_{-j})$, where

$$\mathbb{E}[\widehat{\text{KL}}_j] = D((\mathbf{X}_j, \tilde{\mathbf{X}}_j, \mathbf{X}_{-j}, \tilde{\mathbf{X}}_{-j})||(\tilde{\mathbf{X}}_j, \mathbf{X}_j, \mathbf{X}_{-j}, \tilde{\mathbf{X}}_{-j}))$$

[5] We slightly abuse the notation because $F_j(\cdot|\mathbf{X}_{-j})$ and $\hat{F}_j(\cdot|\mathbf{X}_{-j})$ are either the conditional probability mass functions or conditional densities.

is the usual KL divergence according to Definition A.4 between the two distributions of $(\mathbf{X}_j, \tilde{\mathbf{X}}_j, \mathbf{X}_{-j}, \tilde{\mathbf{X}}_{-j})$ and $(\tilde{\mathbf{X}}_j, \mathbf{X}_j, \mathbf{X}_{-j}, \tilde{\mathbf{X}}_{-j})$. To see why this holds, we refer to Barber et al. (2020, p. 1415–1416). In the case where F_X is known, MX knockoffs satisfy the exchangeability property. To put it in other words, Proposition 4.4 implies $(\mathbf{X}_j, \tilde{\mathbf{X}}_j, \mathbf{X}_{-j}, \tilde{\mathbf{X}}_j) \stackrel{d}{=} (\tilde{\mathbf{X}}_j, \mathbf{X}_j, \mathbf{X}_{-j}, \tilde{\mathbf{X}}_j)$, and since $F_j = \hat{F}_j$, we always have $\widehat{\mathrm{KL}}_j = 0$. When we use an approximate covariate distribution, the observed KL divergence measures to which degree the pairwise exchangeability for covariate j is violated. The following theorem incorporates this violation and describes the inflated FDR control when we use approximate distributions $\hat{F}_j$.

Theorem 4.5 (**Inflated FDR control**). *For any $\epsilon \geq 0$, consider all null features for which $\widehat{\mathrm{KL}}_j \leq \epsilon$ holds. The knockoff+ filter controls the FDR corresponding to these nulls by*

$$\mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{H}_0 \cap (\widehat{\mathrm{KL}}_j \leq \epsilon)|}{|\hat{\mathcal{S}}| \vee 1}\right] \leq q \cdot e^{\epsilon}.$$

This implies that the FDR is upper bounded by

$$\mathrm{FDR} \leq \min_{\epsilon \geq 0} \left\{ q \cdot e^{\epsilon} + \mathbb{P}\left(\max_{j \in \mathcal{H}_0} \widehat{\mathrm{KL}}_j > \epsilon \right) \right\}.$$

The proof works similarly as the sketch of proof explained in Section 3.7 but with approximate statements (see Barber et al. (2020) for a detailed explanation). Moreover, an equivalent version exists for the knockoff filter with inflated modified FDR control. The theorem implies that it is sufficient to control the $\widehat{\mathrm{KL}}_j$'s for FDR control. It bounds the false positives originating from those null features with a sufficiently small observed KL divergence. If $\max_{j=1,\ldots,p} \widehat{\mathrm{KL}}_j$ is small with large probability, then we have FDR control close to q. In other words, if we find estimated conditional distributions $\hat{F}_j$ that are close enough to the true ones F_j in the sense that the $\widehat{\mathrm{KL}}_j$'s are small enough, then the knockoff procedure still achieves reasonable FDR control. In the ideal case where F_X is known, the observed KL divergences are zero for all null variables, and we get FDR control at q by taking $\epsilon = 0$ above.
Another important detail of Theorem 4.5 is the following: We can not expect to have FDR control for all null variables if some of their knockoff copies are not even close to mimic the distribution of X. In fact, we restrict our attention only to the null variables with reasonable knockoff controls, i.e. that have small $\widehat{\mathrm{KL}}_j$, and we achieve FDR control for those. Without making further assumptions, this is all we can get from the theory side. While the exact MX setting is based on the assumption that knockoffs are exact controls for each null feature, and so achieve FDR control, the approximate MX framework provides FDR bounds when counting only the null features with reasonable knockoffs.
Although the theorem is purely theoretic, it has the following practical implications:

- *For a general knockoff sampler*: If the covariate distribution F_X can be estimated accurately by $\hat{F}_X$, our FDR control will be reasonably close to q, and the MX knockoff filter is still an effective method for FDR control. The statement assumes the availability of a valid knockoff sampler.
- *Gaussian knockoffs:* If X is multivariate normal but with a general and unknown covariance matrix, the FDR control is still reasonable as long as the covariance matrix can be accurately estimated.

- *Current second-order MX knockoffs implementation*: If the covariate distribution can be well approximated by a multivariate normal distribution, then Thoerem 4.5 implies that the MX knockoff filter provide reasonable FDR control.

On a high level, this result can be interpreted as evidence for the robustness of the MX knockoff filter when we do not know F_X exactly but can estimate it precisely.
Finally, we want to elaborate on the conservativeness of the presented FDR inflation. Theorem 4.5 holds for any valid score statistic W_j. Barber et al. (2020) emphazise that under this generality, the FDR bound might be more pessimistic than the actual FDR inflation. The intuition behind this is the following: The $\widehat{\mathrm{KL}}_j$'s measure how distinguishable $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ are when viewing the full data, and we require low values $\widehat{\mathrm{KL}}_j$. However, the knockoff filter is based on $(W_1, \dots, W_p)$, and we require that null and knockoff variables are indistinguishable concluded from looking at W_j, which is only a broad summary of the full data and contains less information than the KL divergence. Hence, the KL divergence could be a too strict measure: low values for $\widehat{\mathrm{KL}}_j$ are sufficient but not necessary for reasonable FDR control. For a pre-defined W_j, the actual FDR inflation might be less severe, but there is no theory for that so far.[6]

4.3 Exchangeability properties

We will briefly cover the differences and similarities between the underlying properties of the MX and FX knockoffs, with an overview presented in Table 4.1. A property that directly follows from Definition 4.1 is the pairwise exchangeability of the null features with their knockoffs without affecting the joint distribution of the augmented vector given the response.

Lemma 4.6 (**Pairwise exchangeability of nulls**). *For any* $S \subset \mathcal{H}_0$, $(X, \tilde{X})|Y \overset{d}{=} (X, \tilde{X})_{\mathrm{swap}(S)}|Y$.

As we already pointed out in Section 4.2.3, the PEF property (3.7) of the fixed design resembles the pairwise exchangeability (4.1) for the empirical covariance matrix in a deterministic sense, ensuring its invariance under variable-knockoff swaps. Moreover, we also see the similarity between Lemma 4.6 and the PER property (3.8) of the FX design. Note that Lemma 4.6 also holds for the joint distribution of $(X, \tilde{X}, Y)$, because the marginal distribution of Y is the same on both sides. In addition to these exchangeability requirements, we also have to make again assumptions about the feature importance statistics and the scores. Similar to the FX design, we assume the fairness requirement for $z(\cdot)$/ antisymmetry property for $w_j(\cdot)$. With all these satisfied, the most important properties why knockoffs work at all can be derived for the MX knockoffs as well: the pairwise exchangeability of the importance statistics (Lemma 3.3) and the coin-flipping property (Lemma 3.4). An important difference is that, in contrast to the FX knockoff filter, MX knockoffs do not require the sufficiency property on $\mathbf{W}$ anymore, which implied that $\mathbf{W}$ is only a function of $[\mathbf{X}\ \tilde{\mathbf{X}}]^\top[\mathbf{X}\ \tilde{\mathbf{X}}]$ and $[\mathbf{X}\ \tilde{\mathbf{X}}]^\top\mathbf{y}$. We will discuss the consequences of this small detail in the next section.

[6] See Barber et al. (2020, p. 1420–1421) for a more detailed discussion.

Table 4.1: MX and FX knockoffs, properties comparison

MX knockoffs		FX knockoffs
Pairwise exchangeability (4.1)	$\approx$	Pairwise exchangeability features (3.7)
Pairwise exchangeability nulls (Lemma 4.6)	$\approx$	Pairwise exchangeability response (3.8)
Antisymmetry property (3.6)	$=$	Antisymmetry property (3.6)
–	$\times$	Sufficiency property (3.5)
Pairwise exchangeability Z_j (Lemma 3.3)	$=$	Pairwise exchangeability Z_j (Lemma 3.3)
Coin-flipping property (Lemma 3.4)	$=$	Coin-flipping property (Lemma 3.4)

We use $\approx$ to indicate similar properties/assumptions and $=$ for equal ones.

4.4 Step 2: Measure and compare variable importance

Our simulations in Section 3.9.3 have shown the importance of choosing appropriate score functions W_j since they can considerably increase the power of the selection. The MX framework allows for a greater variety of score functions W_j than the FX setting for two reasons: First, the removal of the sufficiency property allows more flexibility. Second, the MX framework itself does not restrict the reader to Gaussian linear models, and so she can use more sophisticated estimation models. We will discuss some examples of promising score functions subsequently.

Taking the Lasso regression with the LCD statistic as an example, due to the removal of the sufficiency property, we do not require a fixed λ as in the FX design anymore. We can tune the hyperparameter by cross-validation and take the absolute cofficient differences

$$W_j = |\hat{\beta}_j(\lambda_{\text{CV}})| - |\hat{\beta}_{j+p}(\lambda_{\text{CV}})|. \tag{4.3}$$

We would expect that this score function achieves greater power than with a fixed λ due to its optimization. The MX setting also allows for more complicated penalization methods such as Adaptive Lasso with CV (Zou 2006), which provides consistent variable selection under milder conditions than vanilla Lasso. In addition, the weighted penalty $\lambda_j = w_j\lambda$ can reduce the bias of the Lasso coefficients, with smaller weights for larger coefficients. A more accurate estimation of $\hat{\beta}$ improves the variable selection and might result in a knockoff filter with a more liberal threshold T and greater power. Moreover, the researcher could also use Elastic Net (Zou and Hastie 2005) with tuned hyperparameters which could be superior in the case of highly correlated variables. To be precise, both penalization methods could also be applied in the FX design but only in a very restrictive way and without CV to ensure the sufficiency property.

Moving on to non-Gaussian models, besides the possibility of fitting GLMs, the MX framework also allows using a wide range of modern machine-learning models such as random forests. An already implemented score function in the `R knockoff` package is the difference of the absolute variable importances $|Z_j|$ and $|\tilde{Z}_j|$, where Z_j is the total decrease of the impurity measure (residual sum of squares for regression, Gini index for classification) due to splits on that variable j averaged over all trees (Patterson and Sesia 2020). MX knockoffs, therefore, provide some interpretability of black-box machine learning approaches where a well-developed inferential theory does not exist. This is a huge advantage from a (classic) statistician's view who is not only interested in the lowest prediction error but also in reproducible and interpretable results.

A complete different approach to construct score statistics comes from a Bayesian perspective. If the researcher has some prior information, she can encode this within a Bayesian

framework, and applies the score statistic

$$W_j = Z_j - \tilde{Z}_j, \quad Z_j = \mathbb{P}(\beta_j \neq 0 | \mathbf{y}, [\mathbf{X}\ \tilde{\mathbf{X}}]),$$

where Z_j and $\tilde{Z}_j$ are the posterior probabilities that the jth original and knockoff coefficient are non-zero respectively. Candès et al. (2018) show that the Bayesian approach achieves greater power than the LCD score based on CV when accurate prior information is available. Even if the prior is wrong, the MX knockoff filter still controls the FDR but with low power.
Taking the construction of score functions one step further, independent of the feature importance statistics Z_j, we can also vary between different anti-symmetric functions w_j: $Z_j^2 - \tilde{Z}_j^2$, $Z_j \vee \tilde{Z}_j \cdot \text{sign}(Z_j - \tilde{Z}_j)$, or $\log(|Z_j|) - \log(|\tilde{Z}_j|)$ are also possible choices.
The key takeaway of this section is that MX knockoffs can be combined with almost any data-fitting algorithm as long as the fairness condition hold, and it provides rigorous FDR control in finite samples even with black-box models. The variety of possible score statistics seems endless, and it is not trivially clear to which the researcher should stick.

4.5 Discussion and open research questions

We start by briefly summarizing the key differences between FX and MX knockoffs (Table 4.2). The FX design assumes that the response is linked to the covariates through a homoscedastic Gaussian linear model, whereas the MX knockoff filter does not assume anything on that relationship, so it can be considered as model-free. Hence, the researcher can apply knockoffs beyond the linear model such as to GLMs or random forests. The large model flexibility implies a much wider range of possible score functions W_j than in the FX setting, probably leading to power increases. All this comes at a cost of knowing the covariate distribution F_X, or at least an accurate estimation of it, whereas the FX design treats the covariates as fixed. Depending on the data set, this might be a more reasonable assumption than requiring a homoscedastic Gaussian linear model.
In addition, MX knockoffs can be applied to high-dimensions, which is not the case for FX knockoffs. In a subsequent work, Barber and Candès (2019) extend the classic FX to the high-dimensional framework by splitting the data into two groups. The first one is used to fit the variable selection method and pick a set of possible relevant features, while the second data part is used to conduct inference on this pre-selected subset of variables. We will not further discuss this approach and refer to their work for more details.
Last but not least, the removal of the sufficiency condition enables the researcher to use tuned hyperparameters by CV inside their scores, which leads to power increases compared to fixed hyperparameters that she has to choose in the FX design. Finally, Figure 4.1 extends the flowchart of Section 3.6 and summarizes the key steps of the MX knockoff filters.

Table 4.2: Comparison MX vs. FX design

	MX design	FX design
$F_{Y\|X}$	arbitrary	Gaussian
F_X	known	fixed
Dimensionality	no restrictions	$n \geq p$
Choice W_j	enormous choices	restricted choices

As already admitted by Candès et al. (2018, p. 575), their "paper may pose more problems than it solves". We want to discuss some open research questions and present related developments in these directions.

MX knockoff construction: The SCIP algorithm provides a general but impractical method to construct knockoffs. An exact knockoff sampler exists for the Gaussian case, and also the approximate second-order knockoffs are reasonable if the estimated distribution can be well-approximated by a Gaussian distribution (Barber et al. 2020). On the one hand, it would be interesting to derive exact knockoff constructions for other feature distributions F_X. First results exist for Markov Chains and Hidden Markov Models (Sesia et al. 2018), Gaussian Mixture Models (Gimenez et al. 2019) and Graphical Models (Bates et al. 2020). On the other hand, the current knockoff constructions aim to minimize the pairwise correlations between each variable and its knockoff by the choice of $\mathbf{s}$. This might not be the most powerful approach of generating valid knockoffs. Spector and Janson (2021) question the correlation as a minimization metric and propose a new optimization measure, the so-called reconstructability of the features, to create more powerful knockoffs. A different research direction is the construction of other approximate knockoff samplers. Romano et al. (2019) propose a new sampling procedure to approximate knockoffs based on deep learning techniques by also matching higher moments. Although their Deep knockoffs are more powerful and robust than second-order knockoffs, they suffer from their computational intensity. Clearly, the knockoff construction is a research area with much potential for improvement.

Investigation of second-order knockoffs: Although Barber et al. (2020) derive theoretical results for the robustness of second-order knockoffs, and Romano et al. (2019) present two examples where the second-order approximation significantly fails, there is still room for the investigation of further scenarios by simulations. Especially because Barber et al. (2020) admit the conservativeness of their bounds, so the actual FDR inflation might be lower. More concrete examples are needed to examine how large the FDR inflation may be under certain deviations from the Gaussian distribution in practice. This would give researchers a better sense of when to rely on second-order knockoffs.

Comparison of FX and MX knockoffs: Both knockoff procedures rely on different distributional assumptions. In addition, MX knockoffs provide a much more general range of applications. Restricting the attention to the low-dimensional linear case, Candès et al. (2018) showed in one of their simulations the superiority in power of MX over FX knockoffs but based on different score statistics W_j (MX: LCD with λ_{CV}, FX: LLSM). Since LCD with λ_{CV} is a more optimized score statistic, it obviously is more powerful than LLSM. They also generated the features from a multivariate normal distribution, for which an exact knockoff sampler exists. But does the MX knockoff approach achieve greater power than the FX construction in a linear model if we use the same score function and generate covariates a) from a Gaussian distribution or b) from a non-Gaussian distribution? For the latter, we already know that the MX filter leads to an inflated FDR control and also potential power losses. If the MX approach is not uniformly superior, how large can the deviation from the true Gaussian be? Such results provide evidence if the initial approach by Barber and Candès (2015) is completely outdated or still useful under some scenarios. So far, no research examines this in more detail.

Importance measures: In Section 4.4, we have already mentioned the endless possibilities to construct score statistics W_j. While the knockoff filter achieves FDR control for any importance feature, its power will depend on the choice of W_j. It would be fruitful to

compare different importance statistics regarding their power under certain settings and to explore new ones that lead to power increases. Except of Gimenez et al. (2019), who propose a new score the so-called Swap Integral, there has been not much research done in this direction.

Multiple MX knockoffs: The construction of multiple knockoffs $(\tilde{\mathbf{X}}_j^{(1)}, \ldots, \tilde{\mathbf{X}}_j^{(m)})$ for each variable $\mathbf{X}_j$ could increase the method's power because it incorporates more information. For example, among the multiple knockoffs, we could evaluate the rank for which an original feature enters into the model. However, multiple control features could constrain the knockoff construction by itself. In the MX design, for example, some sort of "extended exchangeability property" has to be satisfied: swapping original variables with their knockoffs but also multiple knockoffs with each other does not affect the joint distribution of $(X, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(m)})$. This is undoubtedly harder to fulfill than the "normal" exchangeability property.[7]
A much simpler approach would be to run the knockoff procedure multiple times with different knockoff matrices $\tilde{\mathbf{X}}$, each run yielding its own set of selected variables. It would be interesting to examine if we could aggregate all these sets into one final set to increase the power while maintaining FDR control. Xie and Lederer (2021) aggregate K different knockoff runs with nominal levels $q_1, \ldots, q_K \in [0, 1]$ and $\sum_{k=1}^{K} q_k = q$. The union of the K variable sets leads to power improvements while controlling the FDR at q compared to one knockoff run with q. However, their simulation setting is not tailored to the MX knockoff framework since they treat the data as fixed. Nguyen et al. (2020) derive so-called intermediate p-values π_j from the scores W_j for each of the B knockoff draws and summarize them via quantile aggregation as in Meinshausen et al. (2009). To control the FDR, they apply BH on the aggregated p-values. While their approach has some theoretical weaknesses and violations, it seems to perform slightly better than vanilla knockoffs in simulations, even though the superiority holds only for some parameter settings. Furthermore, they only examine extremely high-dimensional models, and the performance of their method could differ for low-dimensional ones. Gui (2020) proposed the aggregation method ADAGES which achieves a larger power than the previous two aggregation methods. However, he does only study a distributed learning environment, and he does not investigate many parameters settings in his simulation. A comparison of ADAGES with the vanilla MX knockoff filter is also missing. Without further in-depth comparisons, we cannot conclude if one of these aggregated knockoff methods is a useful extension to vanilla MX knockoffs. In addition, none of these methods provides an accessible and user-friendly implementation in `R`, which offers potential for improvement, at least from a practical viewpoint.[8]

In the subsequent section, we will analyse the third open research question by comparing FX and MX knockoffs when the underlying covariate distribution F_X deviates from a multivariate Gaussian one. In Chapter 5, we will try to answer the last open research question about multiple knockoffs. After we have presented the three aggregation methods, we will contribute to the knockoff literature by providing user-friendly implementations of them and comparing the three multiple knockoff procedures in an extensive simulation study with each other.

[7] The literature refers to these approaches as simaltaneous knockoffs. First ideas are developed by Gimenez and Zou (2019) and Emery et al. (2020).

[8] An approach from Candès itself, the derandomized knockoffs, aggregates knockoff draws similar to the stability selection by Meinshausen and Bühlmann (2010). However, it controls either the per family error rate or the k-FWER instead of the FDR (Ren et al. 2020).

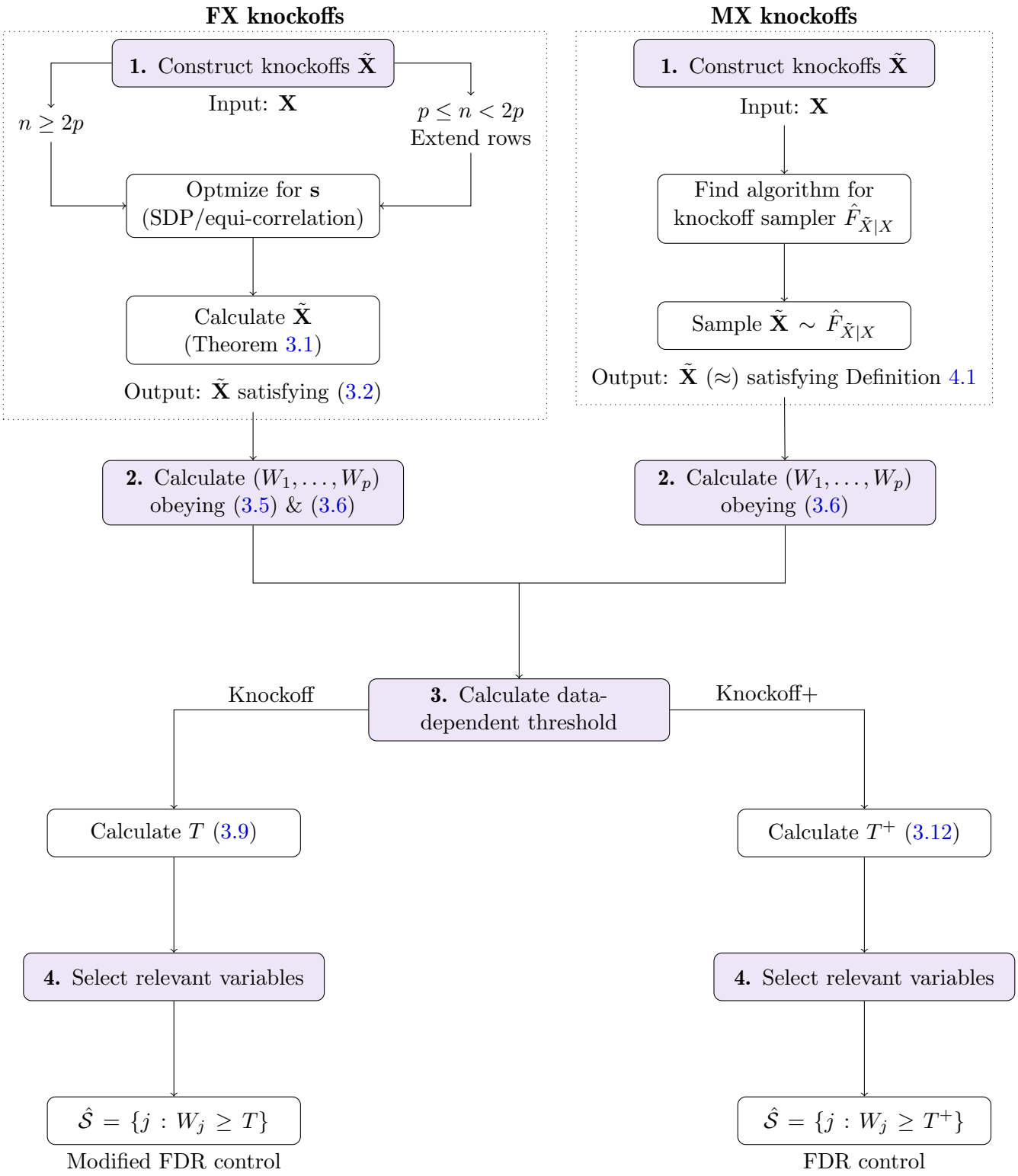


Figure 4.1: Flowchart: Fixed-X and model-X knockoffs

4.6 Key conclusions: FX vx. MX knockoffs in a linear model

In this last section, we want to contribute some findings to answer the (second and) third open research question of Section 4.5. While we will present the simulation design and the key conclusions here only superficially, the full simulation details and results are described in Appendix A.5.

MX knockoffs (Candès et al. 2018) are often seen as the extension of FX knockoffs (Barber and Candès 2015) due to their wider range of score functions and applications. Therefore, the majority of recent theoretical research (e.g. Barber et al. (2020); Huang and Janson (2020); Bates et al. (2020)) as well as empirically-driven publications (e.g. Yu et al. (2021); Chia et al. (2021); Xie and Lederer (2021)) deals with the MX knockoff filter. As discussed in Section 4.5, FX and MX knockoff filters rely on different distributional assumptions. The FX knockoff filter requires a homoscedastic low-dimensional linear model but places no probabilistic assumptions on the covariates in contrast to MX knockoffs. The most popular approach of generating MX knockoffs is the second-order construction which performs well when the covariates are (close to) Gaussian. However, if the origin of the covariates is far from the normality assumption, the FDR control of MX knockoffs might be inflated and we have to expect power losses. So far, there has been no research conducted that shows that MX knockoffs are superior to FX knockoffs in a low-dimensional linear model but under various deviations from the normality assumption of the covariates.

The goal of our analysis is to examine whether FX knockoffs are completely outdated or still useful under some model settings where the second-order MX construction suffers from power losses due to model misspecifications. We will study four different types of model misspecifications in our simulations:

1.) Multivariate t-distribution.

2.) Discretized multivariate normal distribution.

3.) Multivariate skew-normal distribution.

4.) Multivariate normal distribution but with a misspecified $\mathbf{\Sigma}$.

Within those four settings, will vary the degree of misspecification respectively. We will run the FX and MX knockoff filter on the same score, namely the LLSM statistic

$$W_j = Z_j \vee \tilde{Z}_j \cdot \text{sign}(Z_j - \tilde{Z}_j), \quad \begin{aligned} Z_j &= \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\} \\ \tilde{Z}_j &= \sup\{\lambda : \hat{\beta}_{j+p}(\lambda) \neq 0\}. \end{aligned}$$

Additionally, we also perform the MX knockoff filter with the LCD score

$$W_j = |\hat{\beta}_j(\lambda_{\text{CV}})| - |\hat{\beta}_{j+p}(\lambda_{\text{CV}})|,$$

because it is the most applied score, and it has the highest power of all Lasso related statistics.[9]

Turning to the main findings of our analysis, MX knockoffs with the LCD score have a (notably) higher power than FX knockoffs, even when the underlying covariate distribution has wider tails, is discretized, skewed or with estimation errors of the covariance matrix. The FDR was controlled in all model misspecification settings, except when we provide a misspecified $\hat{\mathbf{\Sigma}}$ with underestimated entries compared to the true covariance matrix.

[9] Remember that the LCD score with CV cannot be used with FX knockoffs because the sufficiency property is not satisfied.

Since we know that the LCD statistics with cross-validated λ is more powerful than LLSM, we also compared both knockoff filters based on the LLSM score. Surprisingly, the MX LLSM filter performs much worse than its FX equivalent, sometimes with a power that was more than twice as low. We also tested the LLD and OMP score that we have already investigated in Section 3.9.3, and we observed the same power behaviour of MX and FX knockoffs as with LLSM. This observation might be crucial when future research investigates new scores since they might not perform equally well for FX and MX knockoffs. In summary, MX knockoffs perform better than their FX predecessor. We advise the reader to apply MX knockoffs with the LCD score over FX knockoffs in a low-dimensional linear model, even if the underlying distribution of the covariates is not Gaussian. The method seems to be quite robust to deviations from the normality assumption. Based on our simulation results, we can consider the FX knockoffs to be outdated. It would be also interesting to extend this analysis to high-dimensional linear models by comparing MX knockoffs with the high-dimensional FX knockoff filter (Barber and Candès 2019). We leave this question open for future research.

Chapter 5

Multiple knockoffs

Even though we work on a single data set in practice, we always prefer a procedure with lower variability in FDP, such that it is close to its expectation. We have already mentioned the variability of the FDP in the empirical Bayes part (Section 2.3.5). FDR control does not imply that the FDP lies below the same nominal level. The higher the variability of this random variable, the less confident we can be about our final variable selection. The variability comes from two sources: the randomness from i.) the data sample $(\mathbf{X}, \mathbf{y})$ and ii.) the probabilistic MX knockoff construction $\tilde{\mathbf{X}}$. Running the knockoff algorithm multiple times on the *same* data results in different knockoffs and probably overlapping but slightly different selection sets $\hat{\mathcal{S}}$. These differences will yield to fluctuating power and FDP values across different runs which make inference more irreproducible. Hence, a knockoff filter with a more stable selection of variables is desirable. One natural extension would be to run the procedure multiple times, each time with a different knockoff matrix, and then aggregate the results in a way such that FDR control is still retained while reducing the variability from the knockoff construction. Such an aggregation might not only stabilize the type-I error but can also lead to power increases. If we interpret each knockoff run as its own learning algorithm, the aggregation of all knockoff runs can be seen as an ensemble learning technique in a very broad sense. Ensemble methods are usually known to improve performance, in our case power, compared to a single model (Rokach 2009). However, most ensemble learning techniques are not applicable to the knockoff setting or do not yield finite sample FDR control, and we need more tailored procedures for that problem (Ren et al. 2020). All of the presented knockoff aggregation schemes below claim to stabilize the outcome and improve the power compared to a single knockoff run while controlling the FDR. We will present each aggregation method and our own implementation in `R`, after which we compare those three methods by extensive simulations. Hence, we contribute to the research community since there are neither user-friendly implementations nor a fair simulation comparison of the aggregation procedures so far.
Section 5.1 introduces the aggregation procedure by Xie and Lederer (2021) that we will denote as union knockoffs. Section 5.2 presents p-value knockoffs by Nguyen et al. (2020). Section 5.3 elaborates on the multiple knockoff method ADAGES by Gui (2020). Section 5.4 starts with a qualitative comparison of the three methods before presenting the simulation results.

5.1 Union knockoffs

5.1.1 Theory

Xie and Lederer (2021) propose union knockoffs (uKO) which consist of running the knockoff filter K times with different (decreasing) nominal levels and combine the resulting selection sets (Algorithm 2). More specifically, they select a sequence of nominal levels $q_1, \dots, q_K$ such that their sum equals the target FDR q that we want to control. Then, they draw knockoffs and run either the knockoff or knockoff+ filter with each q_k, K times in total. Since MX knockoffs are stochastically generated, we will have a different knockoff matrix $\tilde{\mathbf{X}}^{(\mathbf{k})}$ in each run. The resulting K selection sets of the knockoff runs are aggregated by taking their union, which we will denote as $\hat{\mathcal{S}}_q^U$.

Algorithm 2 Union knockoffs

Input: Data $(\mathbf{X}, \mathbf{y})$; number of runs K; nominal level $q \in [0, 1]$

1 Choose sequence $\{q_k\}_{k=1}^K \in [0, 1]$ such that $q = \sum_{k=1}^K q_k$.

2 **for** $k \in \{1, \dots, K\}$ **do**

Generate knockoff matrix $\tilde{\mathbf{X}}^{(\mathbf{k})}$.

Apply knockoff/knockoff+ on $([\mathbf{X}\ \tilde{\mathbf{X}}^{(\mathbf{k})}], \mathbf{y})$ with q_k and obtain selection set $\hat{\mathcal{S}}_{q_k}$.

end for

Result: K different selection sets $\hat{\mathcal{S}}_{q_1}, \dots, \hat{\mathcal{S}}_{q_K}$.

3 Aggregate the selection sets by taking their union

$$\hat{\mathcal{S}}_q^U := \bigcup_{k=1}^K \hat{\mathcal{S}}_{q_k}.$$

Output: Aggregated set $\hat{\mathcal{S}}_q^U$

The following theorem summarizes that uKO still retain (modified) FDR control at q.

Theorem 5.1 (**uKO FDR control**). *Let $\{q_k\}_{k=1}^K \in [0, 1]$ be a sequence of nominal levels satisfying $q = \sum_{k=1}^K q_k$, where q is the target FDR of interest. Further let $\hat{\mathcal{S}}_q^U$ be the selected model of the union knockoff+ filter. Then, the union knockoff+ filter guarantees FDR control*

$$\mathbb{E}\left[\frac{|\hat{\mathcal{S}}_q^U \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}_q^U| \vee 1}\right] \leq q.$$

Similarly, if $\hat{\mathcal{S}}_q^U$ is the selected model of the union knockoff filter, modified FDR control is guaranteed, that is

$$\mathbb{E}\left[\frac{|\hat{\mathcal{S}}_q^U \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}_q^U| + q^{-1}}\right] \leq q.$$

Proof union-knockoff+: By Theorem 3.6, each of the K runs achieves FDR control at q_k

$$\mathbb{E}\left[\frac{|\hat{\mathcal{S}}_{q_k} \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}_{q_k}| \vee 1}\right] \leq q_k, \quad k \in \{1, \dots, K\}. \tag{5.1}$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{|\hat{\mathcal{S}}_q^U \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}_q^U| \vee 1}\right] = \mathbb{E}\left[\frac{|(\cup_{k=1}^K \hat{\mathcal{S}}_{q_k}) \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}_q^U| \vee 1}\right] &= \mathbb{E}\left[\frac{|\cup_{k=1}^K (\hat{\mathcal{S}}_{q_k} \cap \mathcal{H}_0|)}{|\hat{\mathcal{S}}_q^U| \vee 1}\right] \\
&\leq \mathbb{E}\left[\sum_{k=1}^K \frac{|\hat{\mathcal{S}}_{q_k} \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}_q^U| \vee 1}\right] \\
&= \mathbb{E}\left[\sum_{k=1}^K \frac{|\hat{\mathcal{S}}_{q_k} \cap \mathcal{H}_0|}{|\cup_{k=1}^K \hat{\mathcal{S}}_{q_k}| \vee 1}\right] \\
&\leq \mathbb{E}\left[\sum_{k=1}^K \frac{|\hat{\mathcal{S}}_{q_k} \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}_{q_k}| \vee 1}\right] \\
&= \sum_{k=1}^K \mathbb{E}\left[\frac{|\hat{\mathcal{S}}_{q_k} \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}_{q_k}| \vee 1}\right] \\
&\leq \sum_{k=1}^K q_k = q,
\end{aligned}$$

where for the first two upper bounds, we use the union bound and the last inequality simply results from (5.1), which is equal to q by defintion. The proof of the modified FDR control works similarly. Theorem 5.1 gives insights into why the sequence of nominal levels has to consist of values $q_k \leq q$. If we applied all K knockoff runs with our target q and then aggregated the results, we would have larger power but an FDR control at Kq. This follows from replacing q_k by q in (5.1). Hence, in line with Theorem 5.1, each individual knockoff run has to be more conservative if their union should control the FDR at q. Although we explain the suggested aggregation procedure with knockoffs, the scheme can be applied to any fitting and selection algorithm that achieves FDR control in order to aggregate multiple selection sets. However, we will restrict our attention to the knockoff framework. By drawing multiple knockoffs and running the selection procedure multiple times, we aim to stabilize the final selection outcome with possible power increases.

The union knockoff procedure requires the user to specify two parameters groups: the number of knockoff runs K and the sequence of nominal levels $q_1, \ldots, q_K$ for each knockoff filter. If we choose $K = 1$, the union knockoff filter reduces to the vanilla knockoff procedure. Xie and Lederer (2021) recommend $K \approx 5$ to balance the effect of higher computational effort and statistical influence on variability, FDR and power. However, they do not further investigate how the power depends on K. Depending on how $q_1, \ldots, q_K$ are specified, more knockoff runs do not necessarily imply higher power. To see this, consider the sequence $q_k = q/K \ \forall k$ and a target FDR of $q = 0.2$. For a large K, say 20, each knockoff filter will be very strict due to its nominal level $q_k = 0.01$, often resulting in no detections at all and therefore low power. A decreasing sequence might suit better to cover knockoff filters with different "strictness-levels". Xie and Lederer (2021) find that $q_k = q/2^{k-1}$ with $K \approx 5$ works well in practice, even though this sequence will not sum up to q. They conclude that the differences between both sequences are rather small in simulations, but for some nominal levels $q_k = q/2^{k-1}$ leads to 10% larger power than $q_k = q/K$ with $K \approx 5$. Their recommended sequence results in higher power since the theoretical FDR bound according to Theorem 5.1 is no longer q but $q(1 + \frac{1}{2} + \ldots + \frac{1}{2^{K-1}})$, i.e. less strict. Although uKO do not theoretically ensure FDR control at q anymore, Xie and Lederer (2021) empirically observe FDR values below q for all their settings. Xie admits in a personal exchange that "This is also a point which we have not found a good

explanation for". We can neither provide a valid argumentation for this observation but a possible intuition in the right direction. We saw in the simulations of Section 3.9.2 that the empirical FDR of the vanilla knockoff+ filter was often considerably lower than q and the FDR of BH, particularly for small sample sizes n or large correlations ρ of a Toeplitz matrix. The conservativeness of knockoffs was also observed by Barber and Candès (2015) and Arias-Castro and Chen (2017). The way the knockoff filter is constructed seems to introduce conservativeness and is one source of why the empirical FDR values of all knockoff-related procedures are often (much) lower than q.[1] An additional source comes from the construction of the decreasing sequence $q_k = q/2^{k-1}$ of union knockoffs itself. The first knockoff run with $q_1 = q$ will result in the same selection set as vanilla knockoffs. All other runs with $q_2, \ldots, q_K$ are more strict due to the decreasing nominal level, and the resulting selection sets tend to consist of only strong signals and fewer false positives. Since the knockoff matrices differ in each run, the selection set S_{q_k} may include new strong signals that were not detected in the previous more liberal knockoff runs $S_{q_1}, \ldots, S_{q_{k-1}}$. At the same time, the smaller q_k becomes, the less likely it is to find many new false positives that were not already in one of the earlier knockoff runs. Thus, even with a theoretical bound that is larger than q, union knockoffs will probably achieve empirical FDR control at q while increasing the power. But we can also expect an empirical FDR value that is slightly larger than the one of vanilla knockoffs. Although the proof of Theorem 5.1 does not rely on any independence assumption, it does not incorporate that the selection sets are dependent and overlapping since they are estimated with similar but not equal data $([\mathbf{X}\ \tilde{\mathbf{X}}^{(\mathbf{k})}], \mathbf{y})$. Somehow accounting for that could improve the theoretical FDR bound. In summary, the optimal values for the number of knockoff runs and the sequence of nominal levels are not well-supported by theory. Besides their two recommendations for K and $\{q_k\}_{k=1}^K$, their choices remain a bit arbitrary to the user.

Xie and Lederer show (for $K = 5$ and $q_k = q/2^{k-1}$) that uKO have larger power than vanilla knockoffs while retaining FDR control, which holds under different dimensionalities, sparsity levels and correlations. The superiority in power is proven for both a Gaussian linear relationship and a logistic regression setting. We encourage the reader to take a look at Appendix A.6, where we provide some concerns related to their simulation.

Besides reducing the variability of selection sets, uKO can also be beneficial for specific data structures. Yu et al. (2021) present two useful biological applications: i.) Assume the data consists of n voxel intensity samples in p anatomical volumes in $n_{sample} = 10$ individuals. They run one knockoff filter with different q_k for each individual and aggregate the $K = 10$ runs by uKO. ii.) Assume an imbalanced classification data set of $n_1 = 1234$ smokers and $n_2 = 15640$ non-smokers for some p variables. They try to reduce the influence of the imbalancedness with uKO by subsampling 1234 observations from the non-smoker group $K = 10$ times and apply the knockoff filter with q_k, before aggregating the selection sets by their union.

5.1.2 Implementation

We are trying to fill the scientific gap of missing multiple knockoffs `R` implementations by introducing our package `multiknockoffs`. Our intention is to provide user-friendly functions of the three presented aggregation schemes such that the user can improve the power compared to vanilla knockoffs. We implement each aggregation method in two possible ways: Either running the multiple knockoff construction, the estimation of the selection sets and their aggregation by executing one function, or performing each of the

[1] See Barber and Candès (2015, p. 2083) for a possible explanation of what causes the conservativeness.

steps manually. The latter approach offers a greater flexibility. For instance, the user can solely apply the aggregation step to other variable selection methods beyond the knockoff framework. In the following, we will discuss the "all-in-one" approach, while the manual steps are illustrated in Appendix B.2.
First, the user can load our package `multiknockoffs` in `R` directly from the GitHub repository by the following code chunk:

```
library(devtools)
install_github("cKarypidis/multiknockoffs")
```

A list of all implemented functions is provided in Appendix B.1. Besides the explanations of the functions given in the following work, each function has its own documentation in `R`, which can be either called by typing "`?`" in front of the function name or by looking directly at the help menu of the package. Figure B.1 presents as an example the documentation of the function `run.uKO` that we will discuss now.
We implement Xie and Lederer's (2021) knockoff aggregation scheme with `run.uKO`, which constructs K knockoff matrices, runs K knockoff filters at different nominal levels q_k and aggregates the resulting selection sets by taking their union.

```
run.uKO(X, y, knockoffs = create.second_order,
        statistic = stat.glmnet_coefdiff, qk = "decseq",
        q = 0.2, K = 5, q_seq = NULL, offset = 1 , sets = FALSE)
```

The parameters that can be supplied to the function are:

- `X` the $n \times p$ design matrix and `y` and the $n \times 1$ response vector.
- `knockoffs` is the function for the knockoff construction. It must take the $n \times p$ data matrix as input and it must return a $n \times p$ knockoff matrix. The user can either choose a knockoff sampler of the `knockoff` package or define it manually. Default: `create.second_order` (see below).
- `statistic` is a function that computes the scores W_j. It must take the data matrix, knockoff matrix and response vector as input and it outputs a vector of computed scores. The user can either choose one score statistic from the `knockoff` package or define it manually. Default: `stat.glmnet_coefdiff` (see below).
- `qk` the sequence of nominal levels. The user can choose between the options `"decseq"` (default) for $q_k = q/2^{k-1}$ or `"ave"` for $q_k = q/K$.
- `q` is the nominal level for the FDR control. Default: 0.2.
- `K` is the number of knockoff runs. Default: 5.
- `q_seq` to define an own sequence supplied by the user, which has to match in length with the number of knockoff runs `K`. If this argument is specified, `qk` and `q` are ignored.
- `offset` either 0 (knockoff) or 1 (knockoff+). Default: 1.
- `sets` logical argument if the K selection sets of each knockoff run before the aggregation should be returned. Default: `FALSE`.

The argument `create.second_order` refers to the approximate second-order knockoff construction (see Section 4.2.4), which is also the default option in the `knockoff` package. Moreover, the default `stat.glmnet_coefdiff` corresponds to the LCD statistic (4.3), where the coefficients are based on a cross-validated λ. The function `run.uKO` outputs a list with the aggregated set $\hat{\mathcal{S}}_q^U$, the number of specified knockoff runs K, the theoretical FDR bound $\sum_{k=1}^{K} q_k$ and (if specified) the individual selection sets of each knockoff run

$\hat{\mathcal{S}}_{q_1}, \ldots, \hat{\mathcal{S}}_{q_K}$. We finish this section with a small numerical example on which we apply `run.uKO`. We generate data with $(n, p) = (400, 200)$ and randomly placed $|\mathcal{S}_0| = 30$ true signals among those.

```
#Generate data
n <- 400
p <- 200
s_0 <- 30

amplitude <- 1
mu <- rep(0,p)
rho <- 0.25
Sigma <- toeplitz(rho^(0:(p-1)))

X <- MASS::mvrnorm(n, mu, Sigma)
nonzero <- sample(p, s_0)
beta <- amplitude * (1:p %in% nonzero)
y <- X %*% beta + rnorm(n)
```

Then, we run `run.uKO` with default settings, i.e. the decreasing sequence $q_k = q/2^{k-1}$ and $K = 5$.

```
res.uKO <- run.uKO(X, y, sets = TRUE)
res.uKO
$Shat
[1]   1   8   9  27  48  49  55  58  72  73  80  83  91  97 108
 115 119 121 126 129 139 142 144 146
[25] 147 151 155 158 163 165 167 171 174 178 181 183 186

$K
[1] 5

$FDRbound
[1] 0.3875

$sets
$sets$S1
[1]   1   8   9  27  48  49  55  58  72  73  80  83  91  97 108
 115 119 121 129 139 142 144 146 147
[25] 151 155 158 163 165 167 171 178 181 183 186

$sets$S2
[1]   8   9  27  48  49  55  58  72  73  91  97 108 115 119 121
126 129 139 142 144 146 147 151 155
[25] 158 163 167 171 174 181 186

$sets$S3
integer(0)

$sets$S4
```

```
integer(0)

$sets$S5
integer(0)
```

Since this example should only illustrate how we can apply the implemented functions and not examine the empirical FDP and power of those, we do not calculate any error measures here. For the investigation of the FDR and power of each method, we refer to the upcoming simulations. However, we briefly want to provide the intuition why three out of the five selection sets are empty in this example. Since we have used the decreasing sequence, the nominal levels of knockoff runs 3–5 are comparably small ($q_3 = 0.05,\ q_4 = 0.025,\ q_5 = 0.0125$). Those runs are very strict such that no variables are selected.

5.2 P-value knockoffs

5.2.1 Theory

Nguyen et al. (2020) construct an aggregation scheme for multiple knockoffs by translating the scores W_j into so-called intermediate p-values. They aggregate these p-values and apply either BH or BY, such that the final set of variables controls the FDR. Since it is the only aggregation procedure based on p-values, we will refer to the procedure as p-value knockoffs (pKO). The procedure starts by generating B knockoff matrices $\tilde{\mathbf{X}}^{(\mathbf{1})}, \ldots, \tilde{\mathbf{X}}^{(\mathbf{B})}$ independently, followed by the computation of the p score statistics $\mathbf{W}^{(\mathbf{b})} = (W_1^{(b)}, \ldots, W_p^{(b)})$ for each data pair $([\mathbf{X}\ \tilde{\mathbf{X}}^{(\mathbf{b})}], \mathbf{y}),\ b = 1, \ldots, B$.[2] The key step of pKO is the formulation of intermediate p-values based on the scores.

Definition 5.2 (**Intermediate p-values**). *Let* $(W_1^{(b)}, \ldots, W_p^{(b)})$ *be the vector containing* p *scores based on the data* $([\mathbf{X}\ \tilde{\mathbf{X}}^{(\mathbf{b})}], \mathbf{y})$. *Then, the intermediate empirical p-value* $\pi_j^{(b)}$ *is defined as*

$$\pi_j^{(b)} = \begin{cases} \dfrac{1 + \#\{l :\ W_l^{(b)} \leq -W_j^{(b)}\}}{p}, & W_j^{(b)} > 0 \\ 1, & W_j^{(b)} \leq 0. \end{cases} \tag{5.2}$$

For non-positive scores, we assign p-values equal to one because negative scores indicate that a knockoff is more important than its original counterpart. Remember that empirical p-values are generally defined as the number of cases where the test statistic is below/above the threshold of interest divided by the total number of cases. In the notion of the knockoff setting, we can use an analogue for positive scores. When we think of $W_j^{(b)}$ as threshold t, then we count the number of cases (here scores) that fall below $-t$, which is nothing else than the estimated number of false positives for that given threshold (see Section 3.5). We increment this quantity by one and divide it by the total number of cases (scores) to obtain the empirical p-value. The addition by one follows from the connection with vanilla knockoff+ as we will see below. Depending on which variable we look at, the threshold $W_j^{(b)}$ varies, which is essentially the same mechanism as the determination of T^+ that we have discussed in the vanilla knockoff+ framework (see Section 3.5). A more theoretical

[2] The number of knockoff matrices B has the same meaning as the number of knockoff runs K in union knockoffs and ADAGES. We use B for pKO to keep the notation close to the original one.

foundation of the (asymptotic) validity of intermediate p-values is given by Lemma 2 and A3 in Nguyen et al. (2020). In essence, the p-values (5.2) will be more accurate the larger the number of original (null) variables is since the empirical cumulative distribution function of the (null) scores, on which the computation of the p-values relies on, will be closer to the true one. A growing sample size n, on the other hand, does not affect their accuracy directly, but the estimation of the scores W_j becomes more precise, which can influence the p-values indirectly to some degree.
As in Meinshausen et al. (2009), we combine the empirical p-values $\pi_j^{(1)}, \dots, \pi_j^{(B)}$ for a certain variable j by quantile aggregation

$$\bar{\pi}_j = \min \left\{ \frac{q_\gamma(\{\pi_j^{(b)} : b = 1, \dots, B\})}{\gamma}, 1 \right\}, \quad \gamma \in (0, 1), \tag{5.3}$$

where $q_\gamma(\cdot)$ defines the empirical γ-quantile function. For an example with $\gamma = 0.5$, the p-values are aggregated to $\bar{\pi}_j$ by taking the sample median of $\{\pi_j^{(b)}\}_{b=1}^B$ and multiplied by a factor of two. Note that the aggregation step in (5.3) does not assume anything about the independence of intermediate p-values. Having obtained the set of aggregated p-values, we can either apply BH or BY to determine the final selection set $\hat{\mathcal{S}}_{pKO}$ obeying FDR control. Algorithm 3 summarizes the complete pKO procedure again.

Algorithm 3 P-value knockoffs

Input: Data $(\mathbf{X}, \mathbf{y})$; number of knockoff draws B; nominal level $q \in [0, 1]$
1 **for** $b \in \{1, \dots, B\}$ **do**
  Generate knockoff matrix $\tilde{\mathbf{X}}^{(\mathbf{b})}$.
  Compute $\mathbf{W}^{(\mathbf{b})}$ based on $([\mathbf{X}\ \tilde{\mathbf{X}}^{(\mathbf{b})}], \mathbf{y})$.
  Compute $\pi_1^{(b)}, \dots, \pi_p^{(b)}$ using (5.2).
end for
2 **for** $j \in \{1, \dots, p\}$ **do**
  Obtain $\bar{\pi}_j$ by aggregation of $\pi_j^{(1)}, \dots, \pi_j^{(B)}$ using (5.3).
end for
3 For ordered p-values $\bar{\pi}_{(1)} \leq \bar{\pi}_{(2)} \leq \dots \leq \bar{\pi}_{(p)}$, determine the threshold index j_0 by applying BH (2.4) or BY (2.5).
Output: Selection Set $\hat{\mathcal{S}}_{pKO} = \{j : \bar{\pi}_{(j)} \leq \bar{\pi}_{(j_0)}\}$

In the last step, where we use BH or BY, the same assumptions and intuitions hold as described in Sections 2.3.2–2.3.3. Although the BH procedure requires independence or the PRDS property of p-values, the authors recommend using BH. In their simulations, they observe that the aggregated p-values of the null features are almost independent, which they interpret as empirical justification to use BH. They believe that the aggregation step helps reducing dependencies between the final p-values. However, they only investigate the dependence for a model with a Toeplitz structure as a covariance matrix. The p-values can be far from independent for other covariance matrices, which should be kept in mind when applying pKO with BH. At first, it might seem counterintuitive to derive p-values and use BH/BY since we argued about the superiority of knockoffs compared to the classical FDR controlling procedures. But we emphasize that the pKO aggregation profits from the same advantages as MX knockoffs: the flexibility to control the FDR for almost any model class regardless of the dimensionality.
We proceed with the most important theoretical results of pKO and their intuitions. The

following proposition shows that vanilla knockoffs are a special case of pKO when applied with BH.

Proposition 5.3. *Define the pKO aggregation procedure as in Algorithm 3. Assume that all non-zero scores are distinct almost surely, that is*

$$\mathbb{P}(W_j = W_{j'},\ W_j \neq 0,\ W_{j'} \neq 0) = 0, \quad \forall j \neq j'.$$

Then, pKO with $\gamma = 1$, $B = 1$ and BH procedure as last step is equivalent to vanilla knockoff+.

Sketch of proof: First, with $\gamma = 1$ and $B = 1$, the aggregation step is trivial, and we end up with only one intermediate p-value π_j for each variable. Looking closely, we see that the count in (5.2) is a decreasing function $f(x) = \#\{l : W_l \leq -x\}$ in x. Sorting the intermediate p-values in ascending order is the same as ordering the scores W_j descendingly, that is $W_{(1)} \geq W_{(2)} \geq \ldots \geq W_{(p)}$. Plugging the definition of intermediate p-values in the BH step (2.4) in, but with the "ordered scores notation", we end up with

$$j_0 = \max\left\{j : \frac{1 + \#\{l :\ W_{(l)} \leq -W_{(j)}\}}{p} \leq \frac{j}{p} q\right\}.$$

We can exclude all ordered intermediate p-values $\pi_{(j)} = 1$ since from $q \in (0, 1)$, it follows $jq/p < 1$. Using the uniqueness of non-zero scores as stated in Proposition 5.3 together with the descending ordering assumption of the scores, the number of scores that are at least as large as the j-th score equals $\#\{l : W_{(l)} \geq W_{(j)}\} = j$.[3] Hence, the fraction simplifies to

$$\begin{aligned} j_0 &= \max\left\{j : \frac{1 + \#\{l :\ W_{(l)} \leq -W_{(j)}\}}{\#\{l : W_{(l)} \geq W_{(j)}\}} \leq q\right\} \\ &= \min\left\{W_{(j)} > 0 : \frac{1 + \#\{l :\ W_{(l)} \leq -W_{(j)}\}}{\#\{l : W_{(l)} \geq W_{(j)}\}} \leq q\right\}, \end{aligned}$$

which holds because looking for the *maximum index* j_0 in the descending sequence is the same as finding the *minimum value* of that sequence. Furthermore, due to the exclusion of all p-values equal to one, we restrict ourselves to positive scores. Without loss of generality, we can write the expression in a slightly different notation

$$T^+ = \min\left\{t > 0 : \frac{1 + \#\{l :\ W_l \leq -t\}}{\#\{l : W_l \geq t\}} \leq q\right\} = \min\left\{t > 0 : \frac{1 + |\mathcal{S}^-(t)|}{|\mathcal{S}^+(t)| \vee 1} \leq q\right\},$$

which is nothing else than the adaptive threshold of the knockoff+ procedure (3.12). □

The sketch of proof shows an important implication that was not clearly stated by Nguyen et al. (2020): If we remove the "+1" in the numerator of (5.2) and follow the same sketch of proof, we end up that pKO with BH equal vanilla knockoffs (3.9).
The derivation of the FDR control of pKO requires an additional assumption on the distribution of the null scores.

Proposition 5.4. *For a fixed b, the score statistics of the null variables $\{W_j^{(b)}\}_{j \in \mathcal{H}_0}$ are independent and follow the same distribution $\mathcal{P}_0$. By the coin-flipping property (Lemma 3.4), $\mathcal{P}_0$ is also symmetric around zero.*

[3] By non-zero scores, we mean $W_j \neq 0$, whereas by null scores we refer to scores of true null variables $\{W_j\}_{j \in \mathcal{H}_0}$.

Proposition 5.4 places a stronger assumption on the null scores than in the vanilla knockoff framework, which only requires the symmetry of the null scores around zero. Finally, we can state the FDR control of pKO by the following theorem.

Theorem 5.5 (**pKO FDR control**). *If Proposition 5.4 holds and* $|\mathcal{H}_0| \geq 2$*, then, for any arbitrary number of knockoff draws* B*, the pKO procedure described in Algorithm 3 controls the FDR for any* $q \in (0,1)$ *in a non-asymptotic regime at*

$$\mathbb{E}\left[\frac{|\hat{\mathcal{S}}_{pKO} \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}_{pKO}| \vee 1}\right] \leq \kappa q,$$

where $\kappa \leq 3.24$ *is a constant introduced in their proof.*
If we consider an asymptotic regime where $p \to \infty$ *and the number of true signals* $|\mathcal{S}_0|$ *is fixed such that* $|\mathcal{H}_0| \to \infty$*, pKO achieve FDR control at*

$$\lim_{p\to\infty} \mathbb{E}\left[\frac{|\hat{\mathcal{S}}_{pKO} \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}_{pKO}| \vee 1}\right] \leq q.$$

Equivalently, we can achieve the same bounds for the modified FDR control by the removal of "+1" in the intermediate p-value computation.

Note that we lose the factor κ on the FDR control which seems notably large at first. Although we could replace q by q/κ in the BH/BY step, Nguyen et al. (2020) observe that pKO control the FDR at q even without the adjustment in all their simulations. In fact, pKO often result in strikingly small empirical FDR values, which is kind of contradictory to the non-asymptotic theoretical FDR bound of Theorem 5.5. We will postpone the discussion of this phenomenon to the end of the section and the subsequent simulation part. Hence, they recommend proceeding without the correction factor. With an increasing number of (null) variables, pKO achieve asymptotic FDR control at q since the intermediate p-values become more accurate. We will not go into the proof due to its complexity but provide some comments instead.[4] The proof does not rely on the same strategy as the one for the FDR control of FX/MX knockoffs that we have presented in Section 3.7. The coin-flipping property only holds for a fixed b, i.e. for $\mathbf{W}^{(\mathbf{b})}$ (as used in Definition 5.4). However, conditionally on *all* $p \cdot B$ scores $\mathbf{W}^{(\mathbf{1})}, \ldots, \mathbf{W}^{(\mathbf{B})}$ simultaneously, Nguyen et al. (2020) cannot derive the coin-flipping property anymore when $B > 1$. Hence, their proof follows different mathematical arguments similar to those taken in the proof of Theorem 3.3 in Meinshausen et al. (2009).
The pKO aggregation requires the user to choose two parameters: The number of knockoff matrices B and the quantile value $\gamma \in (0,1)$. The authors showed in simulations that there is no significant improvement in power for $B \geq 25$. Furthermore, when $B \geq 25$, the power is roughly the same for all values of $\gamma > 0.1$. Hence, they recommend the parameter combination $(\gamma, B) = (0.3, 25)$. If the user wants to automatize and optimize the choice of γ, we recommend the adaptive search by Meinshausen et al. (2009) that selects the quantile value based on the data. Let us write the aggregated p-values (5.3) as a function of γ, i.e. $\bar{\pi}_j(\gamma)$, and let $\gamma_{\min}$ be the lowest γ in the search, then we define the adaptive search by

$$\bar{\pi}_j^* = \min\left\{(1 - \log\gamma_{\min}) \inf_{\gamma\in(\gamma_{\min},1)} \bar{\pi}_j(\gamma), 1\right\},$$

[4] See Appendix A.4–5 in Nguyen et al. (2020) for more details.

where the correction factor $1 - \log \gamma_{\min}$ can be seen as the price that we pay for optimizing over γ. Meinshausen et al. recommend the lower bound $\gamma_{\min} = 0.05$.
Turning to the authors' simulation findings, they show that pKO result in less variable FDP and power values than vanilla knockoffs over the iterations. Hence, pKO are successful in stabilizing FDR and power outcomes. Furthermore, for models with high correlations or large signal-to-noise ratios (SNR), pKO perform slightly better in terms of power, though for most parameter settings, pKO and vanilla knockoffs achieve almost equal power. For non-sparser models (larger $|\mathcal{S}_0|$), vanilla knockoffs lead to more correct detections. One major concern that we have found is that the authors only investigate a simulation design with a relatively large number of variables ($n = 500$ and $p = 1000$), where we expect that the asymptotic regime of Theorem 5.5 starts to kick in at some point. They do not study models with a smaller number of variables, e.g. $p = 100 - 500$, where only the finite sample FDR bound holds true. There, we can also expect significantly lower power because the intermediate p-values will be less accurate for smaller p. It is not clear if pKO provide still similar power values as vanilla knockoffs for smaller models, and we will investigate such scenarios in our final comparison.
Before we finish the discussion about pKO, we want to elaborate on two issues. First, since the authors observe an empirical FDR control at q without the adjustment by κ, the FDR bound in Theorem 5.5 might not be tight. This could be due to the independence assumption of the null scores $\{W_j\}_{j \in \mathcal{H}_0}$, which is only needed to apply Bernstein's inequality in their non-asymptotic FDR control proof (this is also where the constant κ comes from). There are several dependent extensions of Bernstein's inequality (Merlevède et al. 2009; Hang and Steinwart 2017), and the authors believe that Proposition 5.4 could be relaxed into some mixing condition. It even turns out that the empirical FDR values are often more conservative than expected. The authors speculate that the conservativeness comes from the quantile aggregation step (5.3). The somewhat contrasting behaviour of the empirical FDR values and the theoretical FDR bound indicate that there might be some room for improvement on both parts. Second, pKO with BH might not yield FDR control if the aggregated p-values are not independent. Although the authors show that their p-values have correlations close to zero, the independence does not hold in general for all data settings. This should be kept in mind in practice and possibly empirically checked before applying BH. Otherwise, the user can apply the conservative BY procedure as the last step in Algorithm 3 to ensure theoretical FDR control.[5] The mathematical concerns and inaccuracies can be seen as a price of running pKO as an aggregation scheme to reduce the instability of vanilla knockoffs.

5.2.2 Implementation

Similar to the implementation discussed for uKO, we briefly illustrate how the user can perform pKO with our `multiknockoffs` package. We can carry out the whole knockoff construction and aggregation by pKO with `run.pKO`.[6]

```
run.pKO(X, y, knockoffs = create.second_order,
        statistic = stat.glmnet_coefdiff, q = 0.2, B = B,
        gamma = 0.3, offset = 1, method = "BH", pvals = FALSE)
```

[5] In practice, this is often not necessary due to the previously described conservativeness of the method itself.
[6] An example where we run the pKO steps manually is again presented in Appendix B.2.

The user can adjust the following parameters:

- `X` the $n \times p$ design matrix and `y` and the $n \times 1$ response vector.
- `knockoffs` is the function for the knockoff construction. It must take the $n \times p$ data matrix as input and it must return a $n \times p$ knockoff matrix. The user can either choose a knockoff sampler of the `knockoff` package or define it manually. Default: `create.second_order`.
- `statistic` is a function that computes the scores W_j. It must take the data matrix, knockoff matrix and response vector as input and it outputs a vector of computed scores. The user can either choose one score statistic from the `knockoff` package or define it manually. Default: `stat.glmnet_coefdiff`.
- `q` defines the nominal level for the FDR control. Default: 0.2.
- `B` is the number of knockoff runs. Default: 25.
- `gamma` is a value between $(0, 1)$ which defines the quantile value used for the aggregation. If `gamma = NULL`, the adaptive search by Meinshausen et al. (2009) is used. Default: 0.3.
- `offset` either 0 or 1. Determines if an additional "+1" is added in the numerator of (5.2). Default: 1.
- `method` is the FDR controlling method in the last step. Either `"BH"` (default) or `"BY"`.
- `pvals` if the aggregated p-values should be reported (logical). Default: `FALSE`.

If the function is executed with all default values, the pKO filter is performed with the authors' reccomended parameters. It returns the aggregated selection set $\hat{\mathcal{S}}_{pKO}$, the number of knockoff matrices B and (if specified) the vector of aggregated p-values. With the same data example as in Section 5.1.2, we illustrate `run.pKO` with $B = 10$ knockoff draws.

```
res.pKO <- run.pKO(X, y, B=10, pvals = TRUE)
res.pKO
$Shat
[1]   8   9  27  48  49  55  58  72  73  91  97 108 115 119 121
 129 139 142 144 146 147 151 155 158
[25] 167 171 174 181 186

$B
[1] 10

$pvals
[1] 0.12833333 1.00000000 1.00000000 0.41666667 1.00000000
1.00000000 1.00000000 0.01666667
[9] 0.01666667 0.37833333 1.00000000 1.00000000 1.00000000
0.44500000 1.00000000 0.11166667
[17] 1.00000000 1.00000000 1.00000000 0.40666667 1.00000000
1.00000000 1.00000000 1.00000000
...
```

Note that we have only displayed the first 24 aggregated p-values (and not all 300) due to space.

5.3 ADAGES

5.3.1 Theory

Gui (2020) developed a procedure called adaptive aggregation with stability for distributed feature selection (ADAGES) which is a generic aggregation scheme and can be used with any FDR controlling procedure. He applies his algorithm in the context of distributed learning, where we can think of K machines that run variable selection algorithms individually. The main assumption here is that the data set is distributed over all K machines, resulting in K independent selection sets. A representative example would be collaborative research among K hospitals that have their own data due to patients' privacy. However, the FDR control of ADAGES does not rely on the independence assumption, and we can apply ADAGES to aggregate multiple knockoff runs, each estimated on data $([\mathbf{X}\ \tilde{\mathbf{X}}^{(\mathbf{k})}], \mathbf{y})$. To our knowledge, we are the first who run ADAGES on the whole data set without splitting it into K independent subsamples. The only difference between each run comes from the varying knockoff matrix $\tilde{\mathbf{X}}^{(\mathbf{k})}$. In the following, we explain ADAGES with knockoff filters, but as mentioned, any variable selection algorithm can be used.
Assume we run $k = 1, \dots, K$ knockoff+ filter runs resulting in K selection sets $\{\hat{\mathcal{S}}_k : k = 1, \dots, K\}$, each with error $\text{FDR}_k = \mathbb{E}\left[\frac{|\hat{\mathcal{S}}_k \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}_k|}\right]$. Moreover, let

$$m_j = \sum_{k=1}^{K} \mathbb{1}_{\{j \in \hat{\mathcal{S}}_k\}}, \quad j = 1, \dots, p$$

be the count of how often a variable j has been selected over all K knockoff runs. We can define a threshold-based selection rule for the aggregation as

$$\hat{\mathcal{S}}_{(c)} = \{j : m_j \geq c\}, \tag{5.4}$$

with integer $c \in \mathbb{Z}^+$. The following proposition states that the common set operations, unions and intersections, are special cases of the threshold-based aggregation rule (5.4).

Proposition 5.6. *If we choose the threshold-based rule $c = 1$, it equals the union of sets $\hat{\mathcal{S}}_U = \bigcup_{k=1}^{K} \hat{\mathcal{S}}_k$, i.e. $\hat{\mathcal{S}}_U = \hat{\mathcal{S}}_{(c=1)}$. If we choose the threshold-based rule $c = K$, it equals the intersection of sets $\hat{\mathcal{S}}_I = \bigcap_{k=1}^{K} \hat{\mathcal{S}}_k$, i.e. $\hat{\mathcal{S}}_I = \hat{\mathcal{S}}_{(c=K)}$.*

The definition of the aggregated set $\hat{\mathcal{S}}_{(c)}$ implies that its cardinality is a decreasing function of c, since $\hat{\mathcal{S}}_{(c_1)} \subseteq \hat{\mathcal{S}}_{(c_2)}$ for any $c_1 \geq c_2$. Assuming each knockoff run has FDR control at q, the union operation ($c = 1$) will lead to the largest set with the highest power but also with an overall FDR much larger than q. The intersection rule, on the other hand, is very strict with the lowest power, resulting in the smallest set. The choice of c controls the model's complexity and balances the trade-off between FDR and power. A priori, it is unknown to the user which threshold $c \in \{1, \dots, K\}$ is best for her data set, and there is no universal one. The goal of ADAGES is to determine an adaptive threshold c that is optimal for the given data. Hence, the union knockoff filter by Xie and Lederer (2021) is a special case of ADAGES with $c = 1$ but $\text{FDR}_k \leq q_k$ such that $q = \sum_{k=1}^{K} = q_k$.
ADAGES starts by choosing a candidate region for the threshold c by restricting the cardinality of the aggregated set $\hat{\mathcal{S}}_{(c)}$ to be at least as large as the mean cardinality of the K selection sets

$$c_0 = \max\left\{c : |\hat{\mathcal{S}}_{(c)}| \geq \bar{s}\right\}, \quad c_0 \geq 1. \tag{5.5}$$

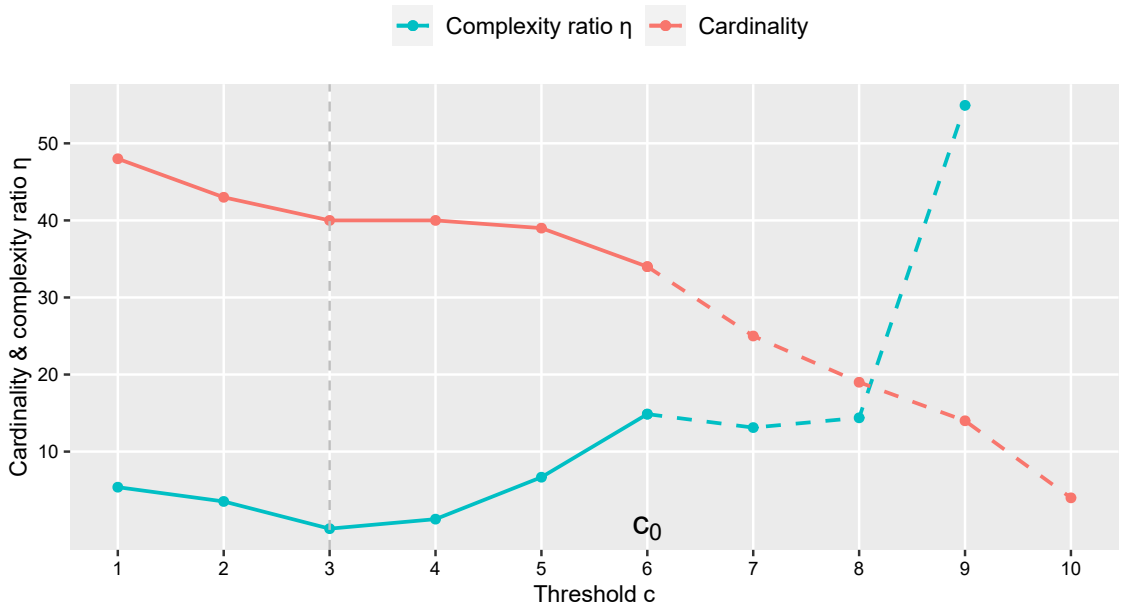
On the one hand, the choice of c_0 is a theoretical requirement for the proof of the FDR control of ADAGES. On the other hand, c_0 intuitively ensures that the aggregated set is not too small to maintain a certain level of power. Hence, ADAGES considers integers between $c \in [1, c_0]$ as thresholds, and we are interested in finding one that balances the tradeoff between FDR and power well. Gui (2020) introduces the complexity ratio as a criterion to choose the adaptive threshold

$$\eta_c = \begin{cases} \dfrac{|\hat{\mathcal{S}}_{(c)}|}{|\hat{\mathcal{S}}_{(c+1)}|}, & |\hat{\mathcal{S}}_{(c+1)}| > 0 \\ \infty, & |\hat{\mathcal{S}}_{(c+1)}| = 0, \end{cases}$$

which characterizes the stability of the aggregated sets for a decreasing sequence of thresholds starting from c_0. The smaller the fraction of cardinalities gets, the smaller the change in the empirical FDP and power. Through a personal exchange with the author, we have found out that the idea of the complexity ratio comes from the eigenvalue ratio test for identifying the number of factors in factor models (Ahn and Horenstein 2013). Finally, the adaptive threshold c^* is that threshold within the candidate region that results in the minimal complexity ratio

$$c^* = \arg\min\{\eta_c : 1 \leq c \leq c_0,\ |\hat{\mathcal{S}}_{(c+1)}| > 0\}, \tag{5.6}$$

which can be interpreted as a sign of a stable selection. We plug c^* into the threshold-based selection rule (5.4) to obtain the final selection set. Figure 5.1 visualizes the determination of the adaptive threshold with the minimization of the complexity ratio as the criterion. At the optimal threshold $c^* = 3$, the change of the two consecutive selection sets $|\hat{\mathcal{S}}_{(3)}|$ and $|\hat{\mathcal{S}}_{(4)}|$ is the smallest, and so the change in FDP and power.



The complexity ratio is scaled by $50 \cdot \log(\eta_c)$ for visualization purposes. Cardinality refers to $|\hat{\mathcal{S}}_{(c)}|$. All dashed points correspond to thresholds $c > c_0$. Model: $n = 500, p = 300, |\mathcal{S}_0| = 30$. ADAGES aggregates $K = 10$ knockoff filters with $2p = 600$ variables in total, each run based on the LCD score.

Figure 5.1: ADAGES: Determination optimal threshold

Algorithm 4 ADAGES knockoffs

Input: Data $(\mathbf{X}, \mathbf{y})$; number of runs K; nominal level $q \in [0, 1]$

1 **for** $k \in \{1, \dots, K\}$ **do**
  Generate knockoff matrix $\tilde{\mathbf{X}}^{(\mathbf{k})}$.
  Apply knockoff/knockoff+ on $([\mathbf{X}\ \tilde{\mathbf{X}}^{(\mathbf{k})}], \mathbf{y})$ and obtain selection set $\hat{\mathcal{S}}_k$.
end for
Result: K different selection sets $\hat{\mathcal{S}}_1, \dots, \hat{\mathcal{S}}_K$.

2 Calculate $m_j = \sum_{k=1}^{K} \mathbb{1}_{\{j \in \hat{\mathcal{S}}_k\}}, \quad \forall j = 1, \dots, p.$

3 **for** $c \in \{1, \dots, K\}$ **do**
  Obtain $\hat{\mathcal{S}}_{(c)} = \{j : m_j \geq c\}$.
  Calculate complexity ratio $\eta_c = \frac{|\hat{\mathcal{S}}_{(c)}|}{|\hat{\mathcal{S}}_{(c+1)}|}$ if $|\hat{\mathcal{S}}_{(c+1)}| > 0$; otherwise $\eta_c = \infty$.
end for

4 Calculate $\bar{s} = \frac{1}{K} \sum_{k=1}^{K} |\hat{\mathcal{S}}_k|$.

5 Find $c_0 = \max \left\{ c : |\hat{\mathcal{S}}_{(c)}| \geq \bar{s} \right\}$.

6 Determine adaptive threshold $c^* = \arg\min\{\eta_c : 1 \leq c \leq c_0\}$.

Output: Aggregated selection set $\hat{\mathcal{S}} = \hat{\mathcal{S}}_{(c^*)} = \{j : m_j \geq c^*\}$

Algorithm 4 summarizes all steps of ADAGES with the complexity ratio criterion again. Other criterias for the choice of c are also possible as long as FDR control is theoretically guaranteed. Gui (2020) suggests another criterion by minimizing the trade-off between the threshold and the model complexity

$$\tilde{c} = \underset{1 \leq c \leq c_0}{\arg\min}\, c|\hat{\mathcal{S}}_{(c)}|, \tag{5.7}$$

and we refer to this approach as modified ADAGES. The idea of minimizing that quantity comes from its occurrence in the lower bound of the proportion of true positives. We will not discuss this criterion further since neither Gui nor we observe that modified ADAGES performs better than ADAGES. There is clearly room for the development and investigation of new criteria for c, and we leave this question open for further research.
Turning to the theoretical FDR control, the following theorem states the FDR bound of ADAGES formally.

Theorem 5.7 (**ADAGES FDR control**). *For any nominal level $q \in (0, 1)$, assume that each knockoff run ensures $\mathrm{FDR}_k \leq q$, for $k = 1, \dots, K$. Then, ADAGES with $c^* \in [1, c_0] \cap \mathbb{Z}^+$ controls the FDR at*

$$\mathbb{E}\left[\frac{|\hat{\mathcal{S}}_{(c^*)} \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}_{(c^*)}| \vee 1}\right] \leq \lambda q,$$

where $\lambda \geq \max_{1 \leq k' \leq K} \frac{|\hat{\mathcal{S}}'_k|}{c^*} \sum_{k=1}^{K} \frac{1}{|\hat{\mathcal{S}}_k|}$ is a theoretical variable from the related proof.

Theorem 5.7 does not only hold for multiple knockoffs but for general K variable selection procedures that achieve $\mathrm{FDR}_k \leq q$ respectively. Due to the generality of ADAGES, the proof does not rely on any strategies that we have presented in the knockoff sketch of proof. We refer to Appendix A.2.4 for the proof of Theorem 5.7. Similar to pKO in finite samples and uKO with the recommended sequence, ADAGES has a theoretical FDR bound usually larger than q but may empirically achieve FDR values below q.

Turning to his simulation results, Gui (2020) compares ADAGES with uKO by Xie and Lederer (2021) and pKO by Nguyen et al. (2020). However, he studies a distributed learning environment, i.e. he splits the sample n into K independent subsamples and applies a knockoff filter on each subsample n_k. He shows that ADAGES' power outperforms the other two aggregation methods while still controlling the empirical FDR at q in simulations with a varying number of runs K and variables p. However, we have some concerns about the representativeness of the simulation design: First, Gui simulates models with many observations $n = 1000$ but only a few variables $p = 15$–90. This probably harms the performance of pKO because the intermediate p-values will be less accurate for a small p. Second, he applies pKO with the conservative BY instead of BH, which also reduces their power considerably. Moreover, Gui does not adopt Xie and Lederer's recommended sequence for q_k but the average sequence $q_k = q/K$, which is usually very conservative in practice, in particular for larger K. Last but not least, since Gui works in the distributed learning environment and deals with a different research question, he does not compare ADAGES with vanilla knockoffs.

5.3.2 Implementation

The multiple knockoff construction and the aggregation with ADAGES can be carried out at once by

```
run.ADAGES(X, y, knockoffs = create.second_order,
           statistic = stat.glmnet_coefdiff, q = 0.2, K = 5,
           offset = 1, type = "ADAGES", sets = FALSE)
```

where the user can customize the following settings:

- `X` the $n \times p$ design matrix and `y` and the $n \times 1$ response vector.
- `knockoffs` is the function for the knockoff construction. It must take the $n \times p$ data matrix as input and it must return a $n \times p$ knockoff matrix. The user can either choose a knockoff sampler of the `knockoff` package or define it manually. Default: `create.second_order`.
- `statistic` is a function that computes the scores W_j. It must take the data matrix, knockoff matrix and response vector as input and it outputs a vector of computed scores. The user can either choose one score statistic from the `knockoff` package or define it manually. Default: `stat.glmnet_coefdiff`.
- `q` defines the nominal level for the FDR control. Default: 0.2.
- `K` is the number of knockoff runs. Default: 5.
- `offset` either 0 (knockoff) or 1 (knockoff+). Default: 1.
- `type` either `ADAGES` (default) or `ADAGES.mod` (see below).
- `sets` logical argument if the K selection sets of each knockoff run before the aggregation should be returned. Default: `FALSE`.

The user can choose between the two criteria for the determination of c by modifiying the argument `type`. The default option `ADAGES` refers to the ADAGES procedure as summarized in detail in the previous section with the complexity ratio criterion (5.6), where `ADAGES.mod` applies the threshold-complexity trade-off criterion (5.7). Although we have not discussed the latter version in great detail, we have implemented the criterion for the sake of completeness. The function `run.ADAGES` returns the aggregated selection set $\hat{\mathcal{S}}_{(c^*)}$, the optimal threshold c^*, the number of knockoff runs K and (if specified) the individual selection sets of each knockoff run. We continue with code showing the execution

of ADAGES with default values on our exemplary data and its output.

```
res.ADAGES <- run.ADAGES(X, y, sets = TRUE)
> res.ADAGES
$Shat
 [1]   8   9  16  27  48  49  55  58  72  73  80  91  97 108 115
 119 121 126 129 139 142 144 146 147
[25] 151 155 158 163 165 167 171 181 186

$c
[1] 2

$K
[1] 5

$sets
$sets$S1
[1]   8   9  27  48  49  55  58  72  73  91  97 108 115 119 121
 126 129 139 142 144 146 147 151 155
[25] 158 167 171 181 186

$sets$S2
[1]   8   9  16  27  48  49  55  58  72  73  80  83  91  97 108
115 119 121 129 133 139 142 144 146
[25] 147 151 155 158 160 163 165 167 171 181 186

$sets$S3
[1]   8   9  27  48  49  55  58  72  73  80  91  97 108 115 119
 121 129 139 142 144 146 147 151 155
[25] 158 165 167 169 171 181 186

$sets$S4
[1]   8   9  16  27  48  49  55  58  72  73  91  97 108 115 119
 121 126 129 139 142 143 144 146 147
[25] 151 155 158 163 165 167 171 181 186

$sets$S5
[1]   8   9  27  48  49  55  58  72  73  91  97 108 115 119 121
 129 139 142 144 146 147 151 155 158
[25] 163 167 171 181 186
```

In this example, the optimal threshold is $c^* = 2$. Hence, the aggregated selection set is not the union but it consists of all variables that occur at least two times across the individual selection sets $\hat{\mathcal{S}}_1, \ldots, \hat{\mathcal{S}}_K$.

5.4 Comparison of the multiple knockoff methods

5.4.1 A qualitative comparison and pre-discussion

Before we conduct the simulation, we start with a qualitative comparison of the three multiple knockoff methods and discuss their advantages and shortcomings (Table 5.1). All arguments given in this subsection are completely based on the theory and observations of the original works and do not include conclusions of our upcoming simulations.
Both uKO and ADAGES can be applied beyond the multiple knockoff framework because their theory only requires a variable selection method that achieves FDR control at q. On the other hand, pKO are not universally applicable since their construction is specifically tailored to the knockoff setting. All three methods rely on the choice of the number of knockoff runs as a hyperparameter. For uKO, the user has to specify an additional parameter group: the sequence of nominal levels $\{q_k\}_{k=1}^K$ which is far from trivial as we have discussed in Section 5.1, and clearly one shortfall of uKO. Moreover, since all three procedures are aggregation methods, we can expect that they reduce the variability of the FDP and power compared to vanilla knockoffs. Out of the three, only uKO theoretically retain finite sample FDR control at q after the aggregation if $\{q_k\}_{k=1}^K = q$. ADAGES and pKO lead to a finite sample FDR control proportional to q. However, all three methods empirically achieve FDR control at q in their original papers, even without the theoretical guarantee for finite samples. Even Xie and Lederer's (2021) recommended sequence with a theoretical bound much higher than q shows empirical FDR values below q. Hence, although a thereotical FDR bound of at most q would be a desirable property of multiple knockoffs, it might not be necessary in practical applications. We will study the empirical FDR control in detail in the upcoming simulations. An advantage of pKO is that they produce asymptotically valid p-values, and the method seems less sensitive for changes in the number of knockoff runs. However, pKO are known to perform worse if the number of null variables is very small. Gui (2020) shows that ADAGES outperforms the other two methods, but he only investigates ADAGES in a distributed learning environment and for very few settings. Furthermore, he neither chooses the recommended settings for uKO nor for pKO, which makes the comparison less representative.
From a computational view, all three aggregation schemes take approximately equally long for the same K (B). The computationally intensive part is the generation of the K (B) knockoff matrices and the estimation of the variable selection method (e.g. Lasso) on the data $([\mathbf{X}\ \tilde{\mathbf{X}}^{(\mathbf{k})}], \mathbf{y})$ for all $k = 1, \ldots, K$. The computational time of the aggregation of the knockoff runs itself is negligible for all three schemes. However, each aggregation method might require a different number of knockoff matrices for optimal results. For example, Xie and Lederer (2021) recommend $K \approx 5$, Nguyen et al. (2020) suggest $B = 25$ and Gui (2020) provides no recommendation but $K \approx 5$ leads to large and stable power values in his simulations. Hence, considering the recommended or optimal hyperparameters of the authors, uKO and ADAGES will take much less time than pKO in practice.

Table 5.1: Qualitative comparison multiple knockoffs

	uKO	pKO	ADAGES
Applications	General	Knockoffs	General
Hyperparameter	Knockoff runs K Sequence $\{q_k\}_{k=1}^K$	Knockoff runs B Optionally γ	Knockoff runs K
Improved stability	Yes	Yes	Yes
FDR control at q	Yes if $\sum_{k=1}^K q_k = q$	Not for finite samples	No
Advantages	Easy to understand	Asymptotically valid p-values Seems less sensitive for large B	Performs best in Gui's (2020) work
Shortcomings	Power can decrease with K Can be very strict Choice of $\{q_k\}_{k=1}^K$ unclear	Can perform poor for small p Only slighty better than MX knockoffs	Performance not well studied Not studied in multiple knockoff setting

5.4.2 Simulation

In this section, we want to compare the presented aggregation methods by simulations. As mentioned in the previous parts, we have some concerns about the simulations conducted in the original papers of each method. Xie and Lederer (2021) treat $(\mathbf{X}, \mathbf{y})$ as fixed such that the randomness comes only from the knockoff matrix construction in their simulation, even though $\mathbf{X}$ is also a random variable in the MX knockoff setting. Across their investigations, they also change more than one parameter simultaneously, which causes more difficulties to understand the behaviour of uKO. Nguyen et al. (2020), on the other hand, generate only large models $(n, p) = (500, 1000)$ where the intermediate p-values are known to be more accurate. Gui (2020) examines a distributed learning environment where he splits the data into K subsamples and runs a knockoff filter on each one. Furthermore, he also draws only small models with $p = 15 - 90$. Although he is the only one who compares ADAGES with pKO and uKO, he does not use the authors' recommendations for pKO and uKO, and he does not cover many parameter settings. Both reasons do make his comparison less representative. A comparison with vanilla knockoffs is also missing in Gui's simulation.
The goal of our investigation is to find out i.) if the aggregation schemes are superior to vanilla knockoffs in terms of power while retaining empirical FDR control at q and ii.) if there is an aggregation procedure that is dominant across all settings. To our knowledge, we are the first, who conduct a fair and extensive comparison of the three methods in the multiple knockoffs setting and with "optimal" parameters for each method. We will generate our base model according to

$$\begin{aligned} \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \quad \Sigma_{j,k} = 0.4^{|j-k|} \quad \forall j \neq k \in \{1, \dots, p\}, \quad \sigma^2 = 1, \\ n = 500, \quad p = 300, \quad |\mathcal{S}_0| = 30, \quad |\mathcal{H}_0| = 270. \end{aligned} \tag{5.8}$$

The 30 non-zero coefficients are selected uniformly at random. Each signal coefficient is set to one before we rescale the coefficient vector to a model with a signal-to-noise ratio of $\|\mathbf{X}\beta\|_2^2 / n\sigma^2 = 3$. The response is drawn according to the linear model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$. For the knockoff construction, we use the default second-order approach discussed in Section 4.2.4, but with an equi-correlated construction instead of SDP since it has resulted in larger power.[7] Each knockoff run is based on the LCD score $W_j = |\hat{\beta}_j(\lambda_{\text{CV}})| - |\hat{\beta}_{j+p}(\lambda_{\text{CV}})|$, i.e. we fit a Lasso regression with data $([\mathbf{X}\, \tilde{\mathbf{X}}], \mathbf{y})$, tune λ by CV and take the coefficient differences of variable j and its knockoff. For the threshold determination, we use knockoff+ (3.12) to ensure theoretical FDR control at $q = 0.2$ of each knockoff run. Then, we determine the selection set of the following four methods:

i.) Vanilla MX knockoffs (KO).
ii.) uKO with $K = 5$ and $q_k = q/2^{k-1}$ for $k = 1, \dots, K$.
iii.) pKO with $(B, \gamma) = (25, 0.5)$.
iv.) ADAGES with $K = 5$.

Once we have obtained the selection sets, we compute the empirical FDP and power of each procedure and average them over $M = 500$ iterations. In each iteration, all aggregation methods utilize the same K knockoff matrices in their estimation, except of pKO which rely on 20 additional constructed ones.[8] Although Nguyen et al. (2020)

[7] The knockoff filter uses SDP knockoffs and not ASDP because $p \leq 500$. In the last part of Section B.2, we show how to modify the implemented functions to create equi-correlated knockoffs instead of solving the (A)SDP optimization.

[8] More precisely, in each iteration, we independently generate 25 knockoff matrices $\tilde{\mathbf{X}}^{(1)}, \dots, \tilde{\mathbf{X}}^{(25)}$ and use $\tilde{\mathbf{X}}^{(1)}$ for vanilla knockoffs, $\tilde{\mathbf{X}}^{(1)}, \dots, \tilde{\mathbf{X}}^{(5)}$ for uKO and ADAGES and $\tilde{\mathbf{X}}^{(1)}, \dots, \tilde{\mathbf{X}}^{(25)}$ for pKO.

recommend $(B, \gamma) = (25, 0.3)$, we observe that our values perform slightly better in our models. Furthermore, the adaptive search for γ proposed by Meinshausen et al. (2009) does not yield a clear improvement over a fixed $\gamma = 0.5$, and so we will stick to that value. For uKO and ADAGES, we choose $K = 5$ because it is the recommendation of Xie and Lederer (2021), and also Gui (2020) uses $K = 5$ in his work. In our opinion, it would be unfair to run each aggregation method on the same number of knockoff runs as it was done by Gui because not all methods could realise their full potential. While the power of pKO increases with B, this is not the case for uKO and ADAGES, as we will see later. Hence, the aggregation methods can have different optimal values for K or B, and we choose to run each method on its own optimal or recommended number of knockoff runs.[9]
Based on the above model (5.8), we vary one of the following parameters in each simulation setting:

i.) Correlation strength $\rho \in \{0.2, 0.3, \dots, 0.7, 0.8\}$.
ii.) SNR $\in \{1, 2, 3, 4, 5\}$.
iii.) Sparsity level $|\mathcal{S}_0| \in \{10, 15, 20, 25, 30, 40, 50, 60, 70\}$.
iv.) Number of knockoff runs $K \in \{2, 5, 8, 10, 15, 20, 25, 40, 50\}$ (or B for pKO).

Figure 5.2 depicts the empirical FDR and power of the four methods over the $M = 500$ iterations for varying correlation strength ρ, SNR, and sparsity level $|\mathcal{S}_0|$ respectively.[10] We observe a similar pattern of the FDR and power curve in every setting. All three aggregation procedures and vanilla knockoffs control the FDR at q for most scenarios. ADAGES, however, does not yield FDR control at q for some "boundary values" of our simulations, i.e. for SNR $= 5$ and $|\mathcal{S}_0| = \{10, 15, 20\}$. Also, uKO slightly exceed the nominal value of $q = 0.2$ at $|\mathcal{S}_0| = 10$, whereas pKO always lie below the nominal value in each simulation setting. We want to remind the reader that none of the three aggregation methods guarantees theoretical FDR control at q but at νq, for $\nu > 1$. We can also observe a ranking with respect to the empirical FDR levels. While uKO result in similar FDR values as vanilla knockoffs, ADAGES is less conservative. The p-value aggregation method pKO, on the other hand, is the most conservative procedure with FDR values below 2.5%. This phenomenon has also been observed by Nguyen et al. (2020) and Gui (2020), which, in their opinion, stems from the conservative quantile aggregation method for the intermediate p-values.
Turning to the power, we first look at the power curve for varying correlation strengths ρ. While for small ρ values all three aggregation methods have approximately the same power, larger differences occur for stronger correlations. The power of uKO almost converges to the power of vanilla knockoffs, but it remains at a higher level for all ρ. ADAGES performs best in terms of power with considerable differences compared to the other methods for increasing ρ. The pKO filter performs better than vanilla knockoffs until $\rho = 0.4$ before the power drastically falls, reaching near-zero values for $\rho \geq 0.7$. This pattern of the power curves repeats for varying SNR and sparsity levels: ADAGES has the largest power followed by uKO, which perform better than vanilla knockoffs. PKO, on the other hand, are better than vanilla knockoffs for some parameter values, but they always have a turning point at which their power falls below the power of vanilla knockoffs. Remember that ADAGES does not achieve empirical FDR control at q in every setting. For SNR $= 5$, all aggregation methods have a power close to one, but only ADAGES exceeds the $q = 0.2$ bound. The same holds true for $|\mathcal{S}_0| = \{10, 15, 20\}$, where the power of each method is approximately one, but ADAGES does not achieve FDR control. In these scenarios, it

[9] We denote the number of knockoff runs with K for uKO and ADAGES and B for pKO.
[10] In each subfigure, we vary one parameter while the others are fixed at the values described in (5.8).

is better to choose uKO or pKO, but we cannot observe the underlying model's sparsity level or SNR in practice.

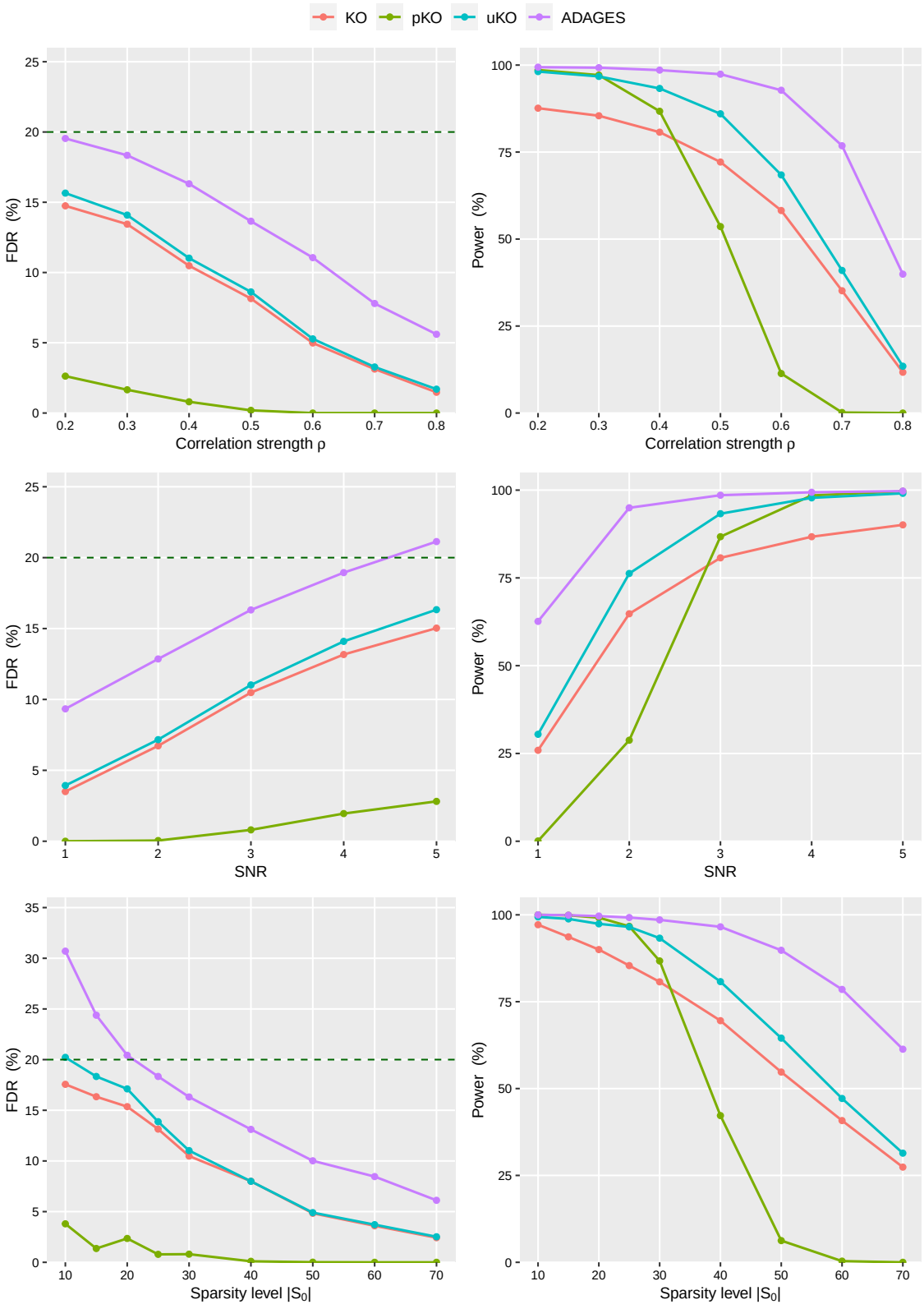


Figure 5.2: Multiple knockoffs: Simulation main results

To check the validity of our chosen number of knockoff runs K, we inspect the FDR and power for a varying K in Figure 5.3. Each aggregation method has an empirical FDR curve below $q = 0.2$ for all values of K. The FDR curve of ADAGES has an interesting pattern. It increases until $K = 5$ before it continuously decreases for larger K. Looking at the power of pKO, we can indeed observe that until $B = 25$, there is a notable increase, whereas it remains almost constant for larger $K = \{40, 50\}$. Hence, $B = 25$ is a valid choice for pKO. Although not clearly obvious from Figure 5.3, the FDR and power of uKO are constant for $K \geq 5$. This property comes from the construction of the decreasing

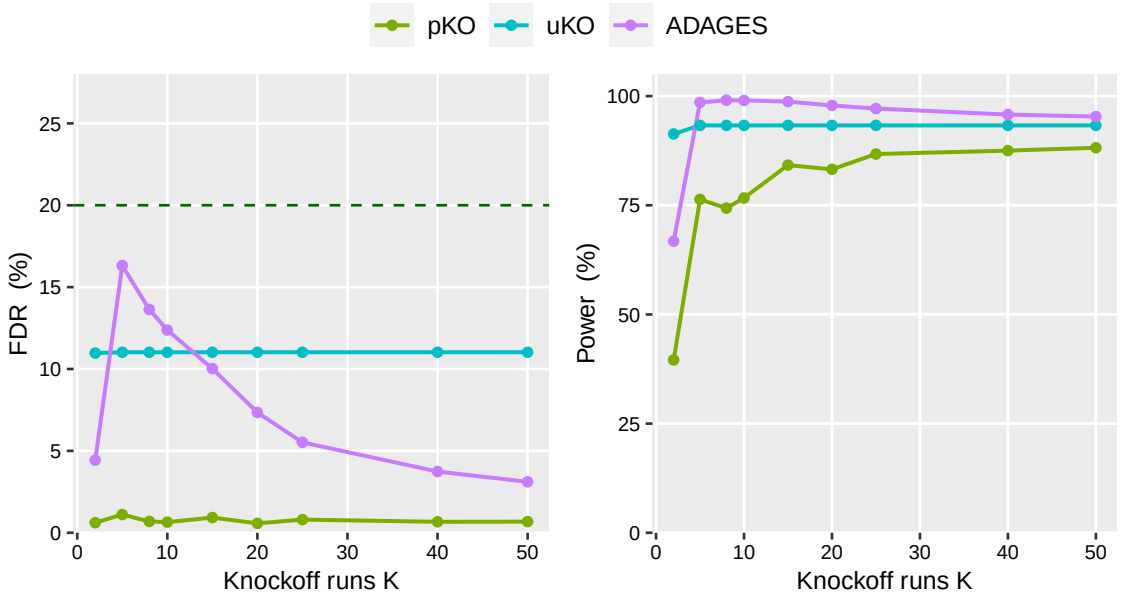


Figure 5.3: Multiple knockoffs: Simulation varying K.

sequence $q_k = q/2^{k-1}$. For larger k, the nominal levels q_k become very small, e.g. for $k = 10$ and $q = 0.2$, we have $q_{10} \approx 0.0004$. The corresponding knockoff run will be very strict with probably no selections or only those with an extraordinarily strong signal. And even in the unlikely case that the selection set is non-empty, the variables might be already in one of the previous selection sets for a larger q_k. Hence, for an increasing K, the FDR and power of uKO with decreasing sequence will not change since no new discoveries are made. ADAGES has a notable power increase for $K = 5$ and stagnates afterwards at around 98%, which justifies the default choice of $K = 5$. But if we take the FDR curve shape into account, we could also think of increasing the knockoff runs to $K = 10$–25 since the power approximately remains stable while the FDR is lower. This could help us achieve an even better FDR control for ADAGES at these "boundary cases" in Figure 5.2. Figure 5.4 illustrates the FDR and power of the main simulation for ADAGES with different K and vanilla knockoffs as a benchmark. Focusing on the boundary cases, which are SNR $= 5$ and $|\mathcal{S}_0| = \{10, 15, 20\}$, we can indeed observe a reduction of the empirical FDR with increasing K, while the power values of all ADAGES methods are roughly equal. In the simulation with varying sparsity levels, at $|\mathcal{S}_0| = 15$ for example, ADAGES$_{K=5}$ has an empirical FDR of 25.5%, whereas ADAGES$_{K=25}$ features an empirical FDR value of 20.3%. Both have approximately a power of one. At $|\mathcal{S}_0| = 30$, although ADAGES$_{K=5}$ and ADAGES$_{K=25}$ control the FDR at $q = 0.2$, the difference between their empirical FDR values even rises to more than ten percentage points, whereas their power values are almost the same. This behaviour of ADAGES implies that it can be better to choose a

larger K than five since the aggregation of more runs seems to make the procedure more conservative but does not significantly affect the power.

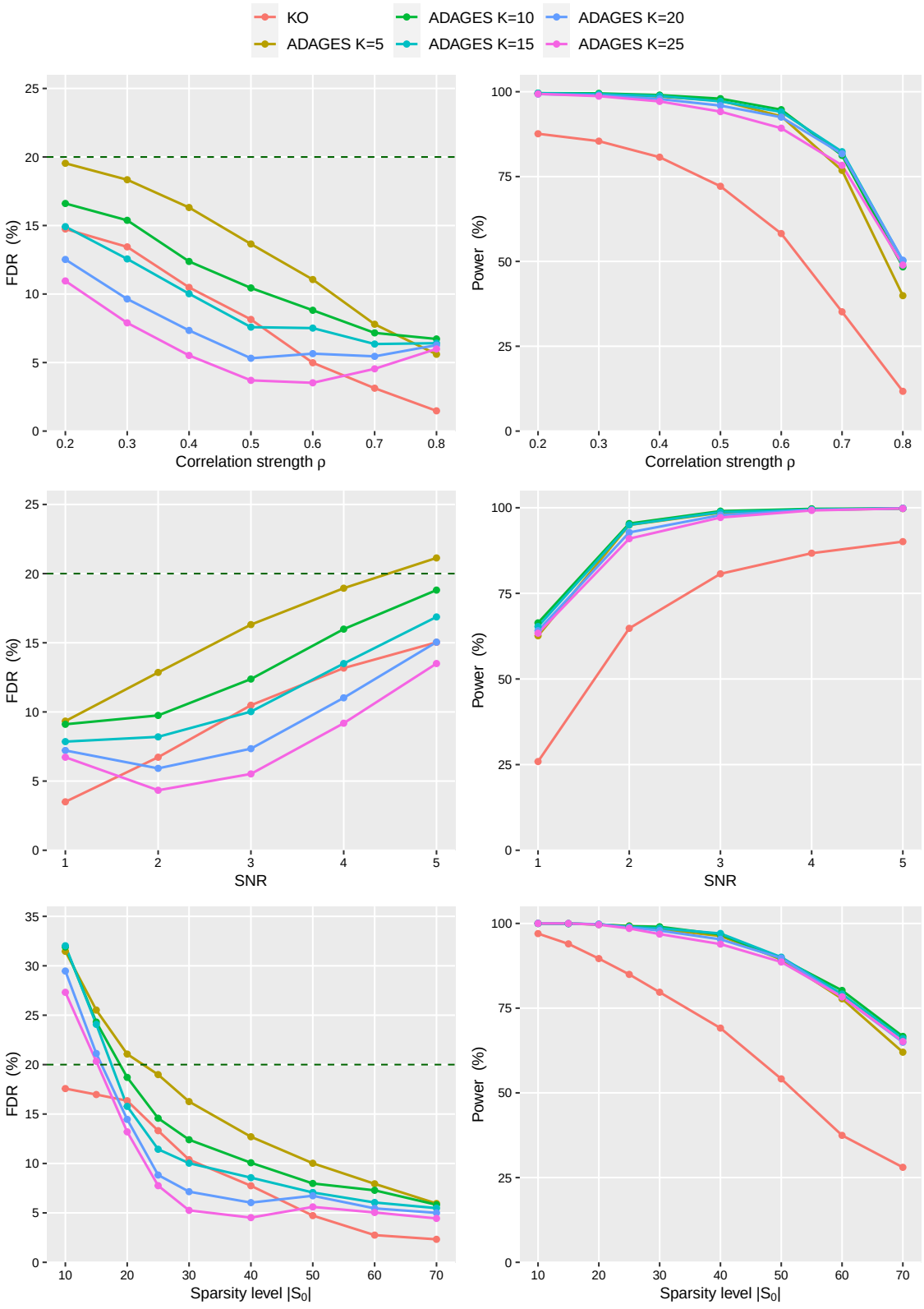


Figure 5.4: Multiple knockoffs: Simulation ADAGES varying K

As mentioned at the beginning of this subsection, we have generated equi-correlated knockoffs since they achieve larger power than SDP knockoffs in this particular experiment. From a personal exchange with Lucas Janson, one of the authors of the model-X knockoff paper (Candès et al. 2018), we figured out that it is often a question of the specific experimental design whether (A)SDP or equi-correlated knockoffs perform better, and there is no comprehensive theory for that so far.[11] Figures A.3 and A.4 depict the FDR and power for exactly the same simulation setting but with the SDP optimization. Besides the general differences in power, we observe that the pKO filter performs considerably different with an SDP construction.[12] In all investigated parameter settings, pKO were very conservative with an empirical FDR close to zero and power values that are notably lower than vanilla knockoffs. ADAGES and uKO are still better than vanilla knockoffs despite the general drop in power. Furthermore, also the pattern for the optimal number of knockoff runs changes for pKO. The power slightly fluctuates around 22%, with minor peaks at $K = \{5, 15, 25\}$. ADAGES and uKO have a similar pattern as for the equi-correlated construction. They experience power increases until $K = 5$, and only minor or no changes for larger K. We interpret the sensitivity of pKO regarding different knockoff constructions as one disadvantage of the method.

Before we close this section, we want to mention some additional conclusions that we have worked out during our simulations but are not shown here: i.) We have also varied the number of variables p, but the power values of all three aggregation methods and vanilla knockoffs quickly approach to one for increasing $p \geq 500$. Since we hold $|\mathcal{S}_0|$ fixed in this experiment, the added features are all null variables. This might be in favour of the underlying correlation structure, where variables that are far away will have a correlation close to zero for large p. ii.) If we use pKO with knockoff instead of knockoff+, i.e. using "+0" instead of "+1" in (5.2), it barely increases the empirical FDR but boosts the power. This observation is another sign of the sensitivity of the intermediate p-values (5.2) and pKO in general. However, it would be an unfair comparison to use pKO with knockoff but knockoff+ for the remaining aggregation methods. iii.) We have also performed uKO with theoretical FDR control by choosing the average sequence $q_k = q/K$, $\forall k$. However, this sequence leads to a very conservative aggregation scheme since each knockoff run is too strict with only a few and often no detections. This observation becomes even more striking for larger K, where each nominal level $q_k = q/K$ is smaller.

In summary, we have seen that multiple knockoffs can improve vanilla knockoffs in terms of power. Although all investigated aggregation methods do not guarantee theoretical FDR control at q, we can still observe an empirical FDR below that nominal level for the majority of our settings. The p-value knockoff aggregation pKO is very conservative in its empirical FDR control. Although pKO can lead to power increases compared to vanilla knockoffs, there is always a turning point in our simulations where the power falls below the one of vanilla knockoffs. In addition, pKO seem sensitive if we use different knockoff construction methods. The aggregation scheme uKO has similar empirical FDR values as vanilla knockoffs but a higher power across all our settings. Although uKO with decreasing sequence do not achieve FDR control in theory either, we observe empirical FDR control for all settings except when the sparsity level was $|\mathcal{S}_0| = 10$. Even there, the empirical

[11] Spector and Janson (2021) argue that minimizing the pairwise correlations $\text{Corr}(\mathbf{X}_j, \tilde{\mathbf{X}}_j)$ is not optimal to maximize the power. They show that minimizing "the ability to reconstruct a variable $\mathbf{X}_j$ by using the variables $\mathbf{X}_{-j}$ and knockoffs $\tilde{\mathbf{X}}$" leads to more powerful knockoffs. Since the optimal knockoff construction theory is fairly new, we stick to the initial approaches by Candès et al. (2018), ASDP and equi-correlated knockoffs, which are used by almost the entire knockoff literature.

[12] Looking at the associated code of Nguyen et al. (2020), they also use the second-order approach with an equi-correlated construction. However, they do not explain their choice further in their work.

FDR only slightly exceeds the 20% bound. Our findings support the conclusion from Xie and Lederer (2021), who use a different simulation strategy, that uKO are an improvement over vanilla knockoffs while retaining FDR control. Gui (2020) concludes that uKO perform very conservative, but he uses the average sequence instead of the recommended decreasing sequence for the nominal levels. We can confirm the conservativeness of the average sequence and recommend using the decreasing sequence. ADAGES performs best in terms of power across all our settings. This comes at a price that it (slightly) exceeds the nominal bound of 20% for extremely small sparsity levels or large SNR values. In those specific settings, all aggregation methods achieved a power of one, but uKO and pKO had FDR values below q, and are therefore preferable. As we have covered more parameter settings and have conducted a fairer comparison than Gui (2020), we can confirm that ADAGES leads indeed to the largest power increases among the three aggregation methods. However, in contrast to Gui, we have also found some settings where ADAGES does not have empirical FDR control anymore.
We finish with our recommendation for the use of the aggregation methods in practice:

- If the user wants to achieve the largest power at a risk of an empirical FDR above q *and* possibly knows from field knowledge that the number of signal variables is not too small and the SNR not too large, she should choose ADAGES with K =10 – 20. Alternatively, for small $|\mathcal{S}_0|$ or large SNR, the user can increase the knockoff runs (e.g. $K = 25$) to reduce the FDR.
- If the user primarily wants an FDR control at q but still improvements over vanilla knockoffs, she should choose uKO.
- We do not recommend using pKO since they have the lowest power in our simulation and their results are very sensitive to the type of knockoff construction.

Of course, these recommendations are just rules of thumb based on the simulation findings and should not be taken for granted. The behaviour of the FDR and power could differ for different model settings. The reader has to keep in mind that we have generated models with a Toeplitz covariance matrix, which is an idealized structure often used in simulation models. Dependencies between variables can be much stronger and more complex in practice. We remind the reader that FX knockoffs sometimes showed a strange behaviour for equi-correlations in the simulations of Section 3.9. Therefore, further research should investigate the performance of multiple knockoffs for other correlation structures.

Chapter 6

Conclusion

In this work, we have introduced the knockoff filter, a variable selection technique that achieves finite sample FDR control without any assumptions on the coefficient sizes, sparsity, feature correlations or noise level. We have started with fixed-X knockoffs (Barber and Candès 2015), which assume a fixed design matrix and can only be applied to low-dimensional Gaussian linear models. Our simulations contribute to research in two ways: i.) We empirically support that knockoffs yield more power than BH and BY while maintaining FDR control for almost all investigated settings. An exception are models with equi-correlated variables where BH is relatively robust, but knockoffs become very conservative with increasing correlation strength. Since the knockoff filter estimates models with $2p$ instead of p variables, the procedure could leverage the "correlation problem" by construction. The BH selection was more powerful than knockoffs and is thus preferable in those situations. Fortunately, it is possible to check the underlying correlation structure in practice, such that the user can evaluate the choice between BH and knockoffs based on the severeness of the correlation strength prior to their application. ii.) We have empirically compared the performance of different score functions. While empirical FDR control is ensured for all score functions as long as they satisfy the antisymmetry and sufficiency property, their choice greatly influences the power. We conclude that the more information and "variable selection character" a score function encodes, the better its performance. The LLSM statistic for example, which is based on Lasso and contains the information when a variable enters the regularization path, performs best, and we recommend using it in practice. There is still a lot of room for further research. It would be profitable to study the theoretical power curves of specific score functions in more detail (Weinstein et al. 2017; Ke et al. 2020) or to develop new statistics that are tailored to certain model settings, e.g. for high correlations, sparse models or specific coefficient structures.
We have continued with a brief description of model-X knockoffs (Candès et al. 2018), where we focus on emphasizing the differences to their predecessor. Due to the probabilistic assumptions placed on the design matrix, MX knockoffs are applicable to almost all non-linear and linear models with no dimensionality restrictions. The aim of this chapter was to create the necessary foundation in order to develop open research questions. One unsolved issue was whether FX knockoffs are outdated or still the superior option compared to MX knockoffs in a (misspecified) linear model. We have shown that the MX knockoff filter with the LCD score achieves greater power than FX knockoffs while controlling the FDR for most of our investigated model misspecifications. In addition, we have also found that the performance of the same score statistic can vary dramatically

between FX and MX knockoffs, which should be kept in mind when developing new scores in the future.

One particular research area which drew our attention was the extension to multiple knockoffs. Similar to the idea of ensemble learning, the aggregation of multiple knockoff runs can increase the power and still maintain FDR control. We have presented the three recent aggregation schemes uKO (Xie and Lederer 2021), pKO (Nguyen et al. 2020) and ADAGES (Gui 2020). Since all three methods are relatively new, there is no representative comparative study yet, nor is there any accessible implementation. We provide an accessible implementation of all three procedures with our `multiknockoffs` package in the statistical software `R`. In the corresponding theoretical sections, we have concluded that none of the aggregation schemes guarantees an FDR of at most q, except uKO with the average sequence. Despite the missing bound at q from the theory side, we have empirically shown that all multiple knockoff methods achieve FDR values below q in almost all our model settings. Moreover, they generally detect more signals correctly than vanilla knockoffs. ADAGES has the most power but suffers from an FDR above q for very sparse models or large SNR values. Next, despite yielding similar FDR values as vanilla knockoffs, uKO have uniformly higher power. The aggregation by pKO, however, was often too conservative, sensitive to changes in the knockoff construction, and has low power. Depending on how much the researcher wants to risk to have an actual FDR above q for a potential higher power, we either recommend ADAGES or uKO.

The discrepancy between the theoretical and empirical results suggest that the theoretical bounds might not be sharp and can be improved. Nguyen et al. (2020) admit that some of their assumptions are solely placed to make their proofs work. These can be possibly relaxed, and so the FDR bound. Furthermore, none of the FDR control proofs of the aggregation methods incorporates that the dependency of the data $([\mathbf{X}\ \tilde{\mathbf{X}}^{(\mathbf{k})}], \mathbf{y})$ used for each knockoff filter will lead to dependent selection sets $\hat{\mathcal{S}}_k$ across the runs $k = 1, \ldots, K$. Accounting for the dependency would probably relax the theoretical bounds.

From a methodological side, there is also plenty of scope for future research. First, specific steps of the aggregation procedures can be modified. For instance, a different criterion to determine the threshold c for ADAGES could lead to a more desirable FDR control while approximately retaining the detection ability of ADAGES. Turning to uKO, trying different sequences of nominal levels could also yield better results. Another potential approach could be to account for the dependencies between the K knockoff runs by constructing a sequential procedure. A knockoff filter at step k could somehow include the information about the previous selections $\hat{\mathcal{S}}_1, \ldots, \hat{\mathcal{S}}_{k-1}$. For example, the information can be encoded by weighting the score functions after their estimation and before determining the threshold T or T^+. The sequential approach would not only lead to a more accurate bound because it incorporates the dependence but could also lead to more power. So far, there exists no work that has tried to design such a sequential multiple knockoff procedure. Last, in Section 4.5, we have mentioned an alternative approach that simultaneously includes multiple knockoffs $(\tilde{\mathbf{X}}_j^{(1)}, \ldots, \tilde{\mathbf{X}}_j^{(m)})$ for each variable $\mathbf{X}_j$ into one model (Gimenez and Zou 2019; Emery et al. 2020). While the construction of simultaneous knockoffs poses its own difficulties because some form of "extended exchangeability property" has to be fulfilled, they could also probably be more powerful than aggregation-based multiple knockoffs. An extensive comparison between both approaches of multiple knockoffs is not conducted so far. The directions of future research are endless, and the potential to improve multiple knockoffs or even the model-X knockoff procedure (see open questions in Section 4.5) seems very promising.

Bibliography

Ahn, S. C. and A. R. Horenstein (2013). Eigenvalue Ratio Test for the Number of Factors. *Econometrica* 81(3), 1203–1227.

Arias-Castro, E. and S. Chen (2017). Distribution-free multiple testing. *Electronic Journal of Statistics* 11(1), 1983–2001.

Azzalini, A. (2018). *The Skew-Normal and Related Families*. N.Y.: Cambridge University Press.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature* 533(7604), 452–454.

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

Barber, R. F. and E. J. Candès (2019). A knockoff filter for high-dimensional selective inference. *The Annals of Statistics* 47(5), 2504–2537.

Barber, R. F., E. J. Candès, and R. J. Samworth (2020). Robust inference with knockoffs. *The Annals of Statistics* 48(3), 1409–1431.

Bates, S., E. Candès, L. Janson, and W. Wang (2020). Metropolized Knockoff Sampling. *Journal of the American Statistical Association*, to appear.

Begley, C. G. and L. M. Ellis (2012). Raise standards for preclinical cancer research. *Nature* 483(7391), 531–533.

Benjamini, Y. and Y. Hochberg (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.

Benjamini, Y., A. M. Krieger, and D. Yekutieli (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* 93(3), 491–507.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testingunder dependency. *The Annals of Statistics* 29(4), 1165–1188.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. N.Y.: Springer.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze* 8, 3–62.

Boyd, S. P. and L. Vandenberghe (2004). *Convex Optimization*. N.Y.: Cambridge University Press.

Bühlmann, P., M. Kalisch, and L. Meier (2014). High-Dimensional Statistics with a View Toward Applications in Biology. *Annual Review of Statistics and Its Application* 1(1), 255–278.

Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Berlin, Heidelberg: Springer.

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

Chang, A. C. and P. Li (2015). Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Usually Not". *Finance and Economics Discussion Series* 2015(83), 1–26.

Chia, C., M. Sesia, C.-S. Ho, S. S. Jeffrey, J. A. Dionne, E. Candes, and R. T. Howe (2021). Interpretable Classification of Bacterial Raman Spectra with Knockoff Wavelets. *IEEE Journal of Biomedical and Health Informatics*. To appear.

Edwards, D. (2000). *Introduction to Graphical Modelling*. 2nd ed. N.Y.: Springer.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. N.Y.: Cambridge University Press.

Efron, B. and T. Hastie (2016). *Computer Age Statistical Inference: : Algorithms, Evidence, and Data Science*. N.Y.: Cambridge University Press.

Emery, K., S. Hasam, W. S. Noble, and U. Keich (2020). Multiple Competition-Based FDR Control and Its Application to Peptide Detection. *Research in Computational Molecular Biology*. Ed. by R. Schwartz. Cham: Springer International Publishing, 54–71.

G'Sell, M. G., S. Wager, A. Chouldechova, and R. Tibshirani (2015). Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78(2), 423–444.

Gimenez, J. R., A. Ghorbani, and J. Zou (2019). Knockoffs for the Mass: New Feature Importance Statistics with False Discovery Guarantees. *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by K. Chaudhuri and M. Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, 2125–2133.

Gimenez, J. R. and J. Zou (2019). Improving the Stability of the Knockoff Procedure: Multiple Simultaneous Knockoffs and Entropy Maximization. *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by K. Chaudhuri and M. Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, 2184–2192.

Grimmett, G. and D. Stirzaker (2001). *Probability and Random Processes*. 3rd ed. N.Y.: Oxford University Press.

Gui, Y. (2020). ADAGES. *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*. ACM.

Hang, H. and I. Steinwart (2017). A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *The Annals of Statistics* 45(2), 708–743.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6(2), 65–70.

Huang, D. and L. Janson (2020). Relaxing the assumptions of knockoffs by conditioning. *Annals of Statistics* 48, 3021–3042.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine* 2(8), e124.

Javanmard, A. and H. Javadi (2019). False discovery rate control via debiased lasso. *Electronic Journal of Statistics* 13(1), 1212–1253.

Ke, Z. T., J. S. Liu, and Y. Ma (2020). Power of FDR Control Methods: The Impact of Ranking Algorithm, Tampered Design, and Symmetric Statistic. To appear.

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

Meinshausen, N., L. Meier, and P. Bühlmann (2009). p-Values for High-Dimensional Regression. *Journal of the American Statistical Association* 104(488), 1671–1681.

Merlevède, F., M. Peligrad, and E. Rio (2009). Bernstein inequality and moderate deviations under strong mixing conditions. *Institute of Mathematical Statistics Collections. High Dimensional Probability V: The Luminy Volume*. Vol. 5. Institute of Mathematical Statistics, 273–292.

Murphy, K. (2012). *Machine learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press.

Nguyen, T.-B., J.-A. Chevalier, B. Thirion, and S. Arlot (2020). Aggregation of Multiple Knockoffs. *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 7283–7293.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* 349(6251), aac4716–aac4716.

Pati, Y., R. Rezaiifar, and P. Krishnaprasad (1993). Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*. IEEE Comput. Soc. Press, 40–44.

Patterson, E. and M. Sesia (2020). *knockoff: The Knockoff Filter for Controlled Variable Selection*. R package version 0.3.3. https://CRAN.R-project.org/package=knockoff.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. 2nd ed. San Francisco: Morgan Kaufmann.

Ren, Z., Y. Wei, and E. Candès (2020). Derandomizing Knockoffs. To appear.

Rokach, L. (2009). Ensemble-based classifiers. *Artificial Intelligence Review* 33(1-2), 1–39.

Romano, Y., M. Sesia, and E. Candès (2019). Deep Knockoffs. *Journal of the American Statistical Association* 115(532), 1861–1872.

Sesia, M., C. Sabatti, and E. J. Candès (2018). Gene hunting with hidden Markov model knockoffs. *Biometrika* 106(1), 1–18.

Soric, B. (1989). Statistical "Discoveries" and Effect-Size Estimation. *Journal of the American Statistical Association* 84(406), 608–610.

Spector, A. and L. Janson (2021). Powerful Knockoffs via Minimizing Reconstructability. *Annals of Statistics*. To appear.

Stagge, J. H., D. E. Rosenberg, A. M. Abdallah, H. Akbar, N. A. Attallah, and R. James (2019). Assessing data availability and research reproducibility in hydrology and water resources. *Scientific Data* 6(1).

Storey, J. D. and R. Tibshirani (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* 100(16), 9440–9445.

Storey, J. D. (2002). A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 64(3), 479–498.

Storey, J. D., J. E. Taylor, and D. Siegmund (2004). Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 66(1), 187–205.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 58(1), 267–288.

Wang, W. and L. Janson (2020). A Power Analysis of the Conditional Randomization Test and Knockoffs. To appear.

Weinstein, A., R. Barber, and E. Candes (2017). A Power and Prediction Analysis for Knockoffs with Lasso Statistics. To appear.

Weinstein, A., W. Su, M. Bogdan, R. Barber, and E. Candès (2020). A Power Analysis for Knockoffs with the Lasso Coefficient-Difference Statistic. To appear.

Westfall, P. H. and S. S. Young (1989). p Value Adjustments for Multiple Tests in Multivariate Binomial Models. *Journal of the American Statistical Association* 84(407), 780–786.

Wu, Y., D. D. Boos, and L. A. Stefanski (2007). Controlling Variable Selection by the Addition of Pseudovariables. *Journal of the American Statistical Association* 102(477), 235–243.

Xie, F. and J. Lederer (2021). Aggregating Knockoffs for False Discovery Rate Control with an Application to Gut Microbiome Data. *Entropy* 23(2), 230.

Yu, L., T. Kaufmann, and J. Lederer (2021). False Discovery Rates in Biological Networks. *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. Ed. by A. Banerjee and K. Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, 163–171.

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(1), 217–242.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association* 101(476), 1418–1429.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2), 301–320.

Appendix A

Supplementary material

A.1 Definitions and theorems

Definition A.1 (**Schur complement**). *Let p and q be non-negative integers. Assume three matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, with dimensions $p \times p$, $p \times q$, and $q \times q$ respectively, that form the symmetric $(p+q) \times (p+q)$ block matrix*

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{C} \end{bmatrix}.$$

If $\mathbf{D}$ is invertible, then the matrix

$$\mathbf{S}/\mathbf{D} \coloneqq \mathbf{A} - \mathbf{B}^\top \mathbf{D}^{-1} \mathbf{B}$$

is the Schur complement of $\mathbf{M}$ with respect to $\mathbf{D}$.

If $\mathbf{A}$ is invertible, then the matrix

$$\mathbf{M}/\mathbf{A} \coloneqq \mathbf{C} - \mathbf{B}^\top \mathbf{A}^{-1} \mathbf{B}$$

is the Schur complement of $\mathbf{M}$ with respect to $\mathbf{A}$.

Theorem A.2 (**Schur complement and positive (semi)definiteness**). *Let $\mathbf{M}$ be a symmetric block matrix with Schur complement $\mathbf{S}$ as in Definition A.1.*

Let $\mathbf{A}$ be positive definite, then the block matrix $\mathbf{M}$ is positive definite if and only if its Schur complement $\mathbf{S}$ is positive definite:

$$\text{If } \mathbf{A} \succ 0, \text{ then } \mathbf{M} \succ 0 \iff \mathbf{S} \succ 0.$$

Let $\mathbf{A}$ be positive definite, then the block matrix $\mathbf{M}$ is positive semidefinite if and only if its Schur complement $\mathbf{S}$ is positive semidefinite:

$$\text{If } \mathbf{A} \succ 0, \text{ then } \mathbf{M} \succeq 0 \iff \mathbf{S} \succeq 0.$$

See Boyd and Vandenberghe (2004, pp. 650–651) for more information about Schur complements.

Definition A.3 (**Conditional moments: Multivariate normal distribution**). *Define* $\mathbf{X} = (X_1, \ldots, X_n)^\top$ *and* $\mathbf{Y} = (Y_1, \ldots, Y_m)^\top$ *to be two n- and m- dimensional multivariate normal distributions forming the (n+m)-dimensional multivariate distributed random vector* $\mathbf{Z}$*, that is*

$$\mathbf{Z} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_\mathbf{X} \\ \mu_\mathbf{Y} \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma_X} & \mathbf{\Sigma_{XY}} \\ \mathbf{\Sigma_{YX}} & \mathbf{\Sigma_Y} \end{pmatrix} \right),$$

where $\mu_\mathbf{X}$ *and* $\mu_\mathbf{Y}$ *are the mean vectors of* $\mathbf{X}$ *and* $\mathbf{Y}$*.* $\mathbf{\Sigma_X}$ *and* $\mathbf{\Sigma_Y}$ *are the variance-covariance matrices of* $\mathbf{X}$ *and* $\mathbf{Y}$*, whereas* $\mathbf{\Sigma_{XY}}$ *defines the covariances between the elements of* $\mathbf{X}$ *and* $\mathbf{Y}$*. Then, the conditional distribution* $\mathbf{Y}|\mathbf{X}$ *is m-dimensional multivariate normal with expectation and covariance matrix equal to*

$$\begin{aligned} \mathbb{E}[\mathbf{Y}|\mathbf{X} = \mathbf{x}] &= \mu_\mathbf{X} + \mathbf{\Sigma_{YX}}\mathbf{\Sigma_X^{-1}}(\mathbf{x} - \mu_\mathbf{X}), \\ \mathrm{Var}[\mathbf{Y}|\mathbf{X} = \mathbf{x}] &= \mathbf{\Sigma_Y} - \mathbf{\Sigma_{YX}}\mathbf{\Sigma_X^{-1}}\mathbf{\Sigma_{XY}}. \end{aligned}$$

See Bishop (2006, pp. 85–88) for more information and a derivation.

Definition A.4 (**Kullback-Leibler divergence**). *Let* P *and* Q *be discrete probability distributions on the same probability space* $\mathcal{X}$*, then the Kullback-Leibler (KL) divergence of* Q *from* P *is defined as*

$$D(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$

There is also a continuous version of the KL divergence, but we only use the discrete equivalent in Section 4.2.5. See Murphy (2012, pp. 56–61) for more information about the KL divergence.

A.2 Proofs and calculations

A.2.1 Proof of FDR properties

Define the FWER as in (2.1), the FDP as in (2.2) and the FDR as in (2.3). Also, recall the the properties

i.) Under the global null $m_0 = m$, it holds that FWER = FDR.
ii.) Under the general case, FWER $\geq$ FDR.

Proof. We proof each property seperately.

i.) Observe that under the global null, all rejections are false discoveries, so $V = R$. The FDP can be split up in two cases:

If $V = 0$, then $R = 0 \Rightarrow \mathrm{FDP} = V/(R \vee 1) = 0$ by definition.

If $V \geq 1$, then $R = V \Rightarrow \mathrm{FDP} = V/R = R/R = 1$.

This can be summarized as

$$\text{FDP} = \mathbb{1}_{\{V \geq 1\}} = \begin{cases} 1, & \text{if } V \geq 1 \\ 0, & \text{if } V = 0. \end{cases}$$

Taking the expectation leads to

$$\text{FDR} = \mathbb{E}[\text{FDP}] = \mathbb{E}[\mathbb{1}_{\{V \geq 1\}}] = 0 \cdot \mathbb{P}(V = 0) + 1 \cdot \mathbb{P}(V \geq 1) = \mathbb{P}(V \geq 1) = \text{FWER}\,.$$

ii.) From Table 2.1 we observe that we always have $m_0 \leq m$ and $V \leq R$.

If $V = 0$, then $R \geq 0 \Rightarrow$ either $R = 0$ and so $\text{FDP} = 0$ by definition or $R > 0$ and so $\text{FDP} = V/R = 0$ since $V = 0$.

If $V \geq 1$, then $R \geq V \Rightarrow \text{FDP} = V/R \leq 1$.

This can be summarized as

$$\text{FDP} = \begin{cases} \leq 1, & \text{if } V \geq 1 \\ 0, & \text{if } V = 0 \end{cases} \leq \mathbb{1}_{\{V \geq 1\}}.$$

Hence, we end up with

$$\text{FWER} = \mathbb{P}(V \geq 1) = \mathbb{E}[\mathbb{1}_{\{V \geq 1\}}] \geq \mathbb{E}[\text{FDP}] = \text{FDR}$$

□

A.2.2 Proof of knockoff construction formula

Construct the knockoff matrix according to $\tilde{\mathbf{X}} = \mathbf{X}(\mathbf{I} - \boldsymbol{\Sigma}^{-1}\,\text{diag}\{\mathbf{s}\}) + \tilde{\mathbf{U}}\mathbf{C}$, then knockoffs will have the desired correlation structure

$$[\mathbf{X} \;\; \tilde{\mathbf{X}}]^\top [\mathbf{X} \;\; \tilde{\mathbf{X}}] = \begin{bmatrix} \mathbf{X}^\top\mathbf{X} & \mathbf{X}^\top\tilde{\mathbf{X}} \\ \tilde{\mathbf{X}}^\top\mathbf{X} & \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \text{diag}\{\mathbf{s}\} \\ \boldsymbol{\Sigma} - \text{diag}\{\mathbf{s}\} & \boldsymbol{\Sigma} \end{bmatrix}.$$

Proof. We will check each entry of the block matrix on its own.

i.) $\mathbf{X}^\top\mathbf{X} = \boldsymbol{\Sigma}$ holds by definition of the Gram matrix.

ii.) For $\mathbf{X}^\top\tilde{\mathbf{X}}$, we use the fact that $\mathbf{X}^\top\tilde{\mathbf{U}} = \mathbf{0}$. Hence, we have

$$\begin{aligned} \mathbf{X}^\top\tilde{\mathbf{X}} &= \mathbf{X}^\top(\mathbf{X}(\mathbf{I} - \boldsymbol{\Sigma}^{-1}\,\text{diag}\{\mathbf{s}\}) + \tilde{\mathbf{U}}\mathbf{C}) \\ &= \mathbf{X}^\top\mathbf{X}(\mathbf{I} - \boldsymbol{\Sigma}^{-1}\,\text{diag}\{\mathbf{s}\}) + \mathbf{X}^\top\tilde{\mathbf{U}}\mathbf{C} \\ &= \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\,\text{diag}\{\mathbf{s}\} \\ &= \boldsymbol{\Sigma} - \text{diag}\{\mathbf{s}\} \end{aligned}$$

iii.) We can use similar arguments for the other off-digonal entry

$$\begin{aligned} \mathbf{X}^\top\tilde{\mathbf{X}} &= \mathbf{X}^\top(\mathbf{X}(\mathbf{I} - \boldsymbol{\Sigma}^{-1}\,\text{diag}\{\mathbf{s}\}) + \tilde{\mathbf{U}}\mathbf{C}) \\ &= \mathbf{X}^\top\mathbf{X}(\mathbf{I} - \boldsymbol{\Sigma}^{-1}\,\text{diag}\{\mathbf{s}\}) + \mathbf{X}^\top\tilde{\mathbf{U}}\mathbf{C} \\ &= \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\,\text{diag}\{\mathbf{s}\} \\ &= \boldsymbol{\Sigma} - \text{diag}\{\mathbf{s}\} \end{aligned}$$

iv.) Turning to the Gram matrix of the knockoffs $\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}$. Let $\mathbf{D} \coloneqq \mathbf{I} - \boldsymbol{\Sigma}^{-1}\operatorname{diag}\{\mathbf{s}\}$ and note again that $\mathbf{S} = \mathbf{C}^\top\mathbf{C} = 2\operatorname{diag}\{\mathbf{s}\} - \operatorname{diag}\{\mathbf{s}\}\boldsymbol{\Sigma}^{-1}\operatorname{diag}\{\mathbf{s}\}$, then

$$
\begin{aligned}
\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} &= (\mathbf{X}\mathbf{D} + \tilde{\mathbf{U}}\mathbf{C})^\top(\mathbf{X}\mathbf{D} + \tilde{\mathbf{U}}\mathbf{C}) \\
&= (\mathbf{D}^\top\mathbf{X}^\top + \mathbf{C}^\top\tilde{\mathbf{U}}^\top)(\mathbf{X}\mathbf{D} + \tilde{\mathbf{U}}\mathbf{C}) \\
&= \mathbf{D}^\top\mathbf{X}^\top\mathbf{X}\mathbf{D} + \mathbf{D}^\top\mathbf{X}^\top\tilde{\mathbf{U}}\mathbf{C} + \mathbf{C}^\top\tilde{\mathbf{U}}^\top\mathbf{X}\mathbf{D} + \mathbf{C}^\top\tilde{\mathbf{U}}^\top\mathbf{U}\mathbf{C} \\
&= \mathbf{D}^\top\boldsymbol{\Sigma}\mathbf{D} + 2\operatorname{diag}\{\mathbf{s}\} - \operatorname{diag}\{\mathbf{s}\}\boldsymbol{\Sigma}^{-1}\operatorname{diag}\{\mathbf{s}\} \\
&= (\mathbf{I} - \boldsymbol{\Sigma}^{-1}\operatorname{diag}\{\mathbf{s}\})^\top\boldsymbol{\Sigma}(\mathbf{I} - \boldsymbol{\Sigma}^{-1}\operatorname{diag}\{\mathbf{s}\}) + 2\operatorname{diag}\{\mathbf{s}\} - \operatorname{diag}\{\mathbf{s}\}\boldsymbol{\Sigma}^{-1}\operatorname{diag}\{\mathbf{s}\} \\
&= (\mathbf{I} - \boldsymbol{\Sigma}^{-1}\operatorname{diag}\{\mathbf{s}\})^\top(\boldsymbol{\Sigma} - \operatorname{diag}\{\mathbf{s}\}) + 2\operatorname{diag}\{\mathbf{s}\} - \operatorname{diag}\{\mathbf{s}\}\boldsymbol{\Sigma}^{-1}\operatorname{diag}\{\mathbf{s}\} \\
&= \boldsymbol{\Sigma} - \operatorname{diag}\{\mathbf{s}\} - \operatorname{diag}\{\mathbf{s}\}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma} + \operatorname{diag}\{\mathbf{s}\}\boldsymbol{\Sigma}^{-1}\operatorname{diag}\{\mathbf{s}\} + 2\operatorname{diag}\{\mathbf{s}\} - \operatorname{diag}\{\mathbf{s}\}\boldsymbol{\Sigma}^{-1}\operatorname{diag}\{\mathbf{s}\} \\
&= \boldsymbol{\Sigma}
\end{aligned}
$$

□

A.2.3 Auxiliary lemma for knockoff theory and its proof

Lemma A.5. *Let $X \sim Bin(n, p)$, where $n \in \mathbb{N}$ being the number of trials and $p \in [0, 1]$ the success probability. Then,*

$$
\mathbb{E}\left[\frac{X}{1+n-X}\right] = \frac{p}{1-p}.
$$

Proof.

$$
\begin{aligned}
\mathbb{E}\left[\frac{X}{1+n-X}\right] &= \mathbb{E}\left[\frac{X}{1+n-X} \cdot \mathbb{1}_{\{X>0\}}\right] \\
&= \sum_{k=1}^{n} \mathbb{P}\{X = k\} \cdot \frac{k}{1+n-k} \\
&= \sum_{k=1}^{n} \binom{n}{k} p^k (1-p)^{n-k} \cdot \frac{k}{1+n-k} \\
&= \sum_{k=1}^{n} p^k (1-p)^{n-k} \cdot \frac{n!}{k!(n-k)!} \cdot \frac{k}{1+n-k} \\
&= \frac{p(1-p)}{p(1-p)} \sum_{k=1}^{n} p^k (1-p)^{n-k} \cdot \frac{n!}{(k-1)!(n-k+1)!} \\
&= \frac{p}{1-p} \sum_{k=1}^{n} p^{k-1} (1-p)^{n-k+1} \cdot \frac{n!}{(k-1)!(n-k+1)!} \\
&= \frac{p}{1-p} \sum_{k=1}^{n} \mathbb{P}\{X = k-1\} \\
&\leq \frac{p}{1-p}
\end{aligned}
$$

In the last step, we use the fact that the sum over all events for X must be 1, and with $k-1$, the sum is smaller or equal 1. □

A.2.4 Proof of ADAGES FDR control

Proof. First, we can re-write the FDR of the model obtained by ADAGES with the law of iterated expectations

$$\mathbb{E}\left[\frac{|\hat{\mathcal{S}}_{(c^*)} \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}_{(c^*)}| \vee 1}\right] = \mathbb{E}\left\{\mathbb{E}\left[\frac{|\hat{\mathcal{S}}_{(c^*)} \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}_{(c^*)}| \vee 1} \;\middle|\; \hat{\mathcal{S}}_1, \ldots, \hat{\mathcal{S}}_K\right]\right\}.$$

We can lower bound the sum of the variable set cardinalities of the K knockoff runs by

$$\sum_{k=1}^{K} |\hat{\mathcal{S}}_k| = \sum_{j=1}^{p} m_j \geq \sum_{j:m_j \geq c^*}^{p} m_j \geq c^* |\hat{\mathcal{S}}_{(c^*)}|$$

where the first inequality comes from bounding the sum of all p counts m_j by only those exceeding the optimal threshold c^*. Since each count is at least c^* and we have $|\hat{\mathcal{S}}_{(c^*)}|$ of them, we can bound the term further. The same inequalities also hold for the number of the false positives of each knockoff run

$$\sum_{k=1}^{K} |\hat{\mathcal{S}}_k \cap \mathcal{H}_0| = \sum_{j \in \mathcal{H}_0} m_j \geq \sum_{j \in \mathcal{H}_0 : m_j \geq c^*}^{p} m_j \geq c^* |\hat{\mathcal{S}}_{(c^*)} \cap \mathcal{H}_0|.$$

We use the false positives inequalities to connect the overall FDR with the individual FDR_k of each knockoff run

$$\mathbb{E}\left\{\mathbb{E}\left[\frac{|\hat{\mathcal{S}}_{(c^*)} \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}_{(c^*)}| \vee 1} \;\middle|\; \hat{\mathcal{S}}_1, \ldots, \hat{\mathcal{S}}_K\right]\right\} \leq \mathbb{E}\left\{\frac{1}{c^*} \sum_{k=1}^{K} \mathbb{E}\left[\frac{|\hat{\mathcal{S}}_k \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}_{(c^*)}| \vee 1} \;\middle|\; \hat{\mathcal{S}}_1, \ldots, \hat{\mathcal{S}}_K\right]\right\}.$$

We can use the requirement of c^* to lie in the candidate region defined in (5.5), that is $|\hat{\mathcal{S}}_{(c^*)}| \geq \frac{1}{K} \sum_{k=1}^{K} |\hat{\mathcal{S}}_k|$. Hence, we obtain

$$\begin{aligned}
\mathbb{E}\left[\frac{|\hat{\mathcal{S}}_{(c^*)} \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}_{(c^*)}| \vee 1}\right] &\leq \mathbb{E}\left\{\frac{1}{c^*} \sum_{k=1}^{K} \mathbb{E}\left[\frac{|\hat{\mathcal{S}}_k \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}_{(c^*)}| \vee 1} \;\middle|\; \hat{\mathcal{S}}_1, \ldots, \hat{\mathcal{S}}_K\right]\right\} \\
&\leq \mathbb{E}\left\{\frac{1}{Kc^*} \sum_{l=1}^{K} \sum_{k=1}^{K} \mathbb{E}\left[\frac{|\hat{\mathcal{S}}_k \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}_l| \vee 1} \;\middle|\; \hat{\mathcal{S}}_1, \ldots, \hat{\mathcal{S}}_K\right]\right\}
\end{aligned}$$

We extend this quantity by a multiplication with $\frac{|\hat{\mathcal{S}}_k|}{|\hat{\mathcal{S}}_k|}$, and so we end up with

$$\begin{aligned}
&= \mathbb{E}\left\{\frac{1}{Kc^*} \sum_{l=1}^{K} \sum_{k=1}^{K} \mathbb{E}\left[\frac{|\hat{\mathcal{S}}_k \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}_k| \vee 1} \cdot \frac{|\hat{\mathcal{S}}_k|}{|\hat{\mathcal{S}}_l|} \;\middle|\; \hat{\mathcal{S}}_1, \ldots, \hat{\mathcal{S}}_K\right]\right\} \\
&= \mathbb{E}\left\{\frac{1}{Kc^*} \sum_{l=1}^{K} \sum_{k=1}^{K} \mathbb{E}\left[\mathrm{FDP}_k \cdot \frac{|\hat{\mathcal{S}}_k|}{|\hat{\mathcal{S}}_l|} \;\middle|\; \hat{\mathcal{S}}_1, \ldots, \hat{\mathcal{S}}_K\right]\right\} \\
&= \mathbb{E}\left\{\frac{1}{Kc^*} \sum_{k=1}^{K} \mathrm{FDP}_k \, \mathbb{E}\left[|\hat{\mathcal{S}}_k| \sum_{l=1}^{K} \frac{1}{|\hat{\mathcal{S}}_l|} \;\middle|\; \hat{\mathcal{S}}_1, \ldots, \hat{\mathcal{S}}_K\right]\right\} \\
&\leq \mathbb{E}\left\{\frac{1}{Kc^*} \sum_{k=1}^{K} \mathrm{FDP}_k \left(\max_{1 \leq k \leq K} |\hat{\mathcal{S}}_k| \sum_{l=1}^{K} \frac{1}{|\hat{\mathcal{S}}_l|}\right)\right\} \\
&= \frac{1}{Kc^*} \sum_{k=1}^{K} \mathrm{FDR}_k \left(\max_{1 \leq k \leq K} |\hat{\mathcal{S}}_k| \sum_{l=1}^{K} \frac{1}{|\hat{\mathcal{S}}_l|}\right).
\end{aligned}$$

Since each of the K knockoff runs achieves FDR control at q, we obtain

$$\mathbb{E}\left[\frac{|\hat{\mathcal{S}}_{(c^*)} \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}_{(c^*)}| \vee 1}\right] \leq \frac{1}{Kc^*} \sum_{k=1}^{K} \mathrm{FDR}_k \left(\max_{1 \leq k \leq K} |\hat{\mathcal{S}}_k| \sum_{l=1}^{K} \frac{1}{|\hat{\mathcal{S}}_l|} \right) \leq \frac{1}{Kc^*} Kq\lambda c^* = \lambda q,$$

where λ is a bound defined as

$$\lambda \geq \max_{1 \leq k \leq K} \frac{|\hat{\mathcal{S}}_k|}{c^*} \sum_{l=1}^{K} \frac{1}{|\hat{\mathcal{S}}_l|}.$$

□

A.3 Figures

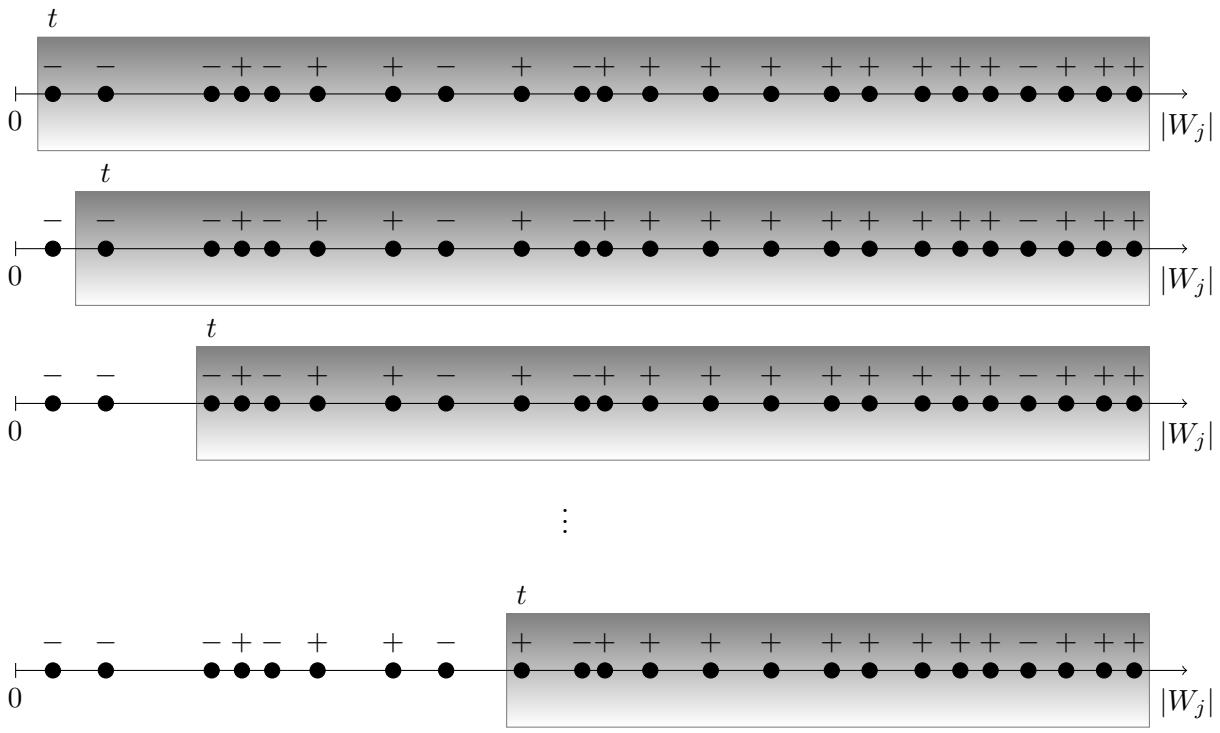
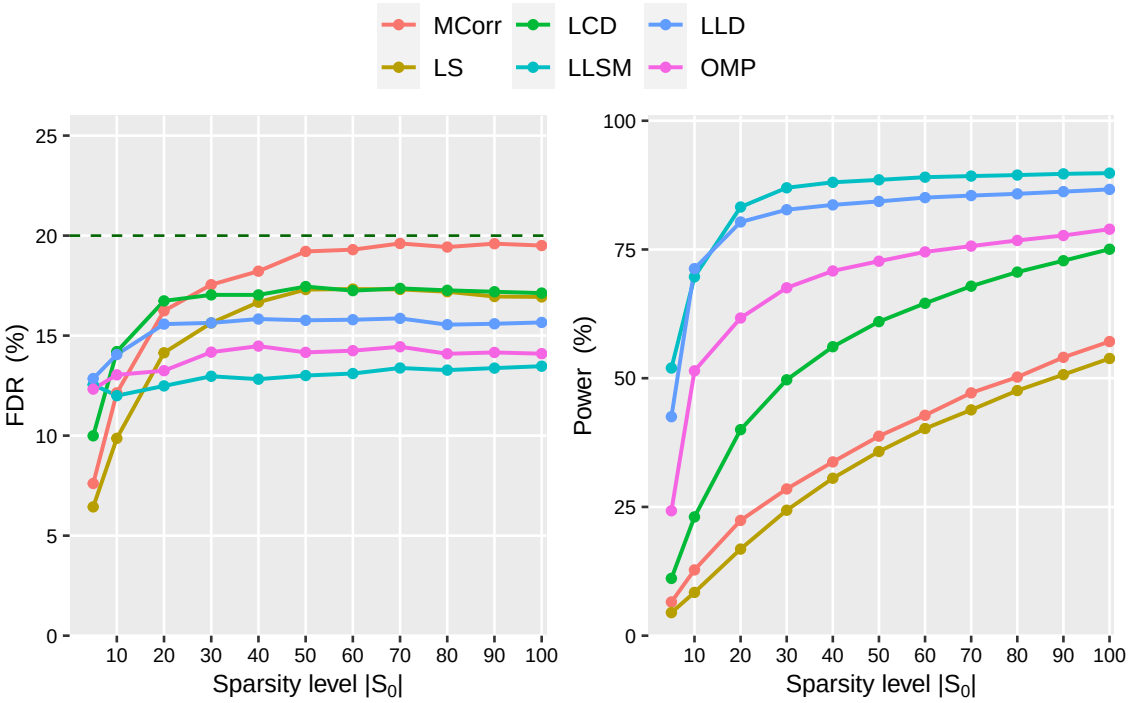


Figure A.1: Knockoffs: View as a process



Nominal level $q = 0.2$. The model is generated with $n = 1000$, $p = 300$, varying $|\mathcal{S}_0|$ and $\rho = 0.25$ in each of the $M = 1000$ trials.

Figure A.2: Simulation score function: Varying sparsity level $|\mathcal{S}_0|$

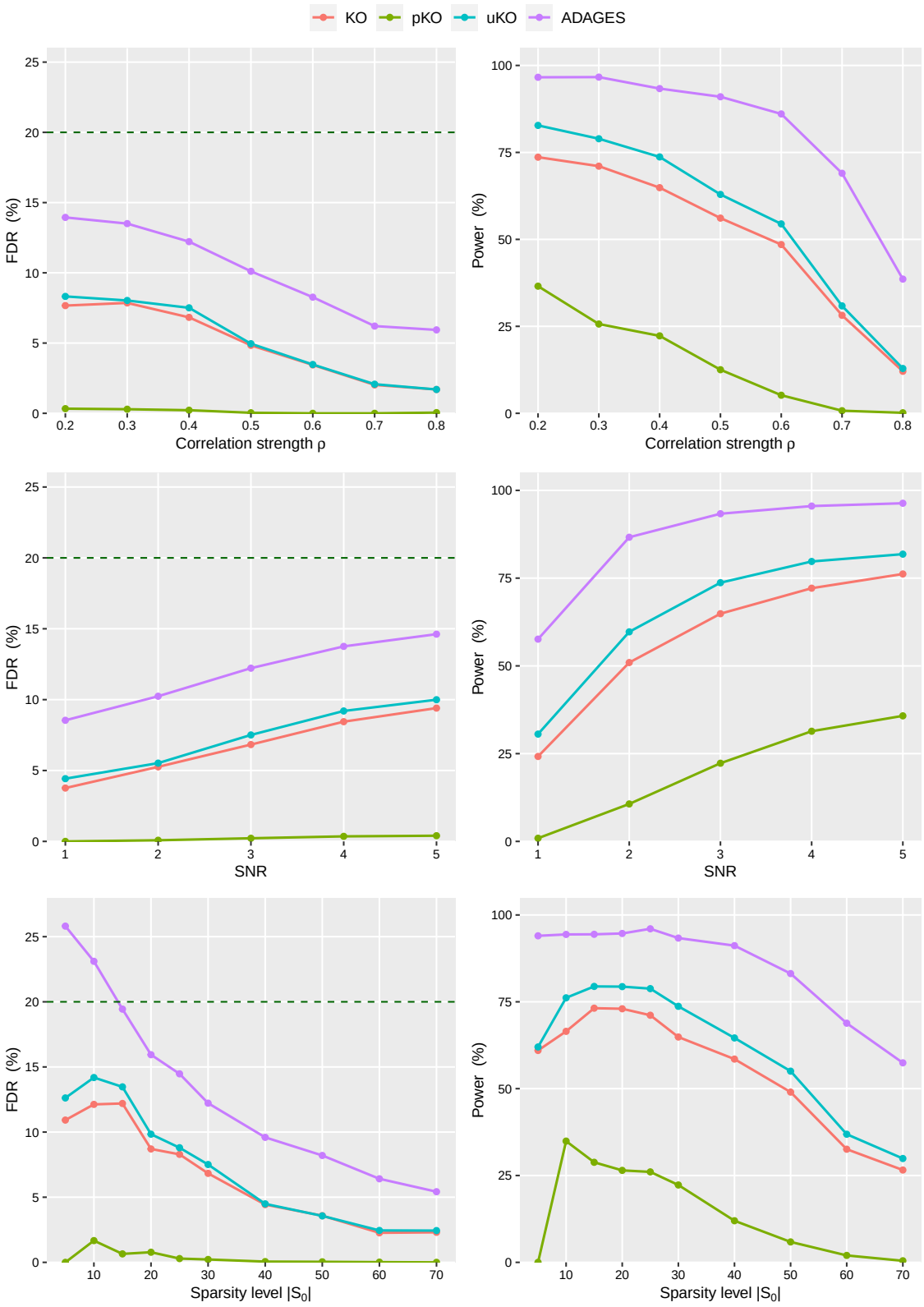


Figure A.3: Multiple knockoffs: Simulation main results, SDP construction

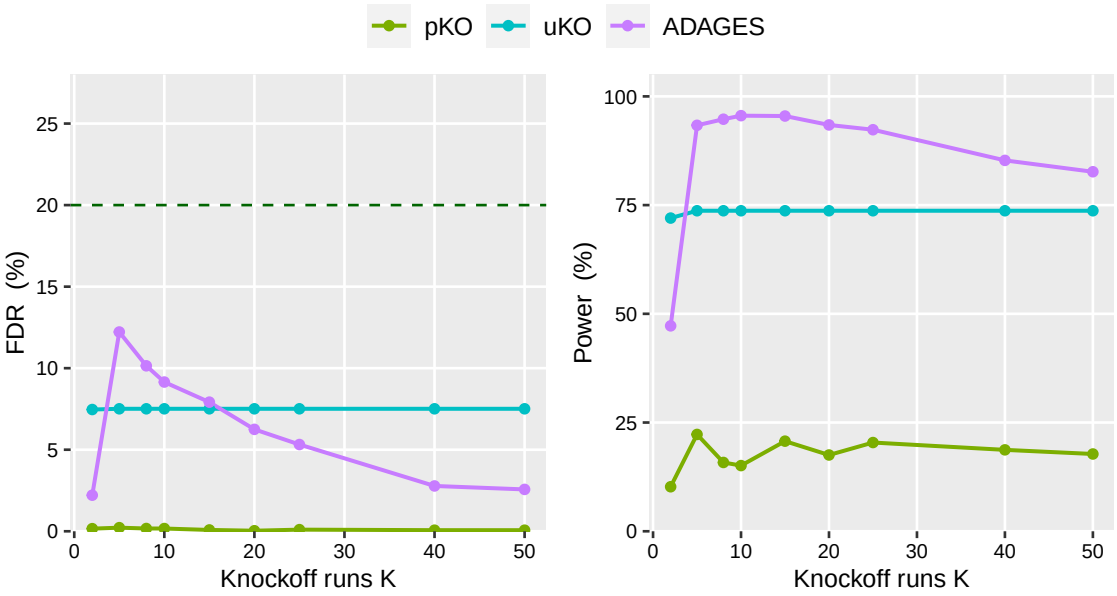


Figure A.4: Multiple knockoffs: Varying K, SDP construction

A.4 Tables

Table A.1: Simulation baseline: Barber & Candès (2015)

Method	FDR	Power	Theoretical FDR control
Knockoff+ (equi-variant)	14.40 %	60.99 %	Yes
Knockoff (equi-variant)	17.82 %	66.73 %	No
Knockoff+ (SDP)	15.05 %	61.54 %	Yes
Knockoff (SDP)	18.72 %	67.50 %	No
BH	18.70 %	48.88 %	No
BY	2.20 %	19.09 %	Yes
BH whitened noise	18.79 %	2.33 %	Yes

A copy of Table 1 in Barber and Candès (2015). The simulated model contains $n = 3000$ observations, $p = 1000$ variables and $|\mathcal{S}_0| = 30$ signals with strength 3.5. Nominal level $q = 0.2$. FDR and power are averages over $M = 600$ iterations. We will omit the BH procedure with whitened noise in our simulations since it was not discussed in this work.

Table A.2: Simulation: FDR and power random signal indices

	ρ	Knockoff	Knockoff+	BH	BY
FDR	0.3	16.98 %	13.31 %	17.86 %	2.98 %
	0.5	14.38 %	10.17 %	18.18 %	2.98 %
	0.7	13.49 %	6.77 %	18.39 %	2.74 %
	0.9	**20.58** %	4.41 %	17.68 %	2.58 %
Power	0.3	71.95 %	65.01 %	59.13 %	25.74 %
	0.5	55.31 %	41.91 %	37.93 %	10.89 %
	0.7	26.64 %	14.04 %	11.99 %	2.23 %
	0.9	6.31 %	1.79 %	1.29 %	0.22 %

Simulation with Toeplitz structure $\Sigma_{j,k} = \rho^{|j-k|}$ and the signal coefficients $\beta_j \neq 0, j \in \mathcal{S}_0$ are chosen randomly.

Table A.3: Simulation score function: FDR for varying correlation structure

	ρ	MCorr	LS	LCD	LLSM	LLD	OMP
$\rho^{\|j-k\|}$	0.3	16.64 %	15.32 %	16.77 %	11.87 %	15.65 %	12.94 %
	0.5	15.59 %	11.25 %	14.72 %	7.95 %	11.82 %	9.82 %
	0.7	11.19 %	7.52 %	9.65 %	2.91 %	4.75 %	5.23 %
	0.9	5.22 %	4.33 %	0.78 %	0.28 %	0.44 %	2.58 %
$\rho_{jk} = \rho$	0.3	13.76 %	12.31 %	16.42 %	12.29 %	15.21 %	13.31 %
	0.5	12.50 %	9.62 %	15.01 %	9.79 %	13.39 %	12.25 %
	0.7	8.99 %	7.39 %	12.12 %	7.25 %	9.17 %	8.79 %
	0.9	6.09 %	3.60 %	6.54 %	4.14 %	5.45 %	4.99 %

Nominal level $q = 0.2$. The model is generated with $n = 1000$, $p = 300$, $|\mathcal{S}_0| = 30$ and varying ρ in each of the $M = 1000$ trials.

Table A.4: Simulation score function: Power for varying correlation structure

	ρ	MCorr	LS	LCD	LLSM	LLD	OMP
$\rho^{\|j-k\|}$	0.3	24.72 %	20.96 %	48.54 %	86.44 %	80.20 %	63.64 %
	0.5	11.60 %	7.94 %	40.98 %	74.70 %	75.96 %	42.22 %
	0.7	3.69 %	2.02 %	36.44%	42.68 %	55.16 %	20.09 %
	0.9	0.62 %	0.32 %	14.99 %	10.06 %	14.53 %	4.71 %
$\rho_{jk} = \rho$	0.3	10.02 %	9.56 %	56.95 %	22.11%	22.89 %	50.47 %
	0.5	6.08 %	3.99 %	43.24 %	10.60 %	12.30 %	27.62 %
	0.7	2.41 %	1.47 %	17.17 %	4.46 %	4.88 %	9.32 %
	0.9	0.57 %	0.34 %	1.9 %	1.00 %	1.20 %	1.37 %

Nominal level $q = 0.2$. The model is generated with $n = 1000$, $p = 300$, $|\mathcal{S}_0| = 30$ and varying ρ in each of the $M = 1000$ trials.

A.5 Simulation: FX vx. MX knockoffs in a linear model

This section presents the details of the simulation as well as the results that we have briefly summarized in Section 4.6. To remind the reader, we will investigate four different types of model misspecifications in our analysis:

1.) Multivariate t-distribution.
2.) Discretized multivariate normal distribution.
3.) Multivariate skew-normal distribution.
4.) Multivariate normal distribution but with a misspecified $\mathbf{\Sigma}$.

In our simulations, we will vary the degree of misspecification within each of the four settings. Our baseline linear model is generated according to

$$\begin{aligned} \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \quad \Sigma_{j,k} = 0.5^{|j-k|} \quad \forall j \neq k \in \{1, \dots, p\}, \quad \sigma^2 = 1, \\ n = 700, \quad p = 300, \quad |\mathcal{S}_0| = 30, \quad |\mathcal{H}_0| = 270. \end{aligned} \tag{A.1}$$

We choose (n, p) such that we are in the normal case of the FX knockoff construction $n \geq 2p$. The 30 non-zero coefficients are selected uniformly at random, and their magnitude is scaled such that the model has an SNR of $\|\mathbf{X}\beta\|_2^2 / n\sigma^2 = 3$. The response is drawn according to the linear model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$. For the FX and MX knockoff construction, we generate SDP knockoffs, and we use knockoff+ (3.12) with a nominal level of $q = 0.1$. We will run the FX and MX knockoff filter with the LLSM statistic

$$W_j = Z_j \vee \tilde{Z}_j \cdot \operatorname{sign}(Z_j - \tilde{Z}_j), \qquad \begin{aligned} Z_j &= \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\} \\ \tilde{Z}_j &= \sup\{\lambda : \hat{\beta}_{j+p}(\lambda) \neq 0\}, \end{aligned}$$

and additionally, we will also perform the MX knockoff filter with the LCD score

$$W_j = |\hat{\beta}_j(\lambda_{\mathrm{CV}})| - |\hat{\beta}_{j+p}(\lambda_{\mathrm{CV}})|.$$

Besides the degree of misspecification in the settings 1.) – 4.), we also vary

i.) the sample size $n \in \{601, 750, 1000, 1250, 1500\}$,
ii.) the sparsity level $|\mathcal{S}_0| \in \{10, 25, 50, 75, 100\}$,
iii.) the correlation strength $\rho \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and
iv.) the SNR $\in \{1, 2, 3, 4, 5\}$.

We conduct each simulation setting with $M = 500$ iterations and compute the FDR and power as averages of those. In the following, we will omit the FDR graphs whenever the FDR is controlled in each parameter setting. We will only display FDR values for additional argumentations or in cases where the FDR control was violated.

1.) Multivariate t-distribution

We start with a non-Gaussian setting by drawing covariates from a multivariate t-distribution. Let $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ and $U \sim \chi^2_\nu$, which are independent of each other, then

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{Z}\sqrt{\frac{\nu}{U}}$$

has a multivariate t_ν-distribution. The degrees of freedom ν determine how "fat" the tails of the distribution are. As ν grows, the multivariate t-distribution gets flatter tails and approaches a multivariate normal distribution. Hence, smaller values of ν encode a stronger model misspecification compared to the normality assumption.
Figure A.5 depicts the average power of the three knockoff filters for various degrees of freedom $\nu \in \{2, 3, 5, 7, 10, 30\}$ over $M = 500$ iterations. Each row of subfigures varies another parameter of the baseline model (A.1), i.e. the first row varies the sample size n, the second the sparsity level $|\mathcal{S}_0|$, the third the correlation strength ρ and the last row the SNR. With increasing degrees of freedom ν – the closer the multivariate t-distribution is to a Gaussian one – the more powerful the three knockoff filters become. Despite the model misspecification, the MX knockoff filter with the LCD statistic has more power than the FX construction in almost every setting. While there are some parameter settings where both are close (e.g. for sparse models $|\mathcal{S}_0| = 10$), there are numerous models where the power of MX LCD is considerably larger. For instance, for a correlation strength of $\rho = 0.7$ and a multivariate t_7-distribution, MX LCD leads to a power of 36%, being almost 30 percentage points larger than the power of FX LLSM.
Even though the FDR control of FX knockoffs does not require any assumptions on the origin of the covariates, their power is influenced by the severeness of fat tails. FX knockoffs have more power the closer the underlying feature distribution is to a multivariate Gaussian one. Although a theoretic power analysis of FX knockoffs is beyond the scope of this work, we can explain this observation intuitively. We have already concluded that the FX knockoff construction closely resembles the second-order approximation of MX knockoffs (compare (3.2) and Definition 4.3). Since the latter improves in quality the closer the underlying distribution is to a multivariate Gaussian, we might expect the same for FX knockoffs.
The most striking observation is the power differences between FX LLSM and MX LLSM. While the LLSM score is the most powerful one for the FX construction (see simulation in Section 3.9.3), it performs extraordinarily poor for MX knockoffs, having the lowest power among the three filters. For example, in a model with $|\mathcal{S}_0| = 25$ signal variables and an underlying t_{30}-distribution, FX LLSM and MX LCD have a power of around 75 – 80%, whereas MX LLSM only achieves a power of approximately 25%. We continue with a follow-up simulation to find out whether these power differences do also hold for other scores that can be applied to both FX and MX knockoff filters. Figure A.6 visualizes the power of the LLSM, LLD and OMP statistics between the FX and MX filter for a varying SNR.[1] Not only for LLSM but also for LLD and OMP the FX construction outperforms the MX filter. The power gaps become even larger with increasing SNR. We conclude that score functions can perform differently well between FX and MX filters. This observation has not been made so far by the knockoff research community and should be kept in mind when designing new scores in the future. Nevertheless, the MX LCD filter (Figure A.5) results in larger power values than those FX filters presented in Figure A.6. This underlines the robustness of the LCD statistic and justifies its application as a robust and powerful score for MX knockoffs.

[1] See Section 3.9.3 for a definition of the LLD and OMP scores.

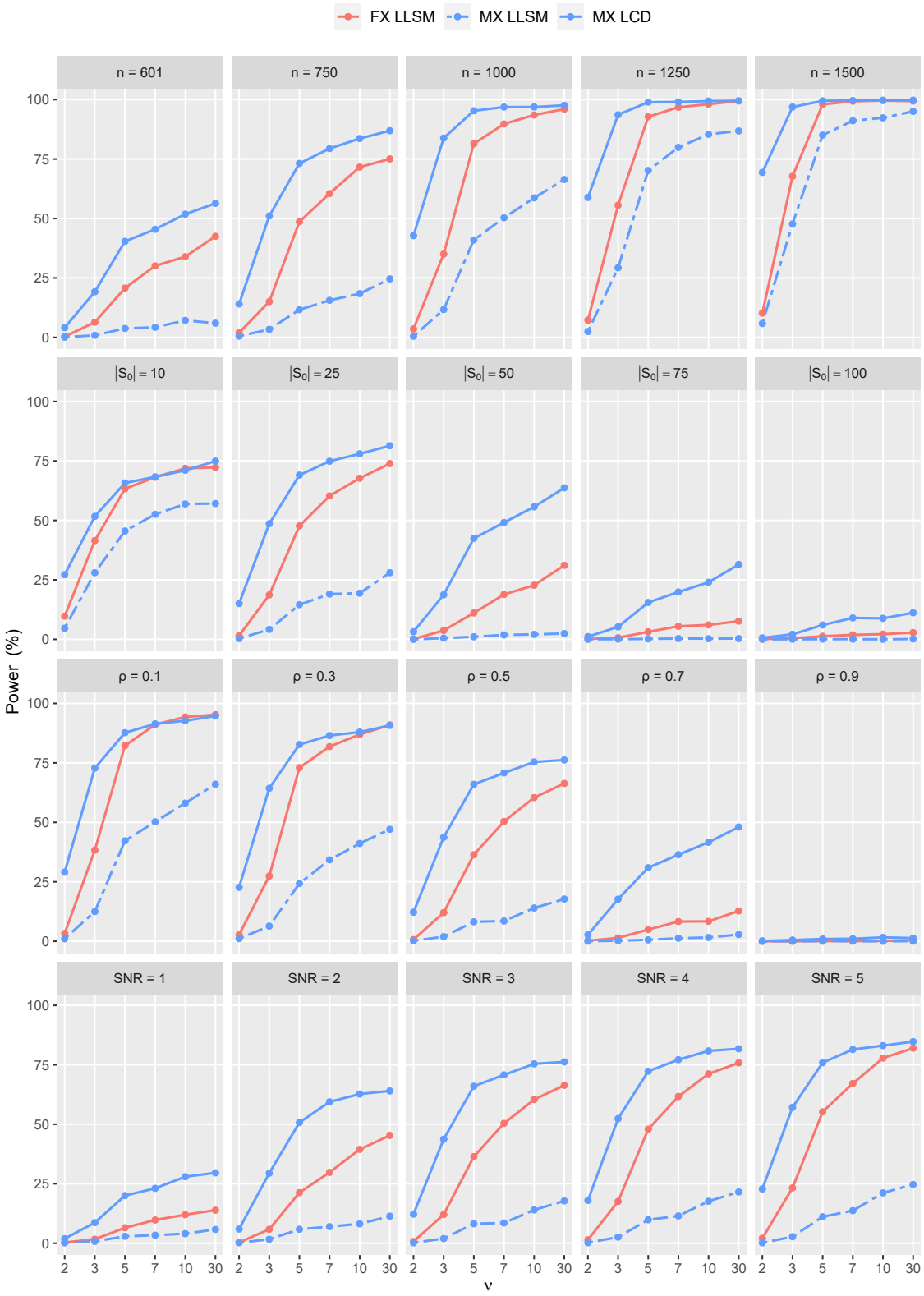


Figure A.5: Model misspecification: Multivariate t-distribution

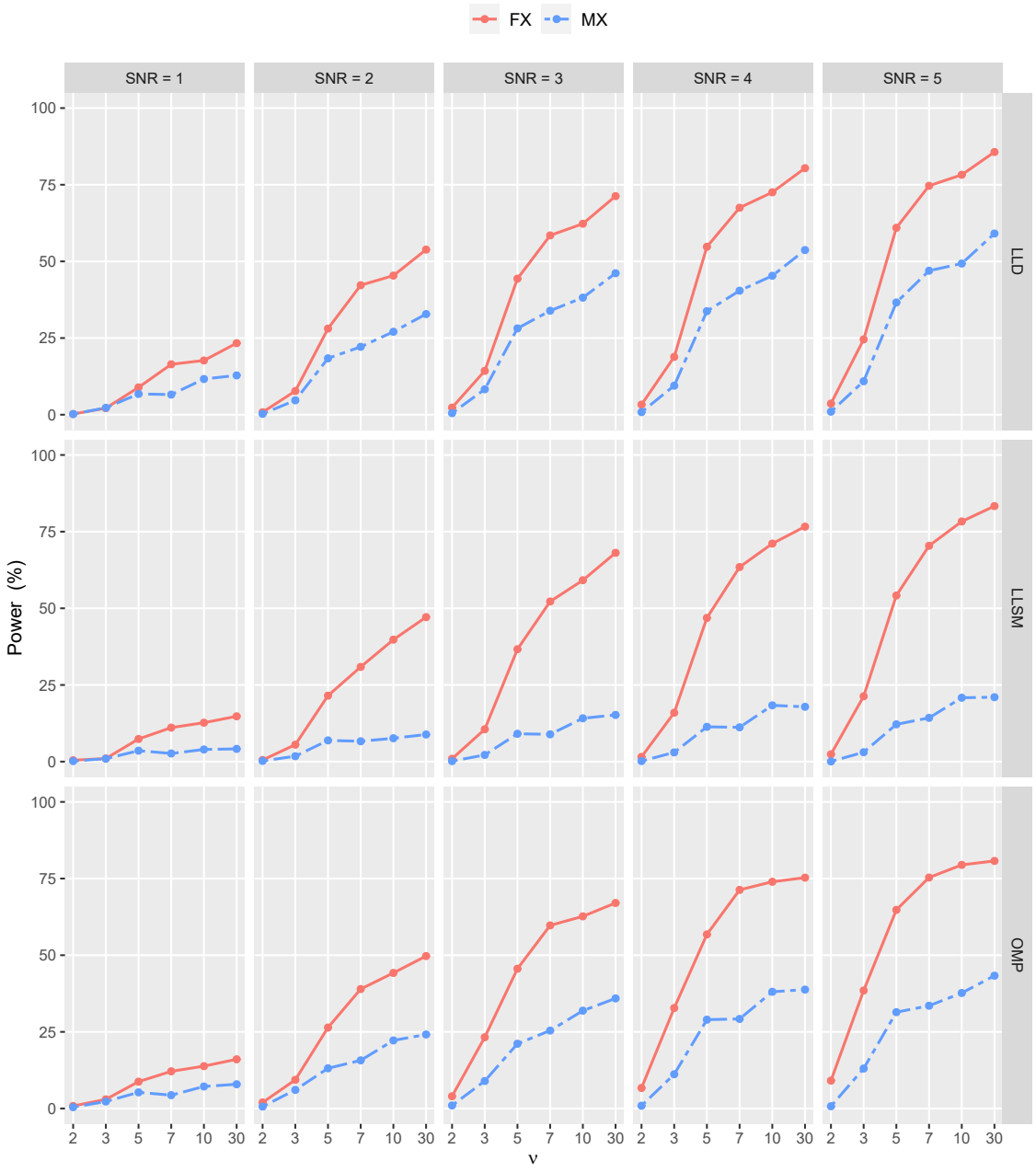


Figure A.6: Model misspecification: Multivariate t-distribution, different scores

2.) Discretization

The next model violations that we are going to investigate are discretized multivariate Gaussian variables at different resolutions, similar to Huang and Janson (2020). We start by drawing covariates from a multivariate normal distribution $\mathbf{X}^{(0)} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, before we discretize each of them by

$$\mathbf{X}_j = \frac{\lfloor \mathbf{X}_j^{(0)} \times G + 1/2 \rfloor}{G}, \quad j = 1, \dots, p.$$

The formula describes that each variable $\mathbf{X}_j^{(0)}$ is rounded to its closest $1/G$-grid value, where G determines the degree of discretization. Small values of G produce a more discrete distribution. For example $G = 0.25$, results in an extremely non-Gaussian distribution with only three different values $\{-4, 0, 4\}$. As $G \to \infty$, the processed covariates become closer to normality $\mathbf{X}_j \to \mathbf{X}_j^{(0)}$.
Figure A.7 shows the average power for different degrees of discretizations $G \in \{0.25, 0.5, 1, 3, 5, 10, 100\}$. Analogously to the previous simulation, MX LCD knockoffs have a larger power than FX LLSM knockoffs in the majority of the settings. MX LLSM knockoffs, however, have a considerably low power again. In contrast to our theoretical intuition, the power has a decreasing and then stagnating pattern with G, even though the covariates become closer to a normal distribution. Huang and Janson (2020) observe the same decreasing pattern with increasing G in a similar simulation. However, they could not find any coherent reasoning for this behaviour either. One possible explanation could be that we intuitively expect the FDR control to be violated for small G and simultaneously to have the highest power at this G since it is not constrained by the same level of FDR control. Figure A.8 illustrates the FDR levels corresponding to the simulation with a varying SNR in Figure A.7. The figure provides evidence in favour of the previous argumentation because we observe an increasing pattern of the FDR and power for smaller G. We have also checked larger values for $G > 100$, but the power and FDR almost remain constant. Although we do not observe FDR control violation in these plots, the same pattern – highest FDR and thus power for small G – may be happening at a scale that is too small to see for knockoffs in this specific setting. Of course, this is only a presumption and should not be taken for granted. However, a theoretical power analysis is beyond the scope of this work since our goal is to compare MX and FX knockoffs under model misspecifications.[2]

[2]First developments of a MX knockoffs power analysis are proposed by Wang and Janson (2020) and Weinstein et al. (2020).

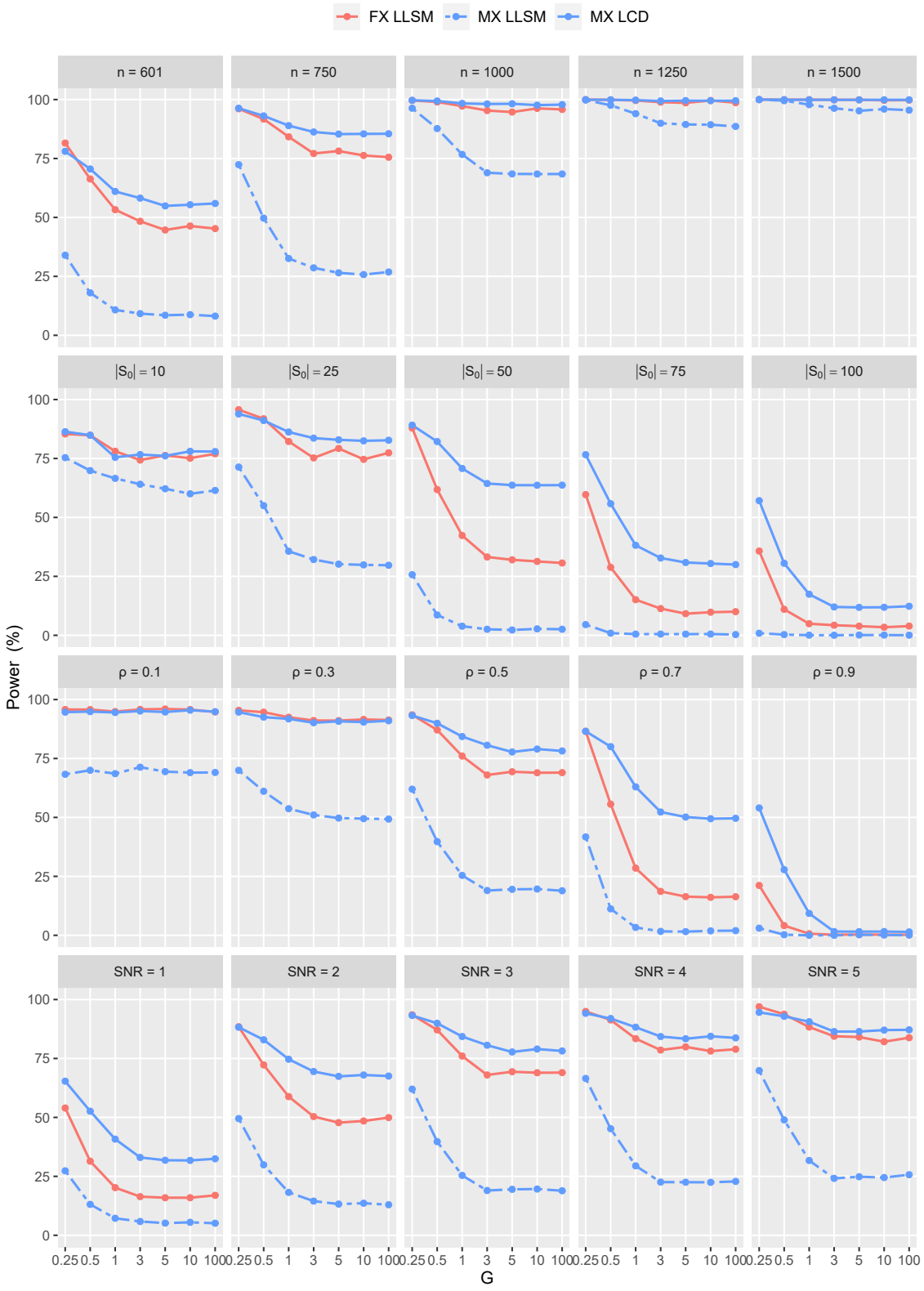


Figure A.7: Model misspecification: Discretization

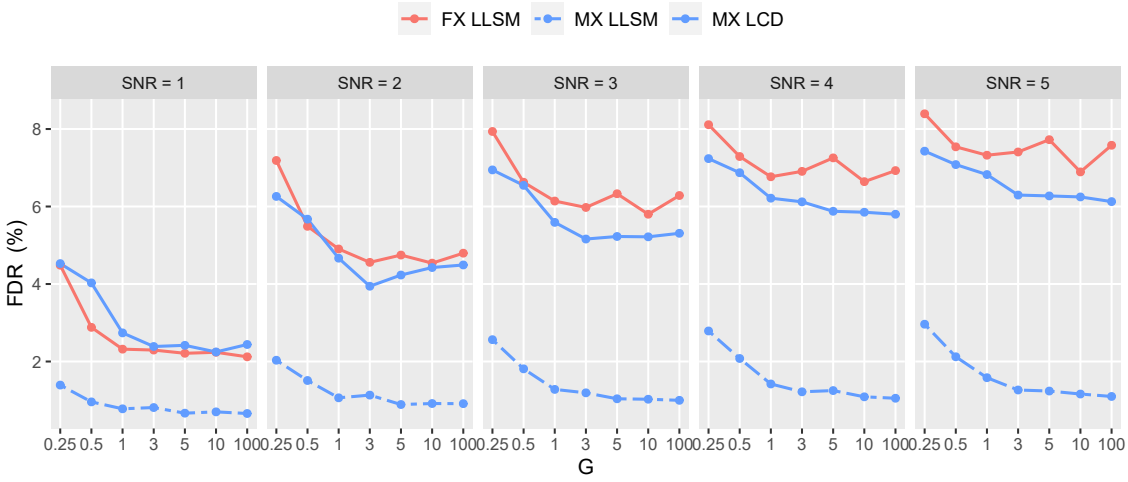


Figure A.8: Model misspecification: Discretization, FDR

3.) Multivariate skew-normal distribution

We proceed by sampling features from a multivariate skew-normal distribution, but we will start with the univariate equivalent. Let $\phi(x)$ be the standard normal probability density function and $\Phi(x)$ its cumulative distribution function, then the probability density function of a skew-normal distribution is defined as

$$f(x) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\alpha\left(\frac{x-\xi}{\omega}\right)\right),$$

with location and scale parameters (ξ, ω) and shape parameter α. Therefore, we can write $X \sim SN(\xi, \omega^2, \alpha)$.[3] The shape parameter α determines the skewness of the distribution, which increases for larger absolute values of α. The distribution is right-skewed (left-skewed) if $\alpha > 0$ ($\alpha < 0$), whereas $\alpha = 0$ results in the normal distribution. The multivariate analogue that we use in our simulations is $\mathbf{X} \sim SN(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha})$, with $\boldsymbol{\xi}, \boldsymbol{\alpha} \in \mathbb{R}^p$ and $\boldsymbol{\Omega} \in \mathbb{R}^{p\times p}$.[4] We choose a location vector of $\mathbf{0} \in \mathbb{R}^p$ and a scale matrix equal to the Toeplitz structure of the baseline model (A.1). We investigate different strengths of right-skewness $\alpha \in \{1, 2, 3, 5, 10, 15\}$, such that the shape vector is $\boldsymbol{\alpha}_j = \alpha, \forall j \in \{1, \dots, p\}$. Figure A.9 displays the power of each parameter setting over $M = 500$ iterations. All three knockoff filters are rather insensitive to changes of α. In addition, we observe the same ranking in power as in the previous two violation settings. Throughout our analysis, we have also tried different scale matrices, larger shape parameters and shape vectors with a random sign $\boldsymbol{\alpha}_j = \pm\alpha, \forall j = 1, \dots, p$, but the power remained almost unaffected for increasing α. Moreover, we have also generated covariates from a multivariate skewed t-distribution (not shown). The power curves were again insensitive to changes of α, but with even larger differences between MX LCD and FX LLSM than in Figure A.9.

[3] Note that (ξ, ω^2) are not the mean and variance of the distribution.

[4] For more information about the univariate and multivariate skew-normal distribution, see Chapters 2 and 5 in Azzalini (2018).

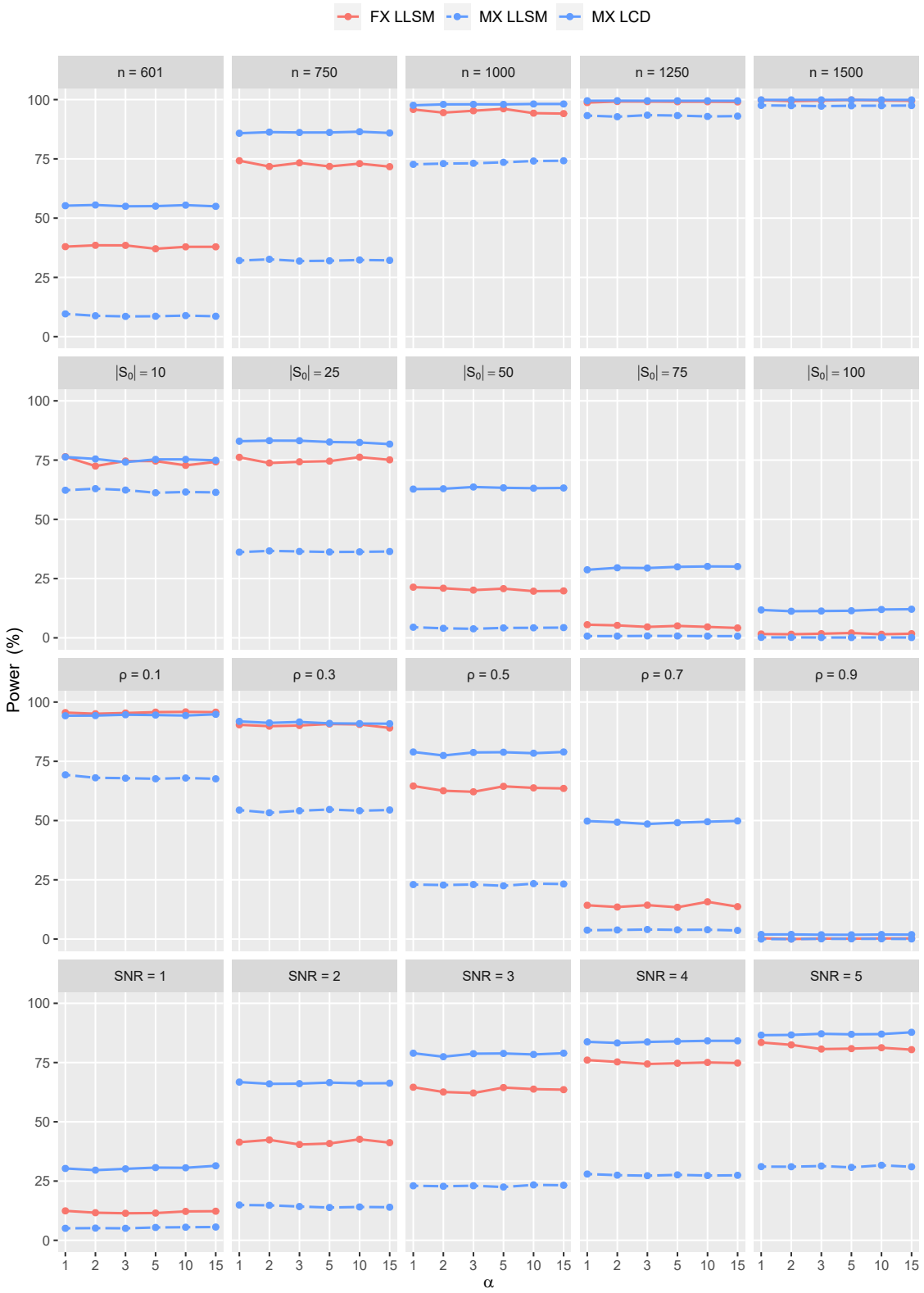


Figure A.9: Model misspecification: Multivariate skew-normal distribution

4.) Misspecified covariance matrix

In the last simulation, we draw features from a multivariate normal distribution, but generate Gaussian knockoffs with a misspecified covariance matrix. In other words, we do not violate the normality assumption but introduce an estimation error. Since the underlying true covariance matrix has a Toeplitz structure, we can construct an estimation error of the covariance matrix by deviating the correlation strength $\tilde{\rho}$ from the true one $\rho = 0.5$ that we have used to generate the covariates. Therefore, our misspeficied covariance matrix has also a Toeplitz structure but with entries $\tilde{\Sigma}_{j,k} = \tilde{\rho}^{|j-k|}$ instead of $\Sigma_{j,k} = \rho^{|j-k|}$, and we define the degree of misspefication as the deviation of the misspecified correlation strength from the truth $\tilde{\rho} - \rho$. We fix the true correlation strength at $\rho = 0.5$ across our simulations and vary the wrong one $\tilde{\rho} \in \{0.1, 0.2, \ldots, 0.8, 0.9\}$.
Figure A.10 presents the power values for different degrees of misspefications $\tilde{\rho} - \rho$. The power of FX knockoffs is independent of $\tilde{\rho} - \rho$ because their construction is not based on the misspecified covariance matrix. For most settings, the power of MX LCD knockoffs is close to 100 % and far higher than the power of FX LLSM knockoffs. This might be the case because the underlying feature distribution is multivariate normal. Moreover, the power of both MX knockoff filters decreases if we overestimate the correlation strength $\tilde{\rho} > \rho$, but increases for an underestimation $\tilde{\rho} < \rho$. The latter scenario indicates that the FDR might not be below our desired nominal level anymore such that the power is not restricted by that control.
Figure A.11 depicts all FDR values corresponding to the simulations of Figure A.10. Indeed, we observe that the FDR is controlled in all settings where we overestimate the correlation strength (or estimate it correctly). On the other hand, the FDR values of the MX knockoff filters have the tendency to exceed the 10% bound for models where we underestimate the correlation strength. In contrast, the FX knockoff filter controls the FDR in every simulation setting.

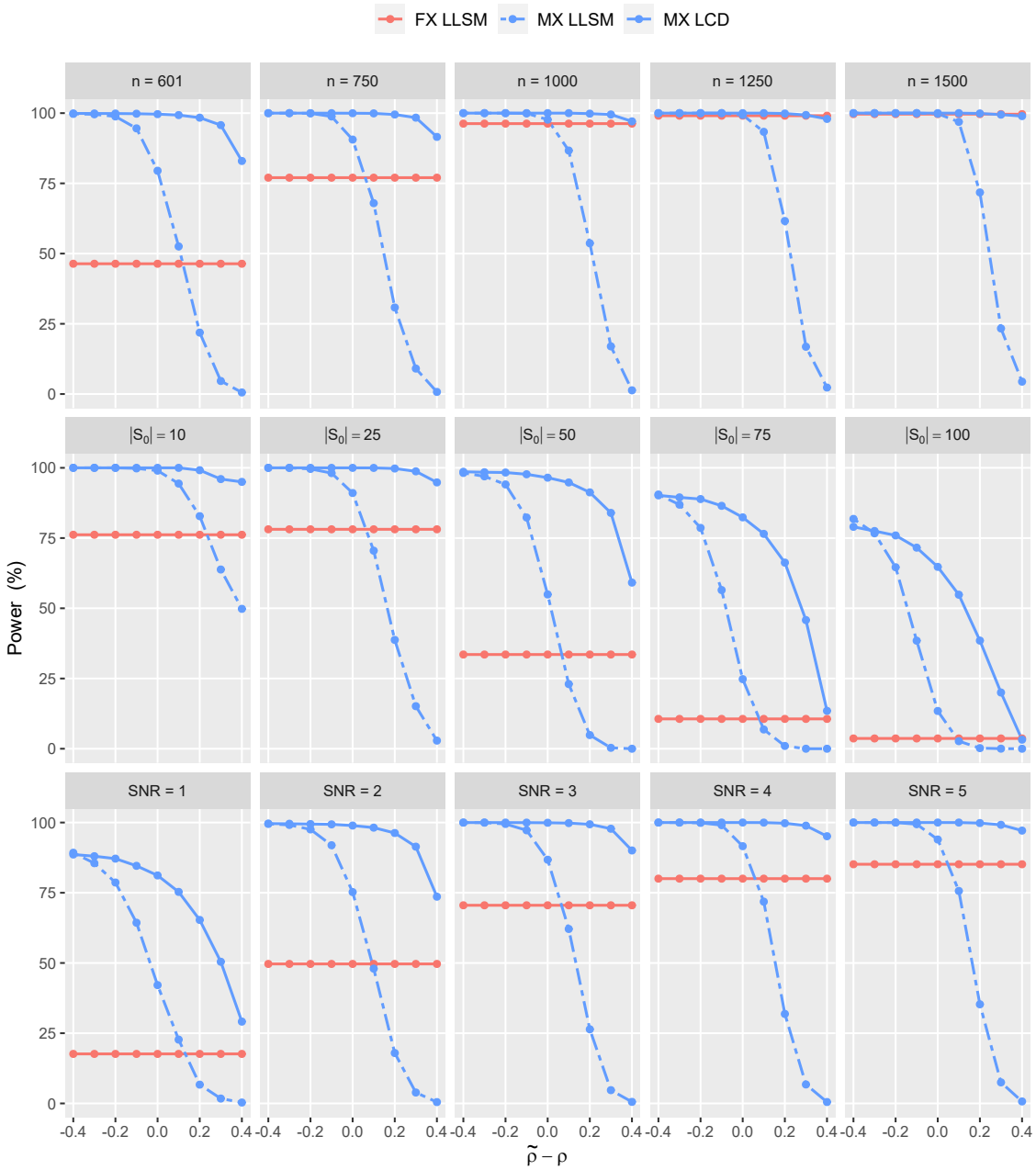


Figure A.10: Model misspecification: Misspecified covariance matrix

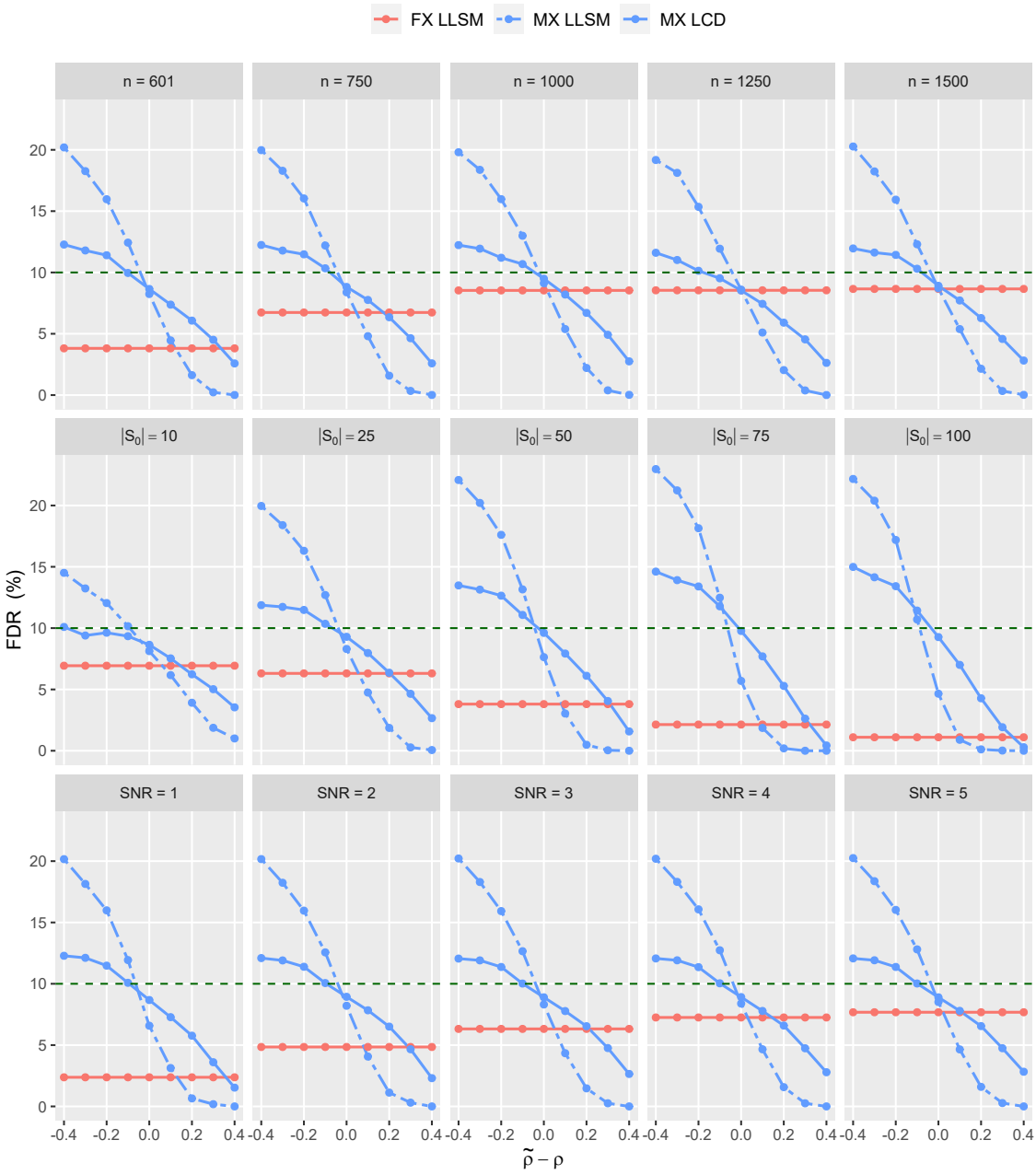
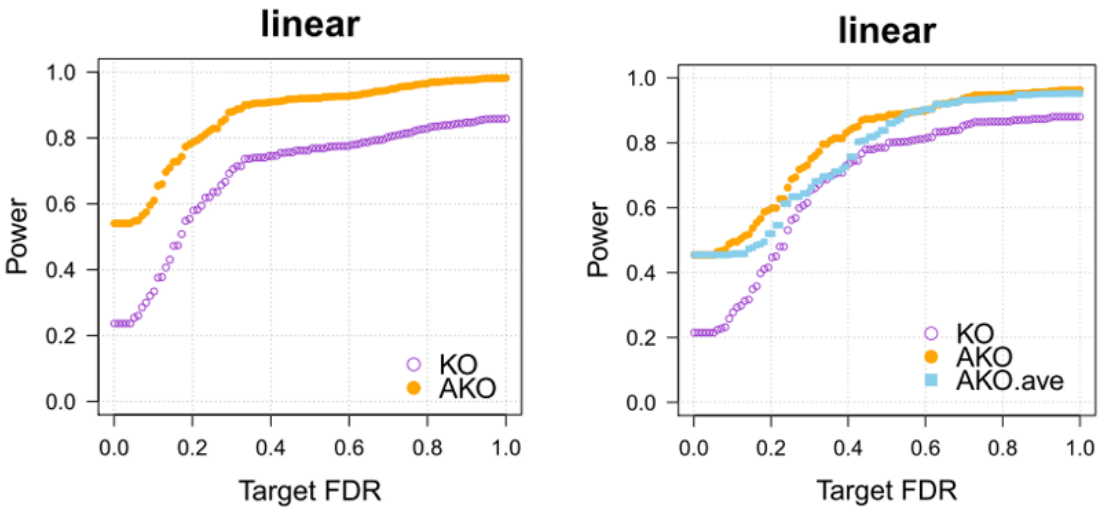


Figure A.11: Model misspecification: Misspecified covariance matrix, FDR

A.6 Union knockoffs: Comments and concerns

In this supplementary section, we will briefly discuss some issues of the simulation conducted in Xie and Lederer (2021).
First, they applied the LLSM score statistic W_j instead of LCD with a cross-validated tuning parameter λ. While this does not affect their main conclusion — union knockoffs are more powerful than the vanilla MX knockoff filter — both union and vanilla knockoffs would have higher power with the LCD score (see Sections 4.6 and A.5).
Looking closely at their simulation code, Xie and Lederer treat the data $(\mathbf{X}, \mathbf{y})$ as fixed, so they do not draw a new covariate matrix and noise in each iteration.[5] The randomness in their simulation comes only from generating new knockoff matrices. The covariates are random in the MX knockoff framework, and both the covariates as well as the noise should be treated as random to cover a larger space of variation. Furthermore, they only use $M = 100$ iterations. Both concerns may affect the reproducibility of their results. Indications for the irreproducibility can be seen by comparing Figure 1a and Figure A2 of their work which we have put together in our Figure A.12. According to the authors, the two figures are based on the same parameter settings. The left subfigure presents their main results, the superiority of union knockoffs over vanilla knockoffs with $K = 5$ and their recommended sequence. In the right subfigure listed in their appendix, they extend their results by running the simulation again but additionally applying union knockoffs with sequence $q_k = q/K$ (AKO.ave). Even though they used the same parameter settings, we observe power differences of union knockoffs (AKO) of up to 20% between both subfigures. This variability stems partly from the small number of iterations but also from treating the data as fixed. It could be that the data generation used in the left figure is just a "good draw". We repeated their simulation setting but with more iterations (not shown here) and achieved power values similar to the right subfigure. In our upcoming



Both figures are based on the same setting: $n = 200$, $p = 100$, $|\mathcal{S}_0| = 20$, $K = 5$. KO = vanilla knockoffs, AKO = union knockoffs with sequence $q_k = q/2^{k-1}$ and AKO.ave = union knockoffs with sequence $q_k = q/K$. Source: Xie and Lederer (2021, Figure 1 & A2).

Figure A.12: Xie and Lederer: Simulations

[5] This can be seen in https://github.com/LedererLab/aggregated_knockoffs/blob/master/Sim_linear_KO_AKO.R.

simulations of Section 5.4, we will investigate union knockoffs in a different setting with more iterations and with a random design matrix and response. We emphasize again, Xie and Lederer's setting may not affect their conclusion: union knockoffs are more powerful in both subfigures. However, the power level of each method between the figures suffers from variability, which also raises the concern that both figures are just depicted because they are good draws. The variability of the underlying setting could be also a problem if we compare union knockoffs with another knockoff aggregation or FDR controlling method that is less sensitive to different data. To confirm with greater certainty that union knockoffs are more powerful than vanilla knockoffs, we should conduct a simulation that covers more randomness.

Appendix B

`multiknockoffs`: Extended examples

B.1 List of all functions

`agg.ADAGES`	Aggregation step ADAGES
`agg.ADAGES.mod`	Aggregation step modified ADAGES
`agg.pKO`	Aggregation step pKO
`agg.union`	Union of selection sets
`multi.knockfilter`	Determination K score matrices and selection sets
`multi.knockoffs`	Construction multiple knockoff matrices
`quantile.aggregation`	Helper function: Quantile aggregation of the p-values
`run.ADAGES`	Whole ADAGES procedure with multiple knockoffs
`run.pKO`	Whole pKO procedure with multiple knockoffs
`run.uKO`	Whole uKO procedure with multiple knockoffs

B.2 Advanced usage

In this section, we present some additional functions for multiple knockoffs that are implemented in our package `multiknockoffs`. They provide more flexibility than those presented in Chapter 5, which construct multiple knockoff matrices, estimate the K score vectors and determine the aggregated selection set at once.
The function `multi.knockoffs` takes the $n \times p$ data matrix $\mathbf{X}$, the number of desired knockoff matrices K and a function that samples the knockoffs as input, and it returns a list with K knockoff matrices $\tilde{\mathbf{X}}^{(\mathbf{1})}, \ldots, \tilde{\mathbf{X}}^{(\mathbf{K})}$.

```
multi.knockoffs(X, K, knockoffs = create.second_order)
```

The argument `knockoffs` requires a function taking the $n \times p$ data matrix as input and returning the $n \times p$ knockoff matrix. The default method is the second-order approximation, which is also the default in the `knockoff` package (Patterson and Sesia 2020). The user can also choose `create.gaussian` from the `knockoff` package or any other user-defined knockoff sampler.
Once we have drawn the list of K knockoff matrices, we can use them to determine the score vectors $\mathbf{W}^{(k)}$ based on the data $([\mathbf{X}\ \tilde{\mathbf{X}}^{(\mathbf{k})}], \mathbf{y})$ $\forall k$ and the corresponding selection sets $\{\hat{\mathcal{S}}_k : k = 1, \ldots, K\}$. More precisely, we can apply `multi.knockfilter`

```
multi.knockfilter(X, Xk, y, q = 0.2, offset = 1,
                  statistic = stat.glmnet_coefdiff)
```

with the arguments:

- `X` the $n \times p$ design matrix.
- `Xk` a list with K elements containing the $n \times p$ knockoff matrices.
- `y` the $n \times 1$ response vector.
- `q` either a scalar or vector of nominal levels. If a scalar is supplied, then the same nominal level is used for each knockoff run. Default: 0.2.
- `offset` either 0 (knockoff) or 1 (knockoff+). Default: 1.
- `statistic` the function to compute the score statistics. The user can either choose any implemented score function of the `knockoff` package or define one by herself. If manually defined, the score function has to take the matrices $\mathbf{X}$, $\tilde{\mathbf{X}}$ and the vector $\mathbf{y}$ as input and must return a vector of score statistics $\mathbf{W}$. Default: LCD score with a λ tuned by CV `stat.glmnet_coefdiff`.

The function outputs a list containing three elements:

- `W.list` a list containing the K score vectors of each knockoff run.
- `Shat.list` a list containing the K selection sets of each knockoff run.
- `q` a vector containing the nominal levels of each knockoff run.

Both functions, `multi.knockoffs` and `multi.knockfilter`, are especially useful if the user wants to conduct simulations comparing the implemented aggregation schemes, or if she wants to investigate a certain step of the multiple knockoff filter estimation more thoroughly.
With the score vectors and selection sets determined, the three presented aggregation methods can be used to obtain the aggregated selection set. The function

```
agg.union(Shat.list)
```

takes a list of K selection sets as input and outputs the union of them. We implement `agg.union` since the idea of taking the union of selection sets can be applied to any selection procedure and not only in the knockoff framework (see Section 5.1.1). The function just requires that the selection sets are stored in a list. The user can also run the three so far presented functions to carry out the union knockoff steps manually by defining the sequence of nominal levels q_k for each knockoff run, that is $q_k = 0.2/2^{k-1}$ in the example below.

```
K <- 5
Xk <- multi.knockoffs(X, K = K)
#Xie and Lederer's recommended sequence
q <- 0.2/2^((1:K)-1)
knock.res <- multi.knockfilter(X, Xk, y, q = q)
uKO.res <- agg.union(knock.res$Shat.list)
```

We continue with the explanation of `agg.pKO`. Although this command does not have much practical benefit, since pKO can only be applied in the knockoff framework, it is useful in simulations, or if the user wants to investigate the behaviour of the aggregated p-values with different options.

```
agg.pKO(W.list, q = 0.2, gamma = 0.3, offset = 1, method = "BH",
        pvals = FALSE)
```

It takes similar arguments as `run.pKO`:

- `W.list` a list with B elements containing the vectors of scores with length p each.
- `q` defines the nominal level for the FDR control. Default: 0.2.
- `gamma` is a value between $(0, 1)$ which defines the quantile value used for the aggregation. If `gamma = NULL`, the adaptive search by Meinshausen et al. (2009) is used. Default: 0.3.
- `offset` either 0 or 1. Determines if an additional $+1$ is added in the numerator of (5.2). Default: 1.
- `method` is the FDR control method in the last step. Either `"BH"` (default) or `"BY"`.
- `pvals` if the aggregated p-values should be reported (logical). Default: `FALSE`.

The command returns the aggregated selection set, the number of knockoff draws B and (if specified) the aggregated p-values. Similar to the example above, the user can run each step of the pKO procedure manually.

```
K <- 5
Xk <- multi.knockoffs(X, K = K)
knock.res <- multi.knockfilter(X, Xk, y)
pKO.res <- agg.pKO(knock.res$W.list)
```

Note that for the manual execution of pKO, the scalar or vector of nominal levels in `multi.knockfilter` can be arbitrary since we only need the list of score vectors whose computation does not depend on the nominal level.
The last aggregation scheme ADAGES can be performed by

```
agg.ADAGES(Shat.list, p = p)
```

which takes a list of K selection sets and the number of variables p of the model as input, and it returns a list with the aggregated set, the optimal treshold c^* and the number of runs knockoff runs K.[1] Similar to `agg.union`, `agg.ADAGES` is not restricted to the knockoff framework, and it can be generally applied to aggregate multiple runs of variable selection procedures or even different variable selection procedures that have FDR control at q respectively. To run ADAGES manually in the multiple knockoff framework, the user can execute the code below.

```
K <- 5
Xk <- multi.knockoffs(X, K = K)
knock.res <- multi.knockfilter(X, Xk, y)
ADAGES.res <- agg.ADAGES(knock.res$Shat.list, p = p)
```

[1] For the sake of completeness, `agg.ADAGES.mod` runs the modified ADAGES aggregation, and it takes the same input arguments as `agg.ADAGES`.

Change of the optimization technique of knockoff samplers

The (A)SDP optimization is used by default when applying the second-order construction in `multi.knockoffs`, `run.uKO`, `run.pKO` and `run.ADAGES`. However, in the simulations of Section 5.4.2, we have seen that equi-correlated knockoffs yielded better power than the (A)SDP construction for that specific model setting. The user can modify the knockoff construction within our functions, e.g. changing the optimization from (A)SDP to equi-correlation. We illustrate this with `multi.knockoffs`, but it works equivalently for the other functions.

```
equi.knock <- function(X) create.second_order(X, method = "equi")
Xk <- multi.knockoffs(X, K = K, knockoffs = equi.knock)
```

Note that `create.second_order` is a function from the original `knockoff` package, which must be loaded beforehand.

Example of help menu

Union knockoff filter

Description

This function runs the whole union knockoff procedure, i.e. it generates multiple knockoff matrices, estimates the score functions and the selection sets of multiple knockoff runs, which are then aggregated by their union to obtain the final selection set.

Usage

```
run.uKO(
  X,
  y,
  knockoffs = create.second_order,
  statistic = stat.glmnet_coefdiff,
  qk = "decseq",
  q = 0.2,
  K = 5,
  q_seq = NULL,
  offset = 1,
  sets = FALSE
)
```

Arguments

X	n x p matrix or data frame of original variables.
y	response vector of length n.
knockoffs	function for the knockoff construction. It must take the n x p matrix as input and it must return a n x p knockoff matrix. Either choose a knockoff sampler of the knockoff package or define it manually. Default: create.second_order (see below).
statistic	function that computes the score vector *W* of length p. It must take the data matrix, knockoff matrix and response vector as input and outputs a vector of computed scores. Either choose one score statistic from the knockoff package or define it manually. Default: stat.glmnet_coefdiff (see below).
qk	sequence of nominal levels. Either choose "decseq" (default) for *q_{k} = q/2^{k-1}* or "ave" for *q_{k} = q/K*.
q	nominal level for the FDR control. Default: 0.2.
K	number of knockoff runs. Default: 5.
q_seq	manual sequence of nominal level which has to match in length with the number of knockoff runs K. If this argument is specified, qk and q are ignored.
offset	either 0 (knockoff) or 1 (knockoff+). Default: 1.
sets	logical argument if the K selection sets of each knockoff run should be returned. Default: FALSE.

Details

This function requires the installation of the knockoff package prior to its execution.

The default knockoff sampler create.second_order is the second-order Gaussian knockoff construction from the knockoff package.

The default score function stat.glmnet_coefdiff is from the knockoff package. It fits a Lasso regression where the regularization parameter λ is tuned by cross-validation. Then, the score is computed as the difference between

$$W_j = |Z_j| - |\tilde{Z}_j|$$

where *Z_j* and *\tilde{Z}_j* are the coefficient estimates for the jth variable and its knockoff, respectively.

The user has to specify either qk together with q to apply one of the pre-defined nominal levels or has to define the argument q_seq for an own sequence of nominal levels.

Value

A list containing following components:

Shat	aggregated selection set.
K	number of knockoff runs.
FDRbound	theoretical FDR bound.
sets	if specified, individual selection sets of each knockoff run.

References

Xie and Lederer (2021). *Aggregating Knockoffs for False Discovery Rate Control with an Application to Gut Microbiome Data*. Entropy 23(2), 230. https://www.mdpi.com/1099-4300/23/2/230/xml

Examples

```
n <- 400; p <- 200; s_0 <- 30
amplitude <- 1; mu <- rep(0,p); rho <- 0.25
Sigma <- toeplitz(rho^(0:(p-1)))

X <- MASS::mvrnorm(n, mu, Sigma)
nonzero <- sample(p, s_0)
beta <- amplitude * (1:p %in% nonzero)
y <- X %*% beta + rnorm(n)

# Basic usage with default arguments
res.uKO <- run.uKO(X, y, sets = TRUE)
res.uKO

# Advanced usage with customized knockoff construction (equi-correlated)
equi.knock <- function(X) create.second_order(X, method = "equi")
res.uKO <- run.uKO(X, y, knockoffs = equi.knock, sets = TRUE)
res.uKO
```

Figure B.1: Example help menu for `run.uKO`

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

The (multiple) knockoff filter: A powerful variable selection with FDR control

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
Karypidis	Chrysovalantis

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
Cologne, 06.09.2021	V. Karypidis

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.