The knockoff filter: A framework for rigorous variable selection with finite-sample FDR control in high-dimensional settings

Mark Milner

Supervised by Henry Reeve
Level 7
20 Credit Points

March 22, 2025

Acknowledgement of Sources

For all ideas taken from other sources (books, articles, internet), the source of the ideas is mentioned in the main text and fully referenced at the end of the report.

All material which is quoted essentially word-for-word from other sources is given in quotation marks and referenced.

Pictures and diagrams copied from the internet or other sources are labelled with a reference to the web page or book, article etc.

Signed Mark Milner

Date 22/03/2025

Abstract

Consider a scenario where we want to model a target variable and have a large set of possible explanatory covariates. To understand its relationship with our covariates and efficiently model the target, we wish to find a smaller subset of covariates that are truly important in influencing the target whilst disregarding the rest.

The challenge in variable selection lies in controlling false discoveries—ensuring that covariates with no true influence on the target are not mistakenly identified as important predictors. The knockoff filter provides a powerful framework for variable selection with rigid control of false discovery rate (FDR). This methodology offers significant practical advantages, particularly in its finite-sample control of false discoveries, as it operates without relying on any asymptotic properties of test statistics common in traditional methods of variable selection. For that reason, it is especially valuable in high-dimensional settings such as in genetic studies, where the number of potential explanatory covariates is comparable to the volume of available data. Importantly, whilst we model Y based on our covariates, our primary objective is not prediction. Instead, we focus on identifying which covariates have a statistically significant influence on Y. In the context of a genetic study, this is akin to determining which genetic variations are provably in the pathway of a disease [4].

Our study will provide a comprehensive review of the literature on knockoffs, along with a detailed description of the methodology itself. This includes the statistical foundations underpinning the mechanism, particularly the proof of FDR control and the construction of knockoffs to ensure necessary exchangeability properties of irrelevant covariates. Finally, we conduct extensive simulations in both traditional and causal frameworks to evaluate the performance of the methodology. By varying model parameters, we simulate a range of real-world scenarios to assess robustness and applicability.

Contents

Chapter 1

Introduction

1.1 Problem statement

We will begin this report by outlining the problem variable selection attempts to solve. Suppose we have some target variable Y that we hope to understand. In a medical context, this may be a phenotype or some key health metric such as blood pressure, cholesterol level or the status or indicator of a disease such as Chrohn's disease or HIV [20] [21]. There are many potential explanatory covariates $X_1, X_2, ..., X_p$ but we are unsure which of them are significant in determining Y and the magnitude of this significance. In our medical example, we would like to establish, given an individual's entire genotype, which genetic variations are truly impactful for the phenotype we are considering. This is exactly the question that variable selection attempts to solve.

Let us formalise the mathematical foundations that will provide the basis of the report. We have that our target Y only depends on a small subset of $X = (X_1, X_2, ..., X_p)$. This subset, denoted $\mathcal{S}$, is referred to as the Markov blanket [26] for our particular variable selection problem. We consider a covariate j to be *null* if, given all the remaining covariate information, X_{-j}, covariate X_j tells us no new information about Y. Formally, we write

$$\text{j is null} \Leftrightarrow Y \perp\!\!\!\perp X_j | X_{-j} \tag{1.1}$$

Intuitively, we want to disregard as many of these null covariates as possible whilst keeping the non-nulls which we refer to as *truly relevant*. Formally, we want to find the smallest subset $\mathcal{S}$ such that

$$Y \perp\!\!\!\perp \{X_j\}_{j \in S^c} \mid \{X_j\}_{j \in S} \tag{1.2}$$

The problem lies in implementing a method of variable selection such that we select as many of the truly relevant covariates whilst at the same time, not obtaining too many false discoveries. We define a false discovery as a covariate that the model deems to be truly relevant but in reality is null. Let us further provide a definition of the False Discovery Rate and the power of a statistical test.

Definition 1 (False Discovery Rate (FDR))**.**

$$FDR := \mathbb{E}\left(\frac{\#False\ Discoveries}{\#Features\ Selected}\right) \tag{1.3}$$

In words, the false discovery rate is the expected proportion of false discoveries among the features selected.

Definition 2 (Power)**.**

$$Power := \mathbb{E}\left(\frac{\#Truly\ Relevant\ Features\ Selected}{\#Truly\ Relevant\ Features}\right) \tag{1.4}$$

In words, the power is the expected proportion of true covariates that are correctly selected.

There lies a trade-off between FDR and power. That is, the less strict our variable selection mechanism, the more likely we are to obtain all our truly relevant covariates and so maximise our power. However, in doing this, we will increase the number of false discoveries we make and raise FDR. The knockoff methodology provides a mechanism by which we obtain approximate FDR control whilst simultaneously obtaining high levels of power.

1.2 Knockoffs as a solution

Knockoffs, denoted $\tilde{X} = (\tilde{X}_1, \tilde{X}_2, ..., \tilde{X}_p)$, are artificial copies of $X = (X_1, X_2, ..., X_p)$, designed to serve as negative controls against the covariates and calibrate the selection procedure. Importantly, these knockoffs are generated without looking at the target Y, ensuring that they are all truly null by construction. In short, we augment X with $\tilde{X}$ and use $[X, \tilde{X}]$ to model Y. The model we choose is based on expert knowledge and the context of our selection problem as well as the the knockoff methodology we impose (as will be seen in 1). The model produces feature importance scores for all feature and knockoff variables, namely $Z = (Z_1, Z_2, ..., Z_p)$ and $\tilde{Z} = (\tilde{Z}_1, \tilde{Z}_2, ..., \tilde{Z}_p)$. As one may expect in the case of truly relevant covariates, their feature importance statistics are likely to be large whereas we would expect null covariates (and indeed all the knockoffs) to have feature importance statistics of smaller magnitude. These feature importance scores are combined into a single, adjusted score $W_j = f(Z_j, \tilde{Z}_j)$ which is then used to compare the relative importance of a covariate X_j with its corresponding knockoff $\tilde{X}_j$.

By an additional exchangeability property of null covariates imposed during knockoff construction, we ensure that the feature importance scores of null covariates are likewise exchangeable with their corresponding knockoffs. That is, one could swap the feature importance statistics of X_j and $\tilde{X}_j$ and this would not influence the value of Y. As will be discussed in following chapters, we see that W_j is constructed such that it is antisymmetric meaning that swapping Z_j and $\tilde{Z}_j$ works to flip its sign. Using this property, we are able to obtain an approximation of the false discovery rate, depending on $W = (W_1, W_2, ..., W_p)$, without knowledge of the ground truth, that is, which covariates are truly relevant and which are null. We can therefore manage calibrate our selection process, based on W, to (approximately) ensure a nominal level of false discovery, q, whilst maintaining a degree power.

1.3 Main objectives

In this paper, we provide a comprehensive review of the literature to date, highlighting key results and methodologies from major works on knockoffs [1][2][20]. To contextualize the methodology, we compare the Model-X and Fixed-X variations, highlighting their assumptions and applicability in different scenarios. We aim to provide a rigorous yet interpretable understanding of the knockoff framework by exploring its statistical foundations. In Chapter 2, We establish key theoretical results, formally defining knockoffs and proving the theorem for false discovery rate (FDR) control. Following this, in Chapter 3, we outline the construction process and provide a proof of the exchangeability property of knockoff statistics, which underpins the validity of the knockoff filter. In Chapter 4, we break down the mechanics of the filter and its role in variable selection, detailing how it identifies truly relevant covariates using knockoff statistics while controlling false discoveries. Finally, in Chapter 5, we validate the framework through simulations, assessing its performance in terms of FDR control and statistical power. We further evaluate its robustness in high-dimensional settings, under model misspecification and with finite volumes data in both traditional and causal frameworks.

1.4 Literature review

This section will discuss the key papers that explore the Knockoffs, highlighting their methodologies, findings and limitations.

1.4.1 Overview

In this review, we aim to provide a comprehensive and clear explanation of the key literature on knockoffs, summarizing the most important developments and insights up to this point. We will

consider two papers, namely *Controling the False Discovery Rate Via Knockoffs* [1] by Rina Foygel Barber and Emmanuel J. Candés and *Panning for gold: 'Model-X' knockoffs for high dimensional controlled variable selection* [2] by Candés et al. The former introduces the knockoff framework as a new means of performing variable selection based on fixed raw data whilst controlling the False Discovery rate (FDR) with a finite number of samples. The latter builds on the original paper by introducing a new alternative paradigm in cases where our covariate information is random and more spcefically, from some known distribution. It thereby establishes a *shift in the burden of knowledge* from the target variable to the covariates themselves. Although both papers describe a knockoff-based technique for variable selection for controlling FDR whilst also maintaining a level of power, the different assumptions underlying each paradigm make each better suited in distinct scenarios.

1.4.2 Methodologies: Fixed-X vs Model-X

The original knockoff paper [1] introduces and implements the Fixed-X paradigm whereas in *Panning for Gold* [2], Candés et al. present the new Model-X framework. In this section, we will provide a high level overview of each. First, suppose that we have a dataset made up of n rows of independent, identically distributed data. In this dataset, there are $p+1$ columns; we have our covariate data matrix, $\mathbf{X}$ which is made up of p columns, $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_p$ where each column corresponds to some covariate. Additionally, we have a final column $\mathbf{Y}$ corresponding to our target variable. Each row i therefore looks like $(\mathbf{X}_{i1}, \mathbf{X}_{i2}, ..., \mathbf{X}_{ip}, \mathbf{Y}_i)$. Now that this is established, let us briefly consider each approach.

1. **The Fixed-X approach**
   As the name suggests, within this framework, our covariate design matrix, $\mathbf{X}$ is fixed. One could think of this as a $n \times p$ matrix of real values from observed studies. From here, we choose to make an assumption about relationship between our covariates and our target variable. For example, in the original Barber and Candés (2015) paper, the assumption made is that our n observations obey a homoscedastic linear model (a linear model where variance of the errors remains constant across all observations):

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \quad \text{where} \quad \beta \in \mathbb{R}^p \text{ and } \varepsilon \sim \mathcal{N}(0, \sigma^2). \tag{1.5}$$

2. **The Model-X approach**
   The Model-X methodology does the exact opposite to the Fixed-X approach in that no assumption is made about the relationship between covariates $\mathbf{X}$ and the target variable $\mathbf{Y}$. Therefore, we do not immediately assume the conditional distribution of the target variable follows some parametric equation. Within this paradigm, our covariate data is not fixed and instead we choose to *model* $\mathbf{X}$. In other words, we assume that our covariates come from some known distribution, $P_{\mathbf{X}}$ and is therefore random (unlike in Fixed-X). Essentially, $\mathbf{X}$ now represents a set of randomly generated variables and one cannot observe the exact values of its entries in the way that one could within the Fixed-X paradigm. As aptly described in in the *Panning for Gold* paper, this paradigm represents a *shift in the burden of knowledge* from the target variable to the covariates. This method has particular use cases that may make it more applicable in certain scenarios than the Fixed-X approach.

Model preference

The assumptions underlying each paradigm make them more applicable to certain scenarios, with each lending itself better to specific contexts than the other due to their inverse assumptions. The Fixed-X approach may be more applicable in situations where we have recorded observational data but are uncertain about the underlying distribution of this data. Alternatively, we may choose to implement Model-X knockoffs in generative modelling or simulation studies, where data is generated according to some distribution. Additionally, we may choose to implement Model-X when we lack target variable data but have sufficient volumes of covariate data to make an approximation about the distribution they come from. That way, we can avoid being forced to make an ill-assumption about the relationship between covariates and target. Additionally, Model-X lends itself to cases of high-dimensionality, common in genetic studies whereas, as will be discussed in later chapters, the Fixed-X approach can only be truly implemented in cases where the number of observations exceeds

the number of covariates ($n > 2p$) and can be extended to the case where ($2p > n$) under further assumptions.

In short, the choice of paradigm asks us to make an assumption about either $\mathbf{X}$ or $\mathbf{Y}$. Clearly, if the covariate/target relationship was known to be linear, this is important information that should be used to implement Fixed-X knockoffs but in practice, this is not always realistic. The purpose of conditional modelling problems is to associate $\mathbf{Y}$, usually some key metric, with $\mathbf{X}$, a set of collected features [2]. Therefore, one would inherently expect that more is understood about these covariates since we use them as a basis to measure our key metric which would suggest that the Model-X approach is more widely applicable.

1.4.3 Key findings

We now proceed with our literature review by highlighting the key findings from each paper and discussing their implications.

Controlling the false discovery rate via knockoffs (2015) [1]

This paper introduces the knockoff filter, a novel method for controlling the FDR when performing variable selection under the Fixed-X approach. In essence, the knockoff filter works by constructing a synthetic variable or *knockoff*, $\tilde{\mathbf{X}}_j$, for each of the p original covariates that exactly mimics the correlation properties of that covariate $\mathbf{X}_j$ such that they can not be distinguished between without considering the target variable, $\mathbf{Y}$. Following this, an individual feature importance statistic is assigned to all original variables and their knockoffs. This statistic is calculated based on the strength of the relationship between that feature (original or knockoff) and the target variable $\mathbf{Y}$ according to the relational model that has been assumed.

In this paper, a lasso model is used with l_1-norm penalized regression to map the combined set of variables and knockoffs, $(\mathbf{X}, \tilde{\mathbf{X}})$, to target $\mathbf{Y}$. Coefficients of irrelevant features are shrunk to zero, and those associated with a relevant feature remain large [18]. Formally, the coefficient of some covariate $\mathbf{X}_j$ in a lasso model are calculated as follows.

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left(\frac{1}{2} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right)$$

The first term in the calculation is the squared residual sum which penalizes the inaccuracy of the linear model whilst the latter term is responsible for the penalization of the complexity of $\hat{\boldsymbol{\beta}}$. λ represents the trade-off between the minimisation of these two terms. The feature importance statistic of variable $\mathbf{X}_j$ is then calculated as $Z_j = \sup\left\{\lambda : \hat{\beta}_j(\lambda) \neq 0\right\}$ along with the feature importance statistic of its corresponding knockoff, $\tilde{\mathbf{X}}_j$, which we define as $\tilde{Z}_j$. This statistic indicates the first time the variable is included in the model (it has a non-zero coefficient) since the value of λ shrinks as the model progresses and the regularization becomes weaker. If Z_j is large, this implies the covariate or knockoff we are currently considering enters the model early in the progression whilst the penalization term is still large. In this case, the variable is likely to be a truly relevant variable. Alternatively, if Z_j is small, this implies the variable we are considering enters the model late in the progression and the coefficient only becomes non-zero when the regularization is weak, suggesting the variable is likely to be null.

The feature importance statistics of covariates and their corresponding knockoff variables are then together used to construct $W_j = f(Z_j, \tilde{Z}_j)$, the knockoff statistic that is used to assess whether that original covariate $\mathbf{X}_j$ is indeed relevant. A threshold value, T, is then constructed based on $\mathbf{W} = (W_1, ..., W_p)$ in such a way that the FDR is controlled whereby large positive values of W_j that surpass threshthold T imply that X_j is truly relevant. Proof of the major theorems for controlled FDR are shown in Chapter 2.

Following discussions of alternative methods of variable selection, the knockoff filter is applied to help detect mutations in HIV associated with drug resistance with the overall conclusion of this experiment suggesting that the knockoff filter performed slightly better than another common alternative variable selection approach (Benjamini-Hochberg Hypothesis testing) but that results were not consistent across different drugs. The Fixed-X paradigm is further applied in simulations to verify that

the method is robust to non-Gaussian noise where the empirical noise distribution is heavily tailed suggesting that a normal distribution is not necessarily appropriate. These tests proved successful with each variable selection method achieving similar power and FDR results.

This paper therefore presents a new and improved method of variable selection with controlled False Discovery Rate (FDR). Barber and Candés discuss potential extensions of their work, considering how the statistical power of FDR-controlled tests could be maximized. They also highlight the problem of high-dimensionality within the Fixed-X approach. To address these challenges, they suggest that continued research will focus on developing knockoff methodology that can handle high-dimensional cases effectively, ensuring both control of FDR and maintaining statistical power.

Panning for gold: 'Model-X' knockoffs for high dimensional controlled variable selection (2018) [2]

The major finding of this paper is the introduction of a novel framework for knockoffs, where the assumptions are modified to provide a more general solution to the variable selection problem where no specific functional form for the underlying data distribution is immediately assumed.

The Model-X paradigm builds upon the foundational theory of FDR (type-1 error) control established in the original Barber and Candès paper for Fixed-X knockoffs. While extending research by imposing assumptions on $\mathbf{X}$ rather than the covariate-target relationship, the Model-X paradigm retains the core theoretical FDR control results without requiring new proofs of these theorems. Instead, its focus shifts toward enhancing the power of these statistical tests.

The new approach presented in this paper is valuable for several reasons. First, the Model-X framework is applicable in high-dimensional settings, where the number of covariates exceeds the amount of available data (something that was not possible with the original Fixed-X approach proposed by Barber and Candés). As previously mentioned, this can be useful in genetic studies where the number of covariates (genes) being studied is vast. Additionally, the assumptions underpinning the Model-X framework ensure its applicability beyond generalized linear models. In essence, this allows us to choose a model that maximizes predictive performance without the need to impose any assumptions, such as a linear relationship. Instead, we can leverage more flexible machine learning, non-parametric models that better capture non-linear relationships in the data. Whilst the knockoff framework is agnostic to model choice in that we should be able to control FDR regardless, the power of the procedure hinges on the model's ability to accurately capture the relationship between target and covariates (and hence accurately assign feature importance statistics) [17]. For example, feature importance statistics can be constructed using random forest or gradient boosting models, such as XGBoost or AdaBoost [25] [11] [16] [20], or from Bayesian variable selection methods [22] as well as generalized additive models [23].

The paper considers the construction of Model-X knockoffs and the problems associated with treating our covariates $\mathbf{X}$ as random. In both frameworks, the construction of the knockoffs $\tilde{\mathbf{X}}$ is in such a way that the correlation structure between covariates in $\mathbf{X}$ is mirrored in $\tilde{\mathbf{X}}$. However, rather than simply performing the necessary calculations on the raw sampled values (as in Fixed-X), the construction must ensure that the correlations between our covariates and their knockoffs are preserved at the distributional level. Both exact and approximate constructions are considered for Gaussian and non-Gaussian data.

The authors suggest potential areas for further research such as computing multiple knockoffs per covariate feature and use the rank with which the original variable enters the model compared to its knockoffs as a feature importance statistic [7]. This could potentially be an improvement (in terms of power) from the "one-bit" comparison (considering which enters the model first out of $\mathbf{X}_j$ and its single knockoff counterpart, $\tilde{\mathbf{X}}_j$) as by comparing $\mathbf{X}_j$ against a richer set of knockoffs, say $\tilde{\mathbf{X}}_j^1, \tilde{\mathbf{X}}_j^2, \tilde{\mathbf{X}}_j^3$, more accurate information about the importance of $\mathbf{X}_j$ could be captured. Similarly, since $\mathbf{X}$ is considered random, we can perform repeated trials whereby we construct several knockoff matrices, each a stochastic realization based on the marginal distribution of $\mathbf{X}$. Each matrix will yield slightly different selection whilst ensuring FDR control allowing for the computation of confidence intervals and mean results on our selection.

1.4.4 Using knockoffs for controlled predictive biomarker identification

In this section, we discuss the application of knockoff-based controlled variable selection in biomarker identification, as presented by Sechidis et al. (2020) [20]. Consider a randomized controlled trial (RCT) designed to determine the optimal genetic makeup of a patient that would lead to an impact in the target variable $\mathbf{Y}$, some key health metric. In this example, covariate variable $\mathbf{X}_j$ could be an indicator for the presence of gene j in the patient. In this context, the emphasis is placed on selecting predictive variables rather than prognostic ones. A prognostic variable refers to one that, independent of treatment, has an impact on the target $\mathbf{Y}$. Alternatively, a predictive variable refers to one that has an impact on the target $\mathbf{Y}$ as a result of taking the treatment. This can be formally described as follows.

$$\mathbb{E}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}, T = t] = h(\mathbf{x}) + g(\mathbf{x})t$$

Where

- $T = 1$ if the subject is in the treatment group and $T = 0$ if the subject is in the control.
- $h(x)$ is the component of the expectation that depends on the prognostic covariates.
- $g(x)$ is the component that captures the interaction between the predictive covariates and the treatment T, showing how the treatment affects the outcome.

Suppose a RCT is being carried out to investigate the effect of a new drug on blood pressure. Test subjects will each be randomly assigned to the treatment group (those that receive a treatment) or the control group (those that don't receive the treatment). In this context, a prognostic variable would be a genetic marker that is associated with a change in blood pressure independent of whether the subject receives the drug or not e.g. weight. A predictive variable would be a genetic marker that leads to a significant change in blood pressure when the subject has received treatment e.g. the interaction of a particular gene variant with the treatment, resulting in a considerable change in blood pressure.

In this context, the aim would be to identify those patients who are more likely to benefit from taking the drug which is analogous to identifying predictive biomarkers as opposed to just identifying prognostic biomarkers that impact blood pressure regardless of treatment. Hence, the random aspect of the trial is crucial; Suppose we had groups A and B and we know that group A is made up of those with optimal blood pressure and group B is made up of those with high blood pressure. Suppose we were to assign all those in group A to the treatment group and all those in group B to the control group. In this case, while we may be able to capture the prognostic biomarkers (those that signal low blood pressure regardless of treatment), we would struggle to capture the predictive ones as we are not observing how genetic markers influence the response to treatment. By randomly assigning subjects in group A and B to either the treatment or control group, we would be able to assess how different genetic markers influence the treatment's effect on blood pressure. If we observe a change in blood pressure for some subject, randomization eliminates the possibility that this change is solely due to the prognostic variables. This paper presents the knockoff methodology applied to predictive variable selection; establishing which biomarkers are responsible for impacting the target $\mathbf{Y}$ as a result of the treatment. This highlights an important application of variable selection in understanding treatment effects in the causal framework.

1.5 Notation

In this section, we formally define some notation that we use throughout this paper. Define $[q] := \{1, \dots, q\}$. We define a Gram matrix [14] as follows. Consider an $n \times p$ matrix $\mathbf{X}$ over $\mathbb{R}$,

$$\mathbf{X} = \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,p} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n,1} & X_{n,2} & \cdots & X_{n,p} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 & \cdots & \mathbf{X}_p \end{bmatrix}$$

The $p \times p$ matrix $\mathbf{X^T X}$ is self-adjoint:

$$\begin{bmatrix} \mathbf{X}_1^T\mathbf{X}_1 & \mathbf{X}_1^T\mathbf{X}_2 & \cdots & \mathbf{X}_1^T\mathbf{X}_p \\ \mathbf{X}_2^T\mathbf{X}_1 & \mathbf{X}_2^T\mathbf{X}_2 & \cdots & \mathbf{X}_2^T\mathbf{X}_p \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}_p^T\mathbf{X}_1 & \mathbf{X}_p^T\mathbf{X}_2 & \cdots & \mathbf{X}_p^T\mathbf{X}_p \end{bmatrix} = \mathbf{\Sigma}$$

$\mathbf{\Sigma}$ is the gram matrix of X. we define diag($\mathbf{s}$) as

$$\text{diag}(\mathbf{s}) = \begin{bmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_m \end{bmatrix}$$

Let $a \vee b$ denote the maximum of a and b for $a, b \in \mathbb{R}$. For random variables X, Y, we have that $Y \perp\!\!\!\perp X$ indicates that Y and X are independent. The equality, $X \stackrel{d}{=} Y$, denotes equality in distribution. The supremum of a set S is the least upper bound. That is, $\sup S =$ the smallest number such that $\sup S \geq s$ for all $s \in S$, and for any $x < \sup S$, there exists $s \in S$ such that $s > x$.

Chapter 2

Statistical foundations

2.1 Key mathematical foundations

In this section, a formal definition of knockoff variables is provided and the process of their construction is described. As in Chapter 1 we suppose that we have a dataset made up of n rows of independent, identically distributed data. In this dataset, there are $p+1$ columns; we have our covariate data matrix, $\mathbf{X}$ which has p columns $\{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_p\}$ which each correspond to some covariate. Additionally, we have a final column $\mathbf{Y}$ corresponding to our target variable. In each row i, we therefore have $(\mathbf{X}_{i1}, \mathbf{X}_{i2}, ..., \mathbf{X}_{ip}, \mathbf{Y}_i)$.

2.1.1 What is a knockoff?

First let us define the concept of a knockoff variable. A knockoff variable is an artificial feature which is designed to mimic the correlation structure within the original features, acting as a negative control to establish whether the original variable is relevant within a particular model by comparing its relevance with its corresponding knockoff.

Given our design matrix $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$, where for each $j \in [p]$ the vector $\mathbf{X}_j$ in $\mathbb{R}^n$ corresponds to the j-th covariate, we let $\tilde{\mathbf{X}}_j$ in $\mathbb{R}^n$ denote the associated knock-off variable. The knockoff variables $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_p)$ are constructed such that they satisfy the following two properties:

1. **Independence from the target:** The knockoff variables $\tilde{\mathbf{X}}$ must be conditionally independent of the target variable Y given $\mathbf{X}$ i.e.
$$\tilde{\mathbf{X}} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{X}.$$
This is an important restriction as by creating knockoffs that are not influenced by $\mathbf{Y}$, one ensures that the knockoffs are neutral and can be fairly compared to the original variables [20]. Due to this independence, we can use these knockoffs to distinguish which of the original variables are relevant or null. We can ensure this property if the knockoff variables are constructed without considering $\mathbf{Y}$. This is equivalent to saying that the knockoffs are all null. Recall, a null variable is conditionally independent of the target given all other covariate information i.e. it provides no additional information about the target conditioned on the fact we have all our other covariates.

2. **Exchangeability Property:** The original covariates $\mathbf{X}_j$ and their corresponding knockoff variables $\tilde{\mathbf{X}}_j$ must be exchangeable under the null hypothesis. Specifically, for each index $j \in \{1, \ldots, p\}$ corresponding to a null covariate, the pair $(\mathbf{X}_j, \tilde{\mathbf{X}}_j)$ must be exchangeable, meaning that swapping $\mathbf{X}_j$ with $\tilde{\mathbf{X}}_j$ does not affect the joint distribution:
$$(\mathbf{X}_1, \ldots, \mathbf{X}_j, \ldots, \mathbf{X}_p, \tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_j, \ldots, \tilde{\mathbf{X}}_p) \overset{d}{=} (\mathbf{X}_1, \ldots, \tilde{\mathbf{X}}_j, \ldots, \mathbf{X}_p, \tilde{\mathbf{X}}_1, \ldots, \mathbf{X}_j, \ldots, \tilde{\mathbf{X}}_p),$$

2.1.2 The knockoff property

In this section, we will state the knockoff property which ensures that the knockoff variables are constructed such that they mirror the relationships between the original variables. Specifically, We

want to create knockoffs such that we ensure that the covariance structure of variables is exactly copied by the knockoffs

1. The correlation between distinct knockoffs $\tilde{\mathbf{X}}_j$ and $\tilde{\mathbf{X}}_k$ (for $j \neq k$) is the same as the correlation between $\mathbf{X}_j$ and $\mathbf{X}_k$.

2. The correlation between $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_k$ (for $j \neq k$) is the same as the correlation between the original variables $\mathbf{X}_j$ and $\mathbf{X}_k$.

Together, these conditions ensure that the joint distribution of $(\mathbf{X}, \tilde{\mathbf{X}})$ is invariant under any swapping of a null variable (i.e. a column) in $\mathbf{X}$ with its knockoff counterpart in $\tilde{\mathbf{X}}$ [1] [20]. In Chapter 3, we will examine why the invariance of the joint distribution under a swapping is important for the knockoff methodology to function correctly.

Recall the definition of the Gram Matrix of $\mathbf{X}$, which we define as $\mathbf{\Sigma}$ where

$$\mathbf{\Sigma} = \mathbf{X^T X} = \begin{bmatrix} \mathbf{X}_1^T\mathbf{X}_1 & \mathbf{X}_1^T\mathbf{X}_2 & \cdots & \mathbf{X}_1^T\mathbf{X}_p \\ \mathbf{X}_2^T\mathbf{X}_1 & \mathbf{X}_2^T\mathbf{X}_2 & \cdots & \mathbf{X}_2^T\mathbf{X}_p \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}_p^T\mathbf{X}_1 & \mathbf{X}_p^T\mathbf{X}_2 & \cdots & \mathbf{X}_p^T\mathbf{X}_p \end{bmatrix}$$

In the case of the first condition which refers to the correlation structure between 2 variables being mimicked by their corresponding knockoffs, the gram matrix is used as a condition for constructing the knockoffs in such a way that they have exactly the same pairwise linear dependencies (correlations) between them as their original counterparts. Essentially, the correlation structure between the original features is preserved in the knockoffs if one ensures that they have equivalent gram matrices. Formally, this condition can be expressed as

$$\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} = \mathbf{X}^\top\mathbf{X} = \mathbf{\Sigma} \tag{2.1}$$

The second condition refers to the correlation between some original variable, $\mathbf{X}_j$, and some distinct knockoff (i.e. not corresponding to that original variable), $\tilde{\mathbf{X}}_k$, being equivalent to the correlation between $\mathbf{X}_j$ and $\mathbf{X}_k$. This is done by ensuring the following condition holds:

$$\mathbf{X}^\top\tilde{\mathbf{X}} = \mathbf{\Sigma} - \text{diag}(\mathbf{s}) = \mathbf{X}^\top\mathbf{X} - \text{diag}(\mathbf{s}) \tag{2.2}$$

Here, $\mathbf{s} \in \mathbb{R}^p$ is some p-dimensional vector that we can choose (we discuss later in Chapter 3 how this choice is made). If this condition holds, it would imply that for all $i \neq j$,

$$(\mathbf{X}^\top\tilde{\mathbf{X}})_{ij} = (\mathbf{X}^\top\mathbf{X})_{ij} \tag{2.3}$$

since the off-diagonal entries of $\text{diag}(\mathbf{s})$ are all zero. Therefore, condition 2 is ensured. However, one can recognize an outcome with this method of construction in that the following must therefore hold:

$$\mathbf{\Sigma}_{ii} = (\mathbf{X}^\top\mathbf{X})_{ii} = (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})_{ii} = 1 \tag{2.4}$$

Which implies the following.

$$(\mathbf{X}^\top\tilde{\mathbf{X}})_{ii} = \mathbf{\Sigma}_{ii} - s_i = 1 - s_i \tag{2.5}$$

Formally, the diagonal elements ($i = j$) are allowed to differ by an s_i of our choosing. In other words, the covariance of covariates and their corresponding knockoffs are not fixed by conditions 1 and 2.

Combining our results (2.1) and (2.2), we can therefore say that $\tilde{\mathbf{X}}$ obeys the following:

$$\begin{bmatrix}\mathbf{X}\ \tilde{\mathbf{X}}\end{bmatrix}^\top \begin{bmatrix}\mathbf{X}\ \tilde{\mathbf{X}}\end{bmatrix} = \begin{bmatrix} \mathbf{X}^\top\mathbf{X} & \mathbf{X}^\top\tilde{\mathbf{X}} \\ \tilde{\mathbf{X}}^\top\mathbf{X} & \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} \end{bmatrix} = \begin{bmatrix} \mathbf{\Sigma} & \mathbf{\Sigma} - diag(\mathbf{s}) \\ \mathbf{\Sigma} - diag(\mathbf{s}) & \mathbf{\Sigma} \end{bmatrix} := \mathbf{G} \tag{2.6}$$

2.1.3 Proof of FDR control

As aforementioned, our main goal is to control the False Discovery Rate, the expected proportion of covariates considered relevant that are in fact not relevant in reality. Informally, FDR is defined as follows:

$$\text{FDR} = \mathbb{E}\left[\frac{\text{False Positives}}{\text{False Positives} + \text{True Positives}}\right]$$

Recall, variable $\mathbf{X}_j$ is truly null if it has no measurable impact on the target variable $\mathbf{Y}$. In a lasso model, this is analogous to saying that the variable's coefficient, β_j, is zero. Our variable selection process will consider the variable $\mathbf{X}_j$ relevant if its associated knockoff statistic (which we discuss in Chapter 4), W_j, surpasses a data-dependent threshold T.

Specificity vs Sensitivity: A trade-off

Initially, it may be an appropriate question to ask why any level of false discovery is allowed at all. Suppose the threshold was set very high in an attempt to filter out truly null variables that still had relatively high associated knockoff statistics. In this case, one would likely exclude some genuinely relevant variables whose knockoff statistics are slightly lower than others, reducing the power of the experiment. Conversely, if the threshold was set very low, all variables that achieve any level of significance would be considered relevant. Although this increases the likelihood of capturing all truly relevant variables, increasing power, it also increases the chance of including more null variables, increasing FDR, as the lower threshold makes it easier for variables to meet the criterion. Here lies our trade-off between specificity (correctly identifying null variables) and sensitivity (correctly identifying relevant variables).

In the knock-off methodology, a data-dependent threshold can be constructed in such a way that the false discovery rate is controlled. This starts by setting an appropriate value for an acceptable level of FDR, q. For our variable $\mathbf{X}_j$, the associated statistic W_j is constructed in such a way that if the variable is truly null, W_j will be distributed symmetrically around zero. This means that, for null $\mathbf{X}_j$, the occurrence of a large positive value of W_j is as probable as the occurrence of large negative values of W_j. Essentially, for truly null variables, one would expect to see a similar number of large positive test statistics as large negative test statistics by the nature of their construction (which is outlined in Chapter 3). We can use this information to approximate the FDR and therefore set a data-dependent threshold that ensures this approximation of the false discovery rate is below our acceptable level of FDR, q. Formally we can define FDR as

$$\text{FDR} = \mathbb{E}\left[\frac{|\{j : \beta_j = 0 \text{ and } W_j \geq T\}|}{|\{j : W_j \geq T\}|}\right]$$

where the threshold T is defined as:

$$T = \min\left\{t \in \mathcal{W} : \frac{1 + |\{j : W_j \leq -t\}|}{|\{j : W_j \geq t\}| \vee 1} \leq q\right\} \tag{2.7}$$

and $\mathcal{W} = \{|W_1|, |W_2|, \ldots, |W_p|\}$.

The main result of the original Barber and Candés paper is that the fixed-X knock-off procedure controls a quantity which dominates the false discovery rate.

Theorem 1 ([1]). *For any $q \in [0, 1]$, the knockoff method satisfies*

$$\mathbb{E}\left[\frac{|\{j : \beta_j = 0 \text{ and } W_j \geq T\}|}{|\{j : W_j \geq T\}| + q^{-1}}\right] \leq \mathbb{E}\left[\frac{|\{j : \beta_j = 0 \text{ and } W_j \geq T\}|}{|\{j : W_j \geq T\}| \vee 1}\right] \leq q,$$

where the expectation is taken over the Gaussian noise z in the model (1.1), while treating $\mathbf{X}$ and $\tilde{\mathbf{X}}$ as fixed.

Theorem 2 ([1]). *For any $q \in [0, 1]$, the knockoff method satisfies*

$$\mathbb{E}\left[\frac{|\{j : \beta_j = 0 \text{ and } W_j \geq T\}|}{|\{j : W_j \geq T\}| \vee 1}\right] \leq q,$$

where the expectation is taken over the Gaussian noise z in the model (1.1), while treating $\mathbf{X}$ and $\tilde{\mathbf{X}}$ as fixed.

To prove Theorem 2, we will the following Lemmas and the Optional Stopping Theorem.

Lemma 3. *Suppose that we have m independent random variables $Z_1, \dots, Z_m$ taking values in $\{-1, 1\}$ with $P(Z_j = \omega) = 1/2$ for $\omega \in \{-1, 1\}$ and $j = 1, \dots, m$. For each $k \in \{0, \dots, m-1\}$ and $\omega \in \{-1, +1\}$, we let*

$$V_k(\omega) := \frac{m - k + \omega \sum_{j=k+1}^{m} Z_j}{2}.$$

Let us also define a filtration $(\mathcal{F}_\ell)_{\ell=0}^m$ by letting $\mathcal{F}_\ell$ denote the smallest σ-algebra such that the random variables $\{V_0(-1), V_0(1)\} \cup \{Z_j\}_{j \le \ell}$ are $\mathcal{F}_\ell$-measurable. Then, the process $(M_k)_{k=0}^{m-1}$ defined by

$$M_k := \frac{V_k(1)}{1 + V_k(-1)},$$

for $k \in \{0, \dots, m-1\}$ is a martingale with respect to $(\mathcal{F}_\ell)_{\ell=0}^m$.

Proof of Lemma 3. We first establish that $\{V_\ell(-1), V_\ell(+1)\}_{\ell \le k}$ are $\mathcal{F}_k$-measurable for each $k \in \{0, 1, \dots, m-1\}$. For $k = 0$, this is immediate from the definition of $\mathcal{F}_0$. For $k \in \{1, \dots, m\}$ and $\omega \in \{-1, 1\}$, we have

$$\begin{aligned} V_k(\omega) &= \frac{1}{2}\left(m - k + \omega \sum_{j=k+1}^{m} Z_j\right) \\ &= \frac{1}{2}\left(m - k + 1 + \omega \sum_{j=k}^{m} Z_j\right) - \frac{1}{2}(1 + \omega Z_k) \\ &= V_{k-1}(\omega) - \frac{1}{2}(1 + \omega Z_k). \end{aligned}$$

Hence, given that Z_k is $\mathcal{F}_k$-measurable, if $\{V_{k-1}(-1), V_{k-1}(+1)\}$ are $\mathcal{F}_{k-1}$-measurable, then $\{V_k(-1), V_k(+1)\}$ are $\mathcal{F}_k$-measurable. Thus, by induction, $\{V_\ell(-1), V_\ell(+1)\}_{\ell \le k}$ are $\mathcal{F}_k$-measurable for each $k \in \{0, 1, \dots, m-1\}$. It follows from the definition of M_k that it is also $\mathcal{F}_k$-measurable.

Furthermore, for each $k \in \{1, \dots, m\}$, we have

$$\mathbb{P}(Z_k = \omega \mid \mathcal{F}_{k-1}) = \mathbb{P}(Z_k = \omega \mid \mathcal{F}_{k-1}, V_{k-1}(\omega)) = \frac{V_{k-1}(\omega)}{V_{k-1}(-1) + V_{k-1}(1)},$$

since $V_{k-1}(\omega) = |\{j \in \{k, \dots, m\} : Z_j = \omega\}|$ and $Z_k, \dots, Z_m$ are conditionally exchangeable. Thus, for each $k \in \{1, \dots, m-1\}$, we have

$$\begin{aligned} \mathbb{E}(M_k \mid \mathcal{F}_{k-1}) &= \mathbb{E}\left(\frac{V_k(1)}{1 + V_k(-1)} \mid \mathcal{F}_{k-1}\right) \\ &= \mathbb{E}\left(\frac{V_{k-1}(1) - \frac{1}{2}(1 + Z_k)}{1 + V_{k-1}(-1) - \frac{1}{2}(1 - Z_k)} \mid \mathcal{F}_{k-1}\right) \\ &= \mathbb{P}(Z_k = -1 \mid \mathcal{F}_{k-1}) \frac{V_{k-1}(1)}{V_{k-1}(-1)} + \mathbb{P}(Z_k = 1 \mid \mathcal{F}_{k-1}) \frac{V_{k-1}(1) - 1}{1 + V_{k-1}(-1)} \\ &= \frac{V_{k-1}(-1)}{V_{k-1}(-1) + V_{k-1}(1)} \frac{V_{k-1}(1)}{V_{k-1}(-1)} + \frac{V_{k-1}(1)}{V_{k-1}(-1) + V_{k-1}(1)} \frac{V_{k-1}(1) - 1}{1 + V_{k-1}(-1)} \\ &= \frac{1 + V_{k-1}(-1)}{V_{k-1}(-1) + V_{k-1}(1)} \frac{V_{k-1}(1)}{1 + V_{k-1}(-1)} + \frac{V_{k-1}(1) - 1}{V_{k-1}(-1) + V_{k-1}(1)} \frac{V_{k-1}(1)}{1 + V_{k-1}(-1)} \\ &= \frac{V_{k-1}(1)}{1 + V_{k-1}(-1)} \\ &= M_{k-1}. \end{aligned}$$

as required. □

Lemma 4. *With the notation of Lemma 1, we have* $\mathbb{E}(M_0) = 1 - 2^{-m}$.

Proof of Lemma 4. Notice that $V_0(\omega) = |\{j \in \{1, \dots, m\} : Z_j = \omega\}|$ for each $\omega \in \{-1, 1\}$. Thus, $V_0(1)$ has a Binomial distribution, and for each $k \in \{0, 1, \dots, m\}$, we have

$$\mathbb{P}\{V_0(\omega) = k\} = 2^{-m} \frac{m!}{k!(m-k)!},$$

and $V_0(-1) = m - V_0(1)$. Hence, we have

$$\mathbb{E}(M_0) = \mathbb{E}\left(\frac{V_0(1)}{1 + m - V_0(1)}\right) = \left(\frac{1}{2}\right)^m \sum_{k=0}^{m} \frac{m!}{k!(m-k)!} \frac{k}{1+m-k}.$$

Simplifying the sum, we get

$$\mathbb{E}(M_0) = \left(\frac{1}{2}\right)^m \sum_{k=1}^{m} \frac{m!}{(k-1)!(m-\{k-1\})!} = 1 - 2^{-m}.$$

□

Recall the definition of a Stopping time.

Definition 3 ([15])**.** *τ is called a Stopping Time with respect to the filtration $(\mathcal{F}_t)_{t\geq 0}$, if the following condition holds:*

$$\{\tau \leq t\} \in \mathcal{F}_t \quad \textit{for all} \quad t \in T$$

Recall the Optional Stopping Theorem for supermartingales.

Theorem 5 (Optional Stopping Theorem for Supermartingales [6])**.** *Let $(M_t)_{t\geq 0}$ be a supermartingale with respect to a filtration $(\mathcal{F}_t)_{t\geq 0}$, and let T be a Stopping time with respect to $(\mathcal{F}_t)_{t\geq 0}$. Assume the following conditions hold:*

1. ***Bounded Stopping Time**: T is almost surely finite, i.e., $P(T < \infty) = 1$.*
2. ***Bounded Increments**: There exists a constant C such that $|M_{t+1} - M_t| \leq C$ for all t.*
3. ***Finite Expectation**: $\mathbb{E}[|M_T|] < \infty$.*

Then,

$$\mathbb{E}[M_T] \leq \mathbb{E}[M_0].$$

Proof of Theorem 2. Given $t \in \mathbb{R}$ let's define $\mathrm{FDP}(t)$ by

$$\mathrm{FDP}(t) := \left[\frac{|\{j : \beta_j = 0 \text{ and } W_j \geq t\}|}{|\{j : W_j \geq t\}| \vee 1}\right],$$

so that the false discovery proportion of the knock-off procedure is $\mathrm{FDP}(T)$. Notice that for all $t \in \mathbb{R}$ we have $|\{j : \beta_j = 0 \text{ and } W_j \leq -t\}| \leq |\{j : W_j \leq -t\}|$, so that

$$1 \leq \frac{1 + |\{j : W_j \leq -t\}|}{1 + |\{j : \beta_j = 0 \text{ and } W_j \leq -t\}|}.$$

Therefore, for any $t \in \mathbb{R}$ we have

$$\begin{aligned}
\mathrm{FDP}(t) &= \frac{|\{j : \beta_j = 0 \text{ and } W_j \geq t\}|}{|\{j : W_j \geq t\}| \vee 1} && (2.8)\\
&\leq \frac{|\{j : \beta_j = 0 \text{ and } W_j \geq t\}|}{|\{j : W_j \geq t\}| \vee 1} \frac{1 + |\{j : W_j \leq -t\}|}{1 + |\{j : \beta_j = 0 \text{ and } W_j \leq -t\}|} && (2.9)\\
&= \frac{1 + |\{j : W_j \leq -t\}|}{|\{j : W_j \geq t\}| \vee 1} \cdot \frac{|\{j : \beta_j = 0 \text{ and } W_j \geq t\}|}{1 + |\{j : \beta_j = 0 \text{ and } W_j \leq -t\}|}. && (2.10)
\end{aligned}$$

Hence, by the the definition of T, substituting into 2.8, it must be that

$$\text{FDP(T)} = \left[\frac{|\{j : \beta_j = 0 \text{ and } W_j \geq T\}|}{|\{j : W_j \geq T\}| \vee 1}\right] \leq q \cdot \frac{|\{j : \beta_j = 0 \text{ and } W_j \geq T\}|}{1 + |\{j : \beta_j = 0 \text{ and } W_j \leq -T\}|} \tag{2.11}$$

Let $W_{(1)}, \ldots, W_{(m)}$ be the the knockoff statistics for the m null covariates $W_1, \ldots, W_m$, reordered in order of magnitude so that

$$|W_{(1)}| \leq \ldots \leq |W_{(m)}|.$$

In Lemma 3 , the notation introduces a collection of random variables, $Z_1, \ldots, Z_m$ corresponding the signs of the $W_{(1)}, \ldots, W_{(m)}$, respectively. Furthermore, for each $k \in \{0, \ldots, m-1\}$ and $\omega \in \{-1, +1\}$, we let

$$V_k(\omega) := \frac{m - k + \omega \sum_{j=k+1}^{m} Z_j}{2},$$

as in Lemma 3. We shall define a filtration $(\mathcal{G}_\ell)_{\ell=0}^m$ by letting $\mathcal{G}_\ell$ denote the smallest σ-algebra such that the random variables $\{|W_{(\ell)}|\}_{\ell=1}^m \cup \{V_0(-1), V_0(1)\} \cup \{Z_j\}_{j \leq \ell}$ are $\mathcal{G}_\ell$-measurable. By Lemma 3, applied conditionally on $\{|W_{(\ell)}|\}_{\ell=1}^m$, the process $(M_k)_{k=0}^{m-1}$ defined by

$$M_k := \frac{V_k(1)}{1 + V_k(-1)},$$

for $k \in \{0, \ldots, m-1\}$ is a martingale with respect to $(\mathcal{G}_\ell)_{\ell=0}^m$. Next, let $\hat{k}$ be the random element of $\{0, ..., m-1\}$ which corresponds to the index of the next smallest null j for which $|W_j| \geq T$, where T is defined in (2.7). Thus, we have

$$\frac{|\{j : \beta_j = 0 \text{ and } W_j \geq T\}|}{1 + |\{j : \beta_j = 0 \text{ and } W_j \leq -T\}|} = \frac{V_{\hat{k}}(1)}{1 + V_{\hat{k}}(-1)} = M_{\hat{k}},$$

Observe that $\hat{k}$ is a stopping time with respect to $(\mathcal{G}_\ell)_{\ell=0}^m$ (as defined in Lemma 3) since $\{\hat{k} \leq k\} \in \mathcal{G}_k$. Clearly, $\hat{k}$ is almost surely finite since m is finite. We have that $|M_{k+1} - M_k| \leq C$ for all $k \in \{0, ...m, m-1\}$ where $C = 3m$ (see Appendix B for how C is derived). Finally, the notion that $\mathbb{E}[|M_{\hat{k}}|] < \infty$ follows from the fact that since $V_k(\omega)$ is finite and thus so is $\mathbb{E}(V_k(\omega))$. Therefore, by the Optional Stopping Theorem for Supermartingales and Lemma 4, we have that

$$\mathbb{E}[M_{\hat{k}}] \leq \mathbb{E}[M_0] = 1 - 2^{-m} \leq 1$$

Therefore, we have that

$$FDR = \mathbb{E}[FDP] \leq q \cdot \mathbb{E}[\frac{|\{j : \beta_j = 0 \text{ and } W_j \geq T\}|}{1 + |\{j : \beta_j = 0 \text{ and } W_j \leq -T\}|}] \leq q \cdot 1 = q$$

□

Now, one can use the proof of the Theorem 2 to prove Theorem 1 simply.

Proof of Theorem 1. Note, since $q \in (0, 1]$, it must be that $q^{-1} \geq 1$. Therefore, $|\{j : W_j \geq T\}| + q^{-1} \geq |\{j : W_j \geq T\}| \vee 1$. Therefore, one must have the following result.

$$\left[\frac{|\{j : \beta_j = 0 \text{ and } W_j \geq T\}|}{|\{j : W_j \geq T\} + q^{-1}|}\right] \leq \left[\frac{|\{j : \beta_j = 0 \text{ and } W_j \geq T\}|}{|\{j : W_j \geq T\}| \vee 1}\right]$$

Now, by the monotonicity of expectations, one is left with the following result (where the second inequality comes from Theorem 2).

$$\mathbb{E}\left[\frac{|\{j : \beta_j = 0 \text{ and } W_j \geq T\}|}{|\{j : W_j \geq T\}| + q^{-1}}\right] \leq \mathbb{E}\left[\frac{|\{j : \beta_j = 0 \text{ and } W_j \geq T\}|}{|\{j : W_j \geq T\}| \vee 1}\right] \leq q$$

□

Chapter 3

Construction of knockoffs

3.1 Construction of knockoffs

In this section, we will outline the process of construction in both the fixed-X and Model-X frameworks

Fixed-X construction

As described in Chapter 1, within this paradigm, we consider $\mathbf{X}$ as a matrix made up of observed data. Therefore, we require that the knockoffs are created in such a way that that the obey the aforementioned exchangeability properties with the exact sample of data we have, $\mathbf{X}$ [3]. As detailed in the original Barber and Candés paper, we will describe the process for constructing knockoff variables $\tilde{\mathbf{X}}$ in this way. We will consider the general case where $n \geq 2p$ although the original paper provides an extension of the method for when $p < n < 2p$.

As before, we have that it must be that $\tilde{\mathbf{X}}$ is constructed such that knockoff property given in section 2.1.2 is guaranteed. According to the original knockoff paper, for $n \geq 2p$, such $\tilde{\mathbf{X}}$ exists and is the following form:

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

where $\tilde{\mathbf{U}} \in \mathbb{R}^{n\times p}$ is an orthonormal matrix such that $\tilde{\mathbf{U}}^\top\mathbf{X} = \mathbf{0}, \mathbf{C} \in \mathbb{R}^{p\times p}$ is a parameter matrix such that $\mathbf{C}^\top\mathbf{C} = 2\operatorname{diag}(\mathbf{s}) - \operatorname{diag}(\mathbf{s})\boldsymbol{\Sigma}^{-1}\operatorname{diag}(\mathbf{s}) \succcurlyeq 0$. and $\mathbf{s} \in \mathbb{R}^p$ is some vector. Let us now examine the steps that lead to this result. We begin with a general form of $\tilde{\mathbf{X}}$. A fundamental result in Linear Algebra is the Orthogonal Decomposition Theorem [8], which states that any vector (or matrix, if we consider its columns as vectors) can be uniquely decomposed into components, each lying in an orthogonal subspace. Therefore, it must be that we can write $\tilde{\mathbf{X}}$ in the following form.

$$\tilde{\mathbf{X}} = \mathbf{X}\mathbf{A} + \tilde{\mathbf{U}}\mathbf{C}$$

Here, $\mathbf{A}, \mathbf{C} \in \mathbb{R}^{p\times p}$. By construction, we have that $\tilde{\mathbf{X}}^\top\mathbf{X} = \boldsymbol{\Sigma} - diag(\mathbf{s})$. Multiplying both sides by $\mathbf{X}$ and subbing in our general form of $\tilde{\mathbf{X}}$, we are left with the following result.

$$\boldsymbol{\Sigma} - diag(\mathbf{s}) = (\mathbf{X}\mathbf{A} + \tilde{\mathbf{U}}\mathbf{C})^\top\mathbf{X}$$

Using established transposition rules [13],

$$\boldsymbol{\Sigma} - diag(\mathbf{s}) = \mathbf{A}^\top\mathbf{X}^\top\mathbf{X} + \mathbf{C}^\top\tilde{\mathbf{U}}^\top\mathbf{X}$$

By construction, we have that $\tilde{\mathbf{U}}^\top\mathbf{X} = \mathbf{0}$ meaning we can simplify the expression such that $\boldsymbol{\Sigma} - diag(\mathbf{s}) = \mathbf{A}^\top\mathbf{X}^\top\mathbf{X} = \mathbf{A}^\top\boldsymbol{\Sigma}$. From here, the multiplication of both sides by $\boldsymbol{\Sigma}^{-1}$ leaves us with the following result:

$$A^\top = I - \boldsymbol{\Sigma}^{-1}\operatorname{diag}(\mathbf{s})$$

Therefore, we are left with our result. In cases where $p < n < 2p$, constructing the matrix $\tilde{\mathbf{U}}$ becomes not possible. A workaround involves extending the data provided we know about the noise level in the data. For example if we know the noise to follow some Gaussian distribution, the response

vector $\mathbf{Y}$ is extended using random Gaussian noise and the design matrix $\mathbf{X}$ is padded with rows of zeros. Essentially, you are creating additional values for the response purely based on noise and treating the covariates as 0. These added rows don't carry any real information but do help to create a new system where $n \geq 2p$, enabling the knockoff filter to be applied as above while preserving the required knockoff properties.

Model-X construction

Within the Model-X framework, we do not have a fixed covariate matrix but instead, $\mathbf{X}$ is random and comes from $P_{\mathbf{X}}$, some known distribution. In other words, $\mathbf{X}$ is not given explicitly in raw data but random variables. We are still required to construct knockoffs $\tilde{\mathbf{X}}$ that satisfy the conditions outlined in chapter 2 however it is important to consider how the way $\mathbf{X}$ is interpreted may alter the significance of these conditions. Clearly, the fixed-X approach ensures that the sample (as in raw data) covariance of $[\mathbf{X}, \tilde{\mathbf{X}}]$ is invariant under the swapping of variables with their knockoff counterparts. Now, we require that the population (under the given distribution) covariance is invariant. Consider the following example of how we could construct Gaussian knockoffs.

Suppose that we are in a situation where we do not have raw data but we know that our distribution of covariates comes from a multivariate normal distribution.

$$\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Recall, our aim is to construct knockoffs such that the joint distribution does not change under column swapping. Intuitively, this suggests that the marginal distribution of knockoffs need to have the same distribution as the original covariates i.e.

$$\mathbf{X} \overset{d}{=} \tilde{\mathbf{X}}$$

As described in Section 2.1.2, a proposal is to construct knockoffs in such a way that their joint distribution is Gaussian with some mean and covariance.

$$(\mathbf{X}, \tilde{\mathbf{X}}) \sim N(\boldsymbol{\mu}, \mathbf{G})$$

Where

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

The diagonal elements are constrained by the fact that we require that both covariance structures of covariates and knockoffs to be equal to $\boldsymbol{\Sigma}$. Furthermore, we require that the covariance between X_j and X_k $(j \neq k)$ is the same as the covariance between X_j and $\tilde{X}_k$. This is encoded in the off diagonal terms of $\mathbf{G}$. Now, to verify the exchangeability property, consider how the joint distribution would change under the swapping of a covariate column $\mathbf{X}_j$, with its corresponding knockoff counterpart $\tilde{\mathbf{X}}_j$. Consider the following example when $p = 3$.

$$X = (X_1, X_2, X_3) \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\tilde{X} = (\tilde{X}_1, \tilde{X}_2, \tilde{X}_3) \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3) \sim N(\boldsymbol{\mu}, \mathbf{G})$$

Where

$$\mathbf{G} = \begin{bmatrix} \sigma_1 & \sigma_{12} & \sigma_{13} & \sigma_1 - s_1 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2 & \sigma_{23} & \sigma_{12} & \sigma_2 - s_2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3 & \sigma_{13} & \sigma_{23} & \sigma_3 - s_3 \\ \sigma_1 - s_1 & \sigma_{12} & \sigma_{13} & \sigma_1 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2 - s_2 & \sigma_{23} & \sigma_{12} & \sigma_2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3 - s_3 & \sigma_{13} & \sigma_{23} & \sigma_3 \end{bmatrix}$$

Suppose we swap covariate X_1 with its corresponding knockoff $\tilde{X}_1$.

$$(\tilde{X}_1, X_2, X_3, X_1, \tilde{X}_2, \tilde{X}_3) \sim N(\boldsymbol{\mu'}, \mathbf{G}')$$

Now, clearly this has no effect on the mean of the joint distribution as the mean of X_1 is equal to the mean of $\tilde{X}_1$ so $\boldsymbol{\mu'} = \boldsymbol{\mu}$. Swapping X_1 and $\tilde{X}_1$ means interchanging the first and fourth rows and columns of $\mathbf{G}$, which results in a new covariance matrix $\mathbf{G}'$. Formally, we apply the permutation matrix $\mathbf{P}$ that swaps the first and fourth entries:

$$\mathbf{P} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Applying this transformation to $\mathbf{G}$, the new covariance matrix is:

$$\mathbf{G}' = \mathbf{P}\mathbf{G}\mathbf{P}^T$$

Expanding $\mathbf{G}'$, we get:

$$\mathbf{G}' = \begin{bmatrix} \sigma_1 & \sigma_{12} & \sigma_{13} & \sigma_1 - s_1 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2 & \sigma_{23} & \sigma_{12} & \sigma_2 - s_2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3 & \sigma_{13} & \sigma_{23} & \sigma_3 - s_3 \\ \sigma_1 - s_1 & \sigma_{12} & \sigma_{13} & \sigma_1 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2 - s_2 & \sigma_{23} & \sigma_{12} & \sigma_2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3 - s_3 & \sigma_{13} & \sigma_{23} & \sigma_3 \end{bmatrix} = \mathbf{G}$$

Therefore, clearly, with this construction, we see that the swapping of a knockoff with its corresponding covariate satisfies the exchangeability property.

Choosing vector s

From our calculations, we have that the diagonal entries of $\mathbf{X}^\top \tilde{\mathbf{X}}$ (which measure the correlations between covariates and their counterpart knockoffs) equate to $1 - s_i$ for covariate i meaning that the variables do not have to perfectly correlate with their corresponding knockoffs. This is beneficial as, by maximising s_i, $\mathbf{X}_i$ and $\tilde{\mathbf{X}}_i$ are less correlated meaning that it will be easier to distinguish between their contribution when it comes to assigning each of them a feature importance statistics (such as the coefficient in a lasso model). As $s_i \rightsquigarrow 0$, these variables would be very highly correlated as $1 - s_i \rightsquigarrow 1$. Therefore, their contribution in the prediction of the model would become increasingly similar. In the example of the Lasso model, this would result in $\mathbf{X}_i$ and $\tilde{\mathbf{X}}_i$ having coefficients of similar magnitude, thus making it increasingly difficult to distinguish between them. This can lead to an increase in both Type I (false discoveries) and Type II errors (false negatives), making it difficult to ensure the model maintains a high level of statistical power (1 - Type 2 error). Therefore, we need to choose s_i as large as possible (such that X_i and $\tilde{X}_i$ are as orthogonal as possible) subject to the constraint that $\mathbf{G}$ is positive semidefinite. This semidefinite condition is necessary as$\mathbf{G}$ represents a covariance matrix, and covariance matrices must be positive semidefinite to ensure that the variances and covariances they represent are non-negative [2].

Now that we have described the construction of knockoffs, we shall use this to prove the exchangeability property of null covariates, which is an essential result in ensuring the exchangeability of feature importance statistics and following this, the coin flip property of the associated knockoff statistics of null covariates.

3.2 Proof of exchangeability of null covariates

Lemma 6. *Suppose we have our covariate data matrix $\boldsymbol{X} = (X_1, X_2, ..., X_p)$. Suppose we have some subset $S \subseteq \{1, 2, ..., p\}$ whereby $\forall i \in S, X_i$ is null (i.e. all labels in S correspond to a null covariate variable). Then*

$$(\boldsymbol{X}, \tilde{\boldsymbol{X}})|Y \stackrel{d}{=} (\boldsymbol{X}, \tilde{\boldsymbol{X}})_{swap(S)}|Y$$

where $(\boldsymbol{X}, \tilde{\boldsymbol{X}}) = (X_1, X_2, ..., X_p, \tilde{X_1}, \tilde{X_2}, ..., \tilde{X_p})$ *and* $(\boldsymbol{X}, \tilde{\boldsymbol{X}})_{swap(S)}$ *is calculated by swapping the entries of* X_j *and* $\tilde{X_j}$ $\forall j \in S$*. For example, suppose* $S = \{1, 2\}$*. Then* $(\boldsymbol{X}, \tilde{\boldsymbol{X}})_{swap(S)} = (\boldsymbol{X}, \tilde{\boldsymbol{X}})_{swap(\{1,2\})} = (\tilde{X_1}, \tilde{X_2}, ..., X_p, X_1, X_2, ..., \tilde{X_p})$

Proof. [3] Since each row in our dataset is independent, we only need to show the proof for a single arbitrary row. Define X as the i^{th} row of $\mathbf{X}$. Here, consider Y_i as the target for row i. Therefore, it suffices to show that

$$(X, \tilde{X})|Y_i \stackrel{d}{=} (X, \tilde{X})_{swap(S)}|Y_i$$

which is analogous to proving

$$p_{((X,\tilde{X})|Y_i)}((x, \tilde{x})|y) = p_{((X,\tilde{X})_{swap(S)}|Y_i)}((x, \tilde{x})_{swap(S)}|y) \tag{3.1}$$

Note the following:

$$p_{((X,\tilde{X})|Y_i)}((x, \tilde{x})|y) = \frac{p_{((X,\tilde{X}),Y_i)}((x, \tilde{x}), y)}{p_{Y_i}(y)}$$

$$p_{((X,\tilde{X})_{swap(S)}|Y_i)}((x, \tilde{x})_{swap(S)}|y) = \frac{p_{((X,\tilde{X})_{swap(S)},Y_i)}((x, \tilde{x})_{swap(S)}, y)}{p_{Y_i}(y)}$$

Therefore, we must have that proving (3.1) is equivalent to proving

$$((X, \tilde{X}), Y_i) \stackrel{d}{=} ((X, \tilde{X})_{swap(S)}, Y_i) \tag{3.2}$$

Now, note the following:

$$p_{((X,\tilde{X}),Y_i)}((x, \tilde{x}), y) = p_{(X,\tilde{X})}((x, \tilde{x})) \cdot p_{Y_i|(X,\tilde{X})}(y|(x, \tilde{x}))$$

By the exchangeability properties in the construction of the knockoffs, we have that $(X, \tilde{X}) \stackrel{d}{=} (X, \tilde{X})_{swap(S)}$. This is equivalent to the equality, $p_{(X,\tilde{X})}((x, \tilde{x})) = p_{(X,\tilde{X})_{swap(S)}}((x, \tilde{x})_{swap(S)})$ Therefore, we must have that proving (3.2) is equivalent to proving

$$Y_i|(X, \tilde{X}) \stackrel{d}{=} Y_i|(X, \tilde{X})_{swap(S)} \tag{3.3}$$

We have that the joint distribution of $(X, \tilde{X})$ is invariant under column swapping. The construction of Model-X knockoffs ensures that swapping columns of X and $\tilde{X}$ preserves the overall distribution of $(X, \tilde{X})$. This is equivalent to a relabelling of the inputs to the conditional distribution and so does not alter the joint distribution. This invariance implies the following.

$$p_{Y_i|(X,\tilde{X})}(y|(x, \tilde{x})_{swap(S)}) = p_{Y_i|(X,\tilde{X})_{swap(S)}}(y|(x, \tilde{x})$$

By construction, we have that, conditionally on our covariates, our knockoff $\tilde{X}$ and target Y are independent of each other. This means that we must have the following result.

$$p_{Y_i|(X,\tilde{X})}(y|(x, \tilde{x})_{swap(S)}) = p_{Y_i|X}(y|x')$$

where $x'_i = x_i$ if $i \notin S$ and $x'_i = \tilde{x}_i$ if $i \in S$. Note, all we have done here is disregard the knockoffs due to independence. By definition, X_j is said to be null if and only if $Y_i \perp\!\!\!\perp X_j|X_{-j}$. Recall, all labels in S correspond to a null covariate variable. X refers to an entire row of covariates in our dataset. Essentially,

$$p_{Y_i|X}(y|x') =: p_{Y_i|X_{1:p}}(y|x'_{1:p})$$

Assume, without loss of generality, that $S = \{1, ..., k\}$ for some $1 \leq k \leq p$. We can safely make this assumption because S is simply holding the column numbers of our null covariates and a simple reordering to format our dataset in such a way that the first k columns are null variables will not impact our calculations. Now, since $1 \in S$, we have that $Y \perp\!\!\!\perp X_1 | X_{-1}$. Hence, we can effectively say that X_1 can take on any value and it should not impact the conditional probability of Y and so all the following terms are equal.

$$p_{Y_i|X_{1:p}}(y|x'_{1:p}) = p_{Y_i|X_{2:p}}(y|x'_{2:p}) = p_{Y_i|X_{1:p}}(y|\tilde{x_1}, x'_{2:p}) = p_{Y_i|X_{1:p}}(y|x_1, x'_{2:p})$$

Therefore, in either case ($X_1 = x_1$ or $X_1 = \tilde{x_1}$), the conditional distribution is the same. We therefore can state the following

$$Y_i|(X)_{swap(S)} \stackrel{d}{=} Y_i|(X)_{swap(S\setminus\{1\})} \tag{3.4}$$

Since the knockoffs, $\tilde{X}$ are independent of Y conditionally on X, we can include them in equation (3.4) to get

$$Y_i|(X, \tilde{X})_{swap(S)} \stackrel{d}{=} Y_i|(X, \tilde{X})_{swap(S\setminus\{1\})}$$

We repeat this process for all the remaining terms in S (values $2, ...k$) and so are left with the following result.

$$Y_i|(X, \tilde{X})_{swap(S)} \stackrel{d}{=} Y_i|(X, \tilde{X}) \tag{3.5}$$

Combining results in (3.1),(3.2),(3.3),(3.5) we have our result.

□

Chapter 4

The knockoff-based variable selection process

Now that we have established our statistical foundations and considered how knockoff variables are constructed both in the Fixed-X and Model-X paradigms, we will present the knockoff-based variable selection process from start to finish. Recall the original setup from Chapter 1 whereby we have n rows of independent, identically distributed data. In this dataset, there are $p + 1$ columns; we have our covariate data matrix, $\mathbf{X}$ which has p columns $\{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_p\}$ which each correspond to some covariate. Additionally, we have a final column $\mathbf{Y}$ corresponding to our target variable. In each row i, we therefore have $(\mathbf{X}_{i1}, \mathbf{X}_{i2}, ..., \mathbf{X}_{ip}, \mathbf{Y}_i)$. We first establish the paradigm we are choosing to implement given the context of our problem and construct our knockoffs, $\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, ..., \tilde{\mathbf{X}}_p$, accordingly as described in Chapter 3. Following this, we concatenate $\mathbf{X}$ and $\tilde{\mathbf{X}}$ to form the augmented design matrix $[\mathbf{X}, \tilde{\mathbf{X}}]$. From here, we begin by calculating feature importance statistics for all variables and knockoffs.

4.1 Calculating feature importance statistics

As the name suggests, the feature importance statistic Z_j (or $\tilde{Z}_j$) associated with some covariate $\mathbf{X}_j$ (or knockoff $\tilde{\mathbf{X}}_j$) is indicative of the level of contribution that the variable provides to the model. To compute these statistics, we initially have to fit a predictive model for the target $\mathbf{Y}$ using the combined design matrix $[\mathbf{X}, \tilde{\mathbf{X}}]$. In the Fixed-X case, we have already established this assumption; for example, we may have assumed a Lasso regression model for the response. In the model-X case, no rigid assumption is made regarding the relationship between covariates and target meaning we can leverage more flexible, non-parametric models that are more capable of modelling non-linear relationships. In either case, based on the model we implement, we can compute the importance scoresZ_j for original variables X_j and $\tilde{Z}_j$ for knockoff variables $\tilde{X}_j$. Let us consider some example cases.

1. **Lasso Regression Model**: We previously defined in Chapter 1 how the feature importance statistics are computed within a lasso regression model. Recall a lasso regression with l_1-norm penalized regression is similar to a standard linear regression model as it models only linear interactions between variables. Formally, coefficients are defined

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left(\frac{1}{2} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right)$$

   This definition ensures that the coefficients of irrelevant features are shrunk to zero. The feature importance statistic for variable $\mathbf{X}_j$ is calculated as

$$Z_j = \sup \left\{ \lambda : \hat{\beta}_j(\lambda) \neq 0 \right\}.$$

   As mentioned earlier, this value will be large for relevant variables, as it is the maximum penalization required for the coefficient $\hat{\beta}_j(\lambda)$ to become zero. Conversely, for irrelevant features, the value will be small, indicating that only a weak penalization is needed to eliminate the feature from the model.

2. **Random Forest**: Now, suppose we are operating within a model-X framework. A common approach for computing variable importance statistics in a random forest model is the computation of *out of bag* variable importance scores (OOB VIMP). The Random Forest [25], is a bagging algorithm that averages the results from many individual decision trees to improve prediction accuracy and reduce over-fitting. Each of these decision trees is trained on a bootstrapped sample of the data, meaning that while all the data is used in training, no individual tree is trained on the entire dataset [10]. The process of computation is as follows.

   (a) Fit the random forest (RF) to model the relationship between $\mathbf{Y}$ and $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_p\}$.

   (b) Fit the random forest (RF') to model the relationship between $\mathbf{Y}$ and $\mathbf{X}' = \{\tilde{\mathbf{X}}_1, \mathbf{X}_2, ..., \mathbf{X}_p\}$. This model is used to isolate the impact that $\mathbf{X}_1$ has on predicting the target $\mathbf{Y}$.

   (c) Consider the k^{th} row of our dataset. This is represented as

$$D_k = \Big\{\{\mathbf{X}_{k,1}, \mathbf{X}_{k,2}, \ldots, \mathbf{X}_{k,p}\}, \{\tilde{\mathbf{X}}_{k,1}, \tilde{\mathbf{X}}_{k,2}, \ldots, \tilde{\mathbf{X}}_{k,p}\}, \mathbf{Y}_k\Big\}$$

   For each row $k \in \{1, \ldots, n\}$

   i. Using the RF model and the covariates $\{\mathbf{X}_{k,1}, \mathbf{X}_{k,2}, \ldots, \mathbf{X}_{k,p}\}$, compute the prediction accuracy (%),$P_{k,1}$, by comparing the prediction given by the model to $\mathbf{Y}_k$, the true value.

   ii. Using the RF' model and the covariates $\{\tilde{\mathbf{X}}_{k,1}, \mathbf{X}_{k,2}, \ldots, \mathbf{X}_{k,p}\}$, compute the prediction accuracy (%), $P'_{k,1}$, by comparing the prediction given by the model to $\mathbf{Y}_k$, the true value.

   iii. Recall, we have n rows of data. Therefore, we can computer the average prediction accuracy over all participants for both models. Formally,

$$Z_1 = \frac{1}{n}\sum_{k=1}^{n} P_{k,1} \text{ and } \tilde{Z}_1 = \frac{1}{n}\sum_{k=1}^{n} P'_{k,1}$$

   iv. Repeat steps (a)-(d) for each feature $\mathbf{X}_2, ..., \mathbf{X}_p$ to compute Z_j and $\tilde{Z}_j$ for $j \in \{2, ..., p\}$

4.2 Computing knockoff statistics

Now we have computed our variable importance statistics Z_j and Z_j for $j \in [p]$. We now introduce the knockoff statistic, $W_j = f(Z_j, \tilde{Z}_j)$ where f is some real valued function. This statistic plays a crucial role in knockoff methodology as it allows for us to distinguish between the variables we deem relevant and those we consider null via the data-dependent threshold, T. Before we examine the particular method behind the construction, recall the exchangeability property of null covariates with their corresponding knockoffs. This will play a pivotal role in the construction of W_j. More specifically, The lemma enables us to prove an important coin flip property related to null covariates.

The construction of $W_j = f(Z_j, \tilde{Z}_j)$ is such that a large positive value of W_j suggests strong evidence against the null hypothesis (covariate X_j is null) since a truly relevant covariate X_j should corresponding feature importance statistic larger than its knockoff which is truly null. Within our framework, if W_j is large and positive, it will exceed the data-dependent threshold T, which we use to categorize covariates as relevant.

An essential property of W_j is that the function f must be antisymmetric, i.e., $f(\tilde{Z}_j, Z_j) = -f(Z_j, \tilde{Z}_j)$ [2]. A common example of this function could be $W_j = f(Z_j, \tilde{Z}_j) = Z_j - \tilde{Z}_j$. To understand why this is important, consider the following cases.

Case 1: Null

Suppose that X_j is null. By construction, we therefore have that for null covariates, the associated statistic W_j is distributed symmetrically around 0 and is therefore equally likely to be positive or negative. In other words, by the construction of the knockoffs ensuring pairwise exchangeability of covariates and their corresponding knockoffs, in the case of null j, due to its antisymmetric property, W_j is equally likely to be positive or negative.

Lemma 7. *Coin Flip property of null covariates [3]*
Define $\mathcal{H}_0$ as the set containing the index values of null variables. Define $\epsilon = (\epsilon_1, ..., \epsilon_p)$ where $\epsilon_i = 1$ for $i \notin \mathcal{H}_0$ and $\epsilon_i = \pm 1$ with probability $\frac{1}{2}$ in each case for $i \in \mathcal{H}_0$. Define $S = \{j : \epsilon_j = -1\} \subset \mathcal{H}_0$. Then, $\forall j \in \mathcal{H}_0, p(sign(W_j) = 1|Y) = p(sign(W_j) = -1|Y) = \frac{1}{2}$.

Proof. Consider $j \in \mathcal{H}_0$ i.e. j corresponds to a null covariate. As stated, $W_j = f(Z_j, \tilde{Z}_j)$. The feature importance statistics Z_j and $\tilde{Z}_j$ are generated based on covariates $(\mathbf{X}, \tilde{\mathbf{X}})$ and target $\mathbf{Y}$. Define the following.

$$W = (W_1, W_2, ..., W_p)$$
$$W_{swap(S)} = (W_1', W_2', ..., W_p')$$

where $W_j' = W_j$ if $j \notin S$ and $W_j' = -W_j$ $j \in S$. By the definition of ϵ, we are required to prove that $W_j \overset{d}{=} \epsilon_j \cdot W_j$. In other words, the distribution is symmetric around 0 for null covariates. Suppose $j \notin S$. In this case, $\epsilon_j = 1$ and so $\epsilon \cdot W_j = W_j$. Since $j \notin S$, $W_j' = W_j$. Therefore, $\epsilon \cdot W_j = W_j'$. Suppose $j \in S$. In this case, $\epsilon_j = -1$ and so $\epsilon \cdot W_j = -W_j$ and since $j \in S$ $W_j' = -W_j$. Therefore, $\epsilon \cdot W_j = W_j'$. Therefore, we have that

$$\epsilon \cdot W = W_{swap(S)} \tag{4.1}$$

From Lemma 6, we have that $((X, \tilde{X}), Y_i) \overset{d}{=} ((X, \tilde{X}_{swap(S)}), Y_i)$ which implies the following:

$$W_{swap(S)} = (W_1', W_2', ..., W_p') = w((\mathbf{X}, \tilde{\mathbf{X}})_{swap(S)}, \mathbf{Y}) \overset{d}{=} w((\mathbf{X}, \tilde{\mathbf{X}}), \mathbf{Y}) = (W_1, W_2, ..., W_p) = W \tag{4.2}$$

Where the second and penultimate equality simply means we generate our knockoff statistics for each $j \in [1, .., p]$ based on the given covariate and target data. Combining results (4.1) and (4.2), we are left with the following result.

$$\epsilon \cdot W \overset{d}{=} W$$

Therefore, we have our result. □

Case 2: Relevant

Now, in the case where X_j is relevant, as aforementioned, we would expect the important statistic Z_j to be large and positive and would expect that the knockoffs important statistic $\tilde{Z}_j$ to be less than Z_j since the knockoff is designed independently of our target and is therefore null by construction. In addition, the correlation between a variable and its knockoff should be small. In the example of the lasso regression model, this ensures that the knockoff does not enter the model early (more specifically, earlier than its actual variable counterpart). This is ensured by choosing an appropriate value of $\mathbf{s}$ during construction as outlined in Chapter 3.

4.3 Calculating a data-dependent threshold

The final step is to construct our data-dependent threshold using the properties of W_j. If variable X_j has a corresponding knockoff statistic W_j that exceeds this threshold then we deem it a relevant variable. If W_j does not exceed this threshold then it is considered a null variable. Recall, our ultimate goal is to control the False Discovery Rate whilst optimising our power the best we can. We define

$$FDR = \mathbb{E}(FDP(T)) \text{ where } FDP(T) = \frac{|\{j \text{ null: } W_j \geq T\}|}{|\{j : W_j \geq T\}|}$$

$$Power = \frac{|\{j \text{ truly relevant: } W_j \geq T\}|}{|\{j \text{ truly relevant}\}—}$$

Now, consider the term $|\{j \text{ null: } W_j \geq T\}|$. Clearly, we cannot directly calculate this as we do not know the ground truth; which variables are truly null and which are truly relevant. Instead, we can estimate this value. From Lemma 7, we have that the sign of W_j is an iid coin flip for null variables which is equivalent to saying that the distribution of W_j is symmetric around 0. Therefore, by uniqueness of cumulative distribution functions,

$$P(j \text{ null: } W_j \geq T) = P(j \text{ null: } W_j \leq -T) \iff |\{j \text{ null: } W_j \geq T\}| \stackrel{d}{=} |\{j \text{ null: } W_j \leq -T\}| \quad (4.3)$$

Now, relevant variables are likely to have large positive values for W_j, which means we can approximate $|\{j \text{ null: } W_j \leq -T\}|$ as $|\{j : W_j \leq -T\}|$, since $|\{X_j \text{ relevant: } W_j \leq -T\}|$ is estimated to be close to 0. Therefore we can estimate our False Discovery proportion with $\hat{\text{FDP}}(T)$

$$\hat{\text{FDP}}(T) = \frac{|\{j : W_j \leq -T\}|}{|\{j : W_j \geq T\}|} \quad (4.4)$$

Now, $\hat{\text{FDP}}(T)$ is a term that we can compute. Suppose we set our false discovery rate to some value $q \in (0, 1]$. Using q, we can then select an appropriate value for T. Using our results from Chapter 2, theorems 1 and 2, we have that we can select an appropriate value of T, namely

$$T = \min\{t \in \{|W_1|, ..., |W_p|\} : \hat{\text{FDP}}(t) = \frac{|\{j : W_j \leq -t\}|}{|\{j : W_j \geq t\}|} \leq q\} \quad (4.5)$$

This choice of T ensures approximate FDR control since $\mathbb{E}(\hat{\text{FDP}}(T)) \leq q$. Notice, we restrict the choices of T to the observed magnitudes $|W_j|$ meaning the threshold is based only on the test statistics present in the dataset as opposed to searching over all positive real numbers which would be computationally intensive and unnecessary. Additionally, by restricting the choices of T, the procedure aligns with the actual test statistic distribution and avoids using arbitrary thresholds. Notice we choose the minimum of this set is to maximise the number of variables we select. This works to maximise our power whilst still ensuring FDR control by choosing the lowest threshold that ensures this approximate FDR control. If we were to select the maximum of the set as our threshold, T, we are more likely to exclude relevant variables, reducing the power and missing true discoveries.

Chapter 5

Experiments on simulated data

5.1 Introduction to the experimental setup

In this Chapter, we introduce the experimental side to this paper as we perform tests using simulated data. The goal of these tests is to validate the Model-X knockoff method. Specifically, FDR serves as a fixed constraint, and our objective is to optimize power within this constraint. In addition to the traditional Model-X framework (which we refer to in our simple simulation study), we consider the application of knockoffs to predictive biomarker identification (see Chapter 1). Thinking of the experiments through the lens of biomarker identification, our focus in section 5.2 will be on identifying prognostic variables, those that are associated with the outcome regardless of treatment (in fact, in this section, no treatment variable will be considered). In section 5.4, our focus shifts to identifying predictive variables, those whose effect on the outcome is influenced by their interaction with treatment. By framing the experiments in this way, we not only validate the methodology but also explore its practical relevance in biomarker discovery. Furthermore, as a way of simulating alternative real-world scenarios, we will vary the parameters within our model such as the number of truly relevant variables and the quantity of data. In addition to this, we will asses how varying feature importance statistics (as discussed in Chapter 4) impacts FDR and power scores. Specifically, we compare the results generated from the feature importance statistics of the lasso and random forest models. All the corresponding code for these results can be found in the GitHub repository provided in appendix A. In the Simple simulation study, we base our experiments on the foundational material provided By Stanford University [21], whereas in the Biomarker study, the code is built upon previous work written by Sechidis et al [20].

5.2 Simple simulation study: Experimental design

In this section, we will discuss the foundations on which we perform variable selection in terms of the data we choose to generate, ensuring that it captures the underlying relationships between the variables while reflecting the complexities of the real world. We begin within the standard variable selection framework (without a treatment variable).

5.2.1 Data generation

Let us first establish how we will generate our synthetic data. Here, there are 3 major considerations; the covariate data, $\mathbf{X}$, the corresponding knockoffs, $\tilde{\mathbf{X}}$ and the target data $\mathbf{Y}$. These steps outline the process for generating such data that will be used in our analysis.

1. Generate Covariate data, $\mathbf{X}$
   The process of generating our covariates depends on the underlying assumptions about the distribution of the covariates. The multivariate normal distribution is a common choice for modelling multivariate data for a multitude of reasons [27] [9]. For one, the multivariate distribution is well-understand mathematically and is not as computationally burdensome as more complex models. Additionally, the multivariate normal distribution often arises naturally in high-dimensional

spaces due to the Central Limit Theorem [5]. We restrict our experiments to the high-dimensional scenarios and therefore have that that the number of rows of data, n, is equal to the number of covariates, $p = 1000$. From here, we randomly select a subset of these covariates that Y will depend on. In our experiments, the default number of truly relevant covariates will be 100.

2. Generate Target Data $\mathbf{Y}$
   As discussed in Chapter 1, a common approach is to model $\mathbf{Y}$ as a linear combination of a subset of variables $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p$. While this simplifies variable selection, it imposes a strong assumption about the relationship between the covariates and the target variable—specifically, that this relationship is strictly linear. This assumption may overlook potential non-linear dependencies, which are particularly relevant in complex, high-dimensional settings. In our simple study experiments, we will assume a linear relationship. However, as discussed in Chapter 3, it is important to highlight that Model-X knockoffs offer a key advantage over the fixed-X approach: they allow for the use of more flexible models that do not strictly require a predefined functional form between $\mathbf{X}$ and $\mathbf{Y}$. This flexibility makes Model-X knockoffs particularly suitable for scenarios where non-linearity is expected, as they relax the need for strong structural assumptions at the outset.

3. Generate Knockoffs $\tilde{\mathbf{X}}$
   Gaussian Synthetic Knockoff variables are constructed as described in Chapter 3.

5.3 Simple simulation study: Key experiments

Within each experiment, we will explore the effects of varying model hyperparameters and the choice of importance statistics to assess their impact on performance, which we measure in terms of power under the constraint of FDR control. These experiments will involve the following factors:

1. Varying the nominal FDR level q. This is a crucial experiment as it justifies the implementation of the knockoff methodology as a means of guaranteeing FDR control.

2. Varying sample size, n. This experiments emphasizes the control of the false discovery rate in a finite sample rather than just asymptotically.

3. Varying the number of truly relevant covariates. This experiment establishes how the knockoff filter ensures FDR control as we change the number of truly relevant covariates within the model and the impact that this has on the power of the tests.

4. Miss-specification: We explore how the knockoff methodology performs when the distribution of the covariates is miss-specified. The theory suggests that we can control FDR while attempting to maximize power. However, it does not directly address whether the inverse holds; that is, can we still achieve FDR control and reasonable power when the knockoffs are generated from a distribution that does not match the one assumed in the knockoff construction?

5.3.1 Varying the nominal FDR level, q

Recall, the nominal FDR level, q, is the maximum level of false discovery that we are tolerating in our variable selection process. For example, in this experiment, setting $q = 0.05$ would mean we aim to control the false discoveries such that no more than 5% of the covariates we deem relevant are null in reality. In our experiment, we vary our values of $q \in \{0.01, 0.05, 0.1, 0.2, 0.5\}$. By varying, we are testing how different levels of false discovery tolerance impact the method's ability to maintain FDR control. Ideally, the guarantee of FDR control is valid regardless of the choice of q. For low values of q, the method will be highly conservative and have a tight control of FDR, resulting in fewer false positives although potentially a reduced level of power (as discussed in Chapter 2) with the converse being true as we raise our value of q.

First, consider Figure 5.1. From the plot, we observe that the false discovery proportion (FDP) increases as we raise the value of q. This is expected, as a higher value of q means that the selection criteria for covariates become less strict, leading to more variables being selected, including some false positives. This confirms that the knockoff-based variable selection mechanism is functioning properly

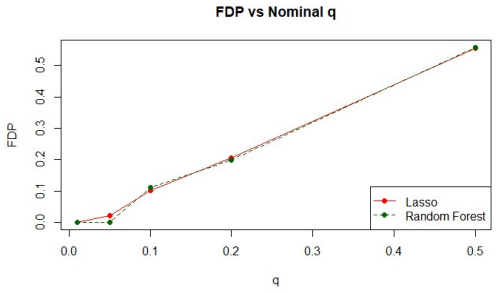

Figure 5.1: A plot measuring FDP as we vary our nominal level of false discovery, q.

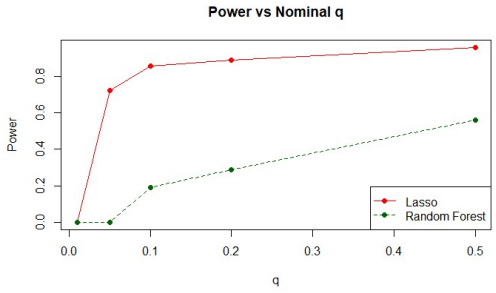

Figure 5.2: A plot measuring power as we vary our nominal level of false discovery, q.

by controlling the false discovery rate (FDR). Next, consider the power plot in Figure 5.2. As q increases, we observe that the power of the test also increases. This aligns with intuition: relaxing the relevance criteria associated with higher values of q reduces the likelihood of missing truly relevant variables. However, we also see that as q continues to increase, the rate of power increase slows down and looks to plateau. This suggests that after a certain point, further relaxation of the relevance threshold leads to diminishing returns in terms of power. Let us now perform multiple tests for each value of q and compute the mean values of power and FDP. By averaging the results across multiple (5) tests, we can better assess the overall performance of the variable selection procedure at each q. for each value of q, we have also produced a 95% confidence interval (with the critical value from the standard normal distribution) to give an indication of the variability of the results. Although we only perform 5 repeated experiments for each value of q due to computational constraints, the consistency of the results suggests that the asymptotic confidence interval remains reasonably valid. we compute the confidence intervals for each metric (FDP and power) at a given q, $m(q)$ with mean $\bar{m}$ and standard deviation $\sigma_{m(q)}$

$$\text{CI} = \bar{m} \pm 1.96 \times \frac{\sigma_{m(q)}}{\sqrt{5}}$$

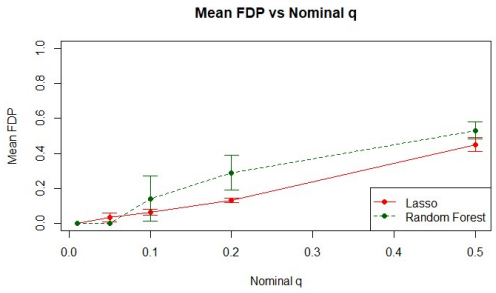

Figure 5.3: A plot of the mean proportion of false discoveries as q is varied.

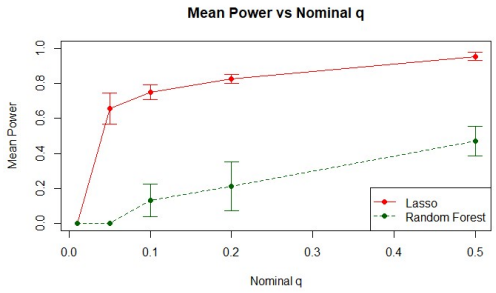

Figure 5.4: A plot of the mean power score as q is varied.

As is expected and as we will see in many of our tests, the lasso model outperforms the random forest model in terms of power since our data construction is designed such that there are only linear interactions between variables thus making the linear model's assumptions well-suited for the task. However, the random forest model, which is a non-parametric approach, meaning it does not inherently assume solely linear interactions, naturally struggles more in terms of power. As is seen in biomarker paper [20], when non-linear interactions are introduced between variables, the filter based on knockoff statistics from the non-parametric Random Forest model outperforms the linear interaction filter in terms of power, while still maintaining control over the false discovery rate (FDR). Now that we see the

framework can control the false discovery rate, we consider how performance differs when we increase the volume of data we have.

5.3.2 Varying sample size, n

In the following experiment, we vary the quantity of data that we have in order to perform variable selection. The goal of this test is to demonstrate how knockoff-based methods can successfully control the false discovery rate (FDR) even with a finite sample size, rather than relying on asymptotic properties.

In classical variable selection techniques, asymptotic results assume that n grows and p remains fixed. This allows error terms to shrink, leading to well-behaved estimates and valid hypothesis tests [24]. However, in the case where p is large relative to n, much of the traditional variable selection techniques fail. These techniques rely on the central limit theorem that states that the sum of a large number of independent random variables tends to a Gaussian distribution irrespective of the distribution of the original variables. In high-dimensional or finite sample settings, the error terms become less well controlled and therefore any central-limit theorem based variable selection approach breaks down in validity. In contrast, as described in the methodology, knockoff-based procedures do not rely on asymptotic normality. Instead, they provide valid variable selection even when p is comparable with n maintaining rigorous FDR control in finite samples. We will conduct an experiment to explore the impact of sample size on FDR control. Simply, we will increase n and analyse how this impacts the proportion of false discoveries and power.

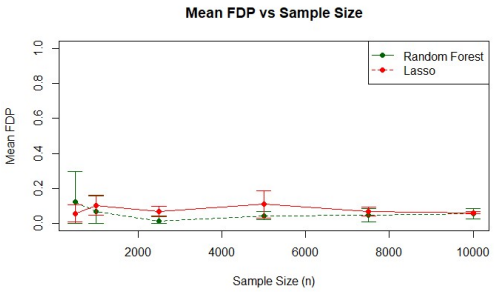

Figure 5.5: A plot of the mean proportion of false discoveries as n is varied

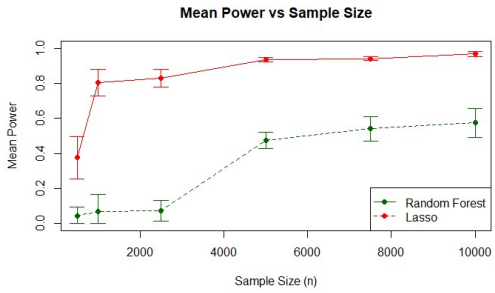

Figure 5.6: A plot of the mean power score as n is varied

Throughout the experiments, we have that the nominal false discovery rate, q, is set to a default value of 0.1. From figure 5.5, it is clearly that we obtain good FDR control even when it is not the case when $n >> p$. This is due to the mechanism by which variables are selected (as discussed in 4) being unconcerned by the number of rows of data. However, the power plot, figure 5.6, suggests that as we increase the sample size, we see improved power scores. More data points to estimate the underlying relationship between predictors and outcome leads to more reliable estimates of coefficients. This means that truly relevant variables are less likely to be missed by the model with more data. Now, we turn our attention to how varying the number of truly relevant covariates affects the performance of the knockoff method. After examining how the number of relevant covariates influences performance, we now consider how the knockoff methodology behaves when the assumptions underlying the covariate distribution are misspecified.

5.3.3 Varying the total number of truly relevant covariates

In this experiment, we vary the total number of truly relevant covariates. In other words, we vary the number of variables that influence the outcome Y. In the context of a simple linear model, this is equivalent to increasing the number of covariates with non-zero coefficients. Specifically, we will consider the case where we have $n = 1000$ rows of data and $p = 1000$ covariates, as in previous setups. We will then iterate through the set $\{50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$ to vary the

number of truly relevant covariates in the model and evaluate how this impacts FDR control and the power score. Consider the following results plots.

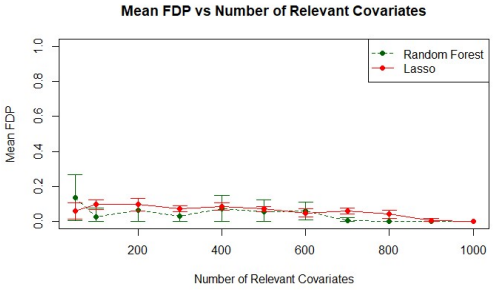

Figure 5.7: A plot of the mean proportion of false discoveries as the number of relevant covariates is varied

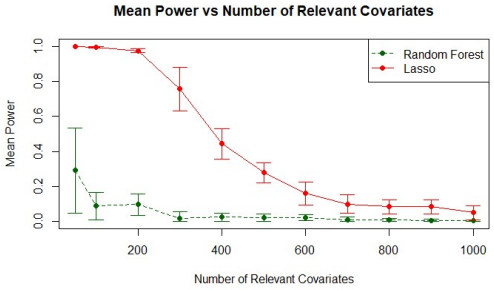

Figure 5.8: A plot of the mean power score as the number of relevant covariates is varied

From the above plots, we observe some interesting behaviour, particularly in terms of our power. In figure 5.7, we see a steady control of the false discoveries around 0.1, our nominal rate of false discovery in this experiment. Notice, as we increase the number of truly relevant variables, the false discovery proportion tends closer to 0. This can be justified by considering the extreme cases for our number of relevant variables. Recall, the false discovery proportion is the proportion of variables considered relevant by the model that are null in reality. In the case where the number of truly relevant variables is 1000 (i.e. all the variables in the model are relevant), any variable deemed relevant by the model is indeed truly relevant. Essentially, the number of truly null variables in the model is 0 and thus so will the number of false discoveries and the FDP. However, in the other cases where the number of truly relevant covariates is less than the total number of covariates in the model, we are introducing the possibility for false discoveries to be made. When fewer true positives are present, it becomes increasingly likely that the the model selects false positives; as discussed in section 4.3.

In figure 5.8, we observe the power decreasing as we increase the number of truly relevant variables we involve within the model. Recall, the power is the number of truly relevant selected variables divided by the total number of truly relevant variables. Clearly, if we increase the number of truely relevant variables within the model, the denominator of this fraction will increase. Therefore, in order to maintain a high level of power, the number of (truly relevant) covariates we select must also increase. That is, the number of covariates who's corresponding knockoff statistic surpasses the threshold,T, must increase. However, the data-dependent nature of T means that this is not possible. Recall the definition of T.

$$T = \min\{t \in \{|W_1|, ..., |W_p|\} : \hat{\text{FDP}}(t) = \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}} \leq q\}$$

This threshold is designed with the intention to keep an approximation of the false discovery proportion below a certain level, q, whilst also providing the most lenient threshold to maximise power. Let us consider the cases of 50 truly relevant variables and 1000 truly relevant variables. In both cases, we have set the $q = 0.1$, our fixed constraint. A truly relevant variable X_j is expected to have a larger knockoff statistic W_j by construction. For example, in the case where we have 1000 truly relevant variables, we would expect to see high values of W_j for all 1000 covariates. As a result, the threshold T, which is the smallest value of $|W_j|$ that ensures FDP is controlled at a fixed level q, must also increase. In order to select more variables, T would need to be reduced. However, reducing T could lead to a rise in FDP as more null variables would be included, potentially violating the q-constraint (importantly, the model is unaware of the ground truth; the number of truly relevant variables and so is not willing to risk violating the q-constraint despite the fact that all the variables are relevant in this case). Given that the model prioritizes controlling FDP (to avoid false positives) over maximizing power (selecting all true positives), reducing T is not possible. Therefore, to maintain the fixed FDP

constraint, the only option is to sacrifice power, meaning fewer truly relevant variables are selected. Consider the following results taken from a single iteration of various cases.

Number of Relevant Covariates	T	FDP	Power
50	0.0310	0.107	1.000
300	0.075	0.129	0.740
1000	0.316	0.000	0.085

Table 5.1: A subset of results taken from a single iteration

Number of Relevant Covariates	T'	FDP	Power
50	0.025	0.153	1.000
50	0.020	0.286	1.000
300	0.065	0.162	0.760
300	0.055	0.188	0.777
1000	0.270	0.000	0.120
1000	0.220	0.000	0.156

Table 5.2: Results taken when we forcefully alter the value of T to T'

Clearly, in cases where the number of truly relevant covariates is very high, reducing the threshold value leads to an increase in power. However, in practice, knockoff methodology and variable selection are typically applied in high-dimensional settings, where the number of truly relevant covariates constitutes only a small subset of the entire set of covariates. In such cases, as demonstrated by the results in tables 5.1 and 5.2, the trade-off between power and false discovery rate becomes more apparent.

5.3.4 Miss-specification

In this experiment, we explore a one parameter family of departures from the Gaussian Distribution in our covariates. Here, we are constructing knockoffs as before where the algorithm "believes" that the data is Gaussian and constructs Gaussian knockoffs accordingly. However, in reality some of the covariates will be sampled from an alternative distribution. For example, we may choose a heavy-tailed non-Gaussian distribution with unit variance and standard mean such as a t-distribution (See figure 5.9). This a natural choice when considering misspecification as we try and simulate outlier data which Gaussian distributions approximate less well.

The process of constructing covariates will work as follows. Suppose we are in the process of constructing one of our covariates X_j where $j \in \{1, 2, ..., p\}$. Previously, this was sampled from a standard normal distribution but now, with probability λ, X_j is sampled from the alternative distribution, $\mathcal{A}$. Formally, we have

$$P_\lambda(X_j) = (1 - \lambda)\mathcal{N}(0, 1) + \lambda\mathcal{A},$$

We would expect that as we increase the value of λ, making it increasingly likely that the data is sampled from the alternative distribution $\mathcal{A}$, the model's ability to correctly detect true covariates and disregard null covariates may worsen as the model assumes that the data follows a normal distribution $\mathcal{N}(0, 1)$, but as λ increases, the data becomes more heavily influenced by the non-Gaussian distribution $\mathcal{A}$, leading to more misinterpretations of the covariates' true structure. However, the major theoretical results of the knockoff methodology states that we get successfully FDR control when no miss-specification is occurring. More importantly, the theory does not implicitly suggest that the framework will not support FDR control under miss-specification. Note, for these experiments, we have only generated statistics from the lasso model due to computation constraints. Consider the following results.

We consider the results of the misspecification experiment using a T-distribution with varying degrees of freedom (heavier tails). Consider Figure 5.10: under no misspecification ($\lambda = 0$), the false discovery proportion (FDP) remains tightly bound around the nominal rate $q = 0.1$, indicating proper FDR control. However, for $\lambda > 0$ (introducing misspecification), FDP varies. While some cases, such as the T-distribution with 2 degrees of freedom, maintain reasonable FDR control even with high misspecification, others exhibit spikes in false discoveries at different levels of λ, highlighting

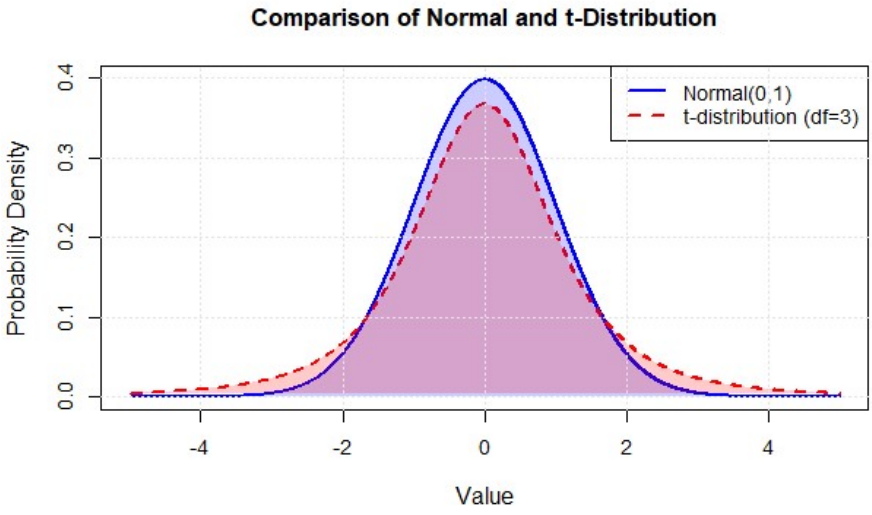

Figure 5.9: The probability distribution functions for the standard normal distribution and the t-distribution with 3 degrees of freedom

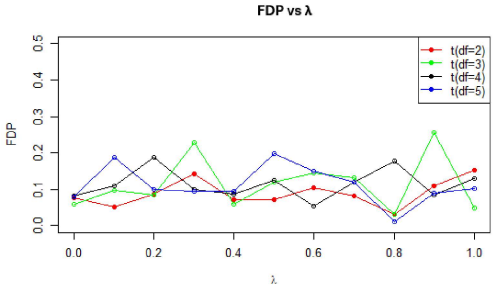

Figure 5.10: A plot of the false discovery proportion of each test as the degree of misspecification is varied under a T-distribution

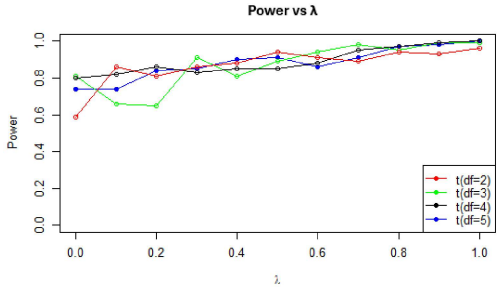

Figure 5.11: A plot of the power score of each test as the degree of misspecification is varied under a T-distribution

the method's sensitivity to distributional assumptions. Regardless, in most cases, the level of false discovery remains close to the nominal rate indicating that the method generally maintains good control over false discoveries, even under varying degrees of misspecification.

Figure 5.11 displays some interesting behaviour. We observe a general trend of improved power as λ increases, introducing greater misspecification into the model. This trend continues until all covariates, originally following a t-distribution, are fully miss-specified as normally distributed. The steady increase in power may be attributed to the nature of heavy-tailed distributions, which introduce covariates with more extreme values. If these covariates are truly relevant, they can enhance signal detection despite the mismatch in assumed distribution. The presence of extreme values provides disproportionately large amounts of information about the relationship between a covariate and the target, as they dominate other terms in the linear summation, making it easier to identify relevant covariates. Although a Gaussian distributional assumption is made, the knockoff methodology may exhibit enough robustness to accommodate this misspecification and in fact, the results suggest that they perform better in terms of power. Therefore, Although the fixed constraint of FDR control within the knockoff framework is partially violated, the improvement in power suggests that, even when the exact distributions of the covariates are uncertain, variable selection via knockoffs remains a potentially viable approach.

We now perform some cross-sectional analysis where we fix λ (and so the level of miss-specification)

and then vary q. That way, we get a sense of how well we are able to ensure false discovery control for varying levels of miss-specification under heavier tailed distributions.

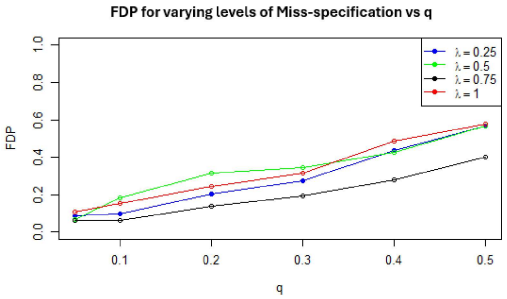

Figure 5.12: T-distribution cross-sectional analysis: how well does the knockoff filter control false discoveries below q for varying levels of miss-specification?

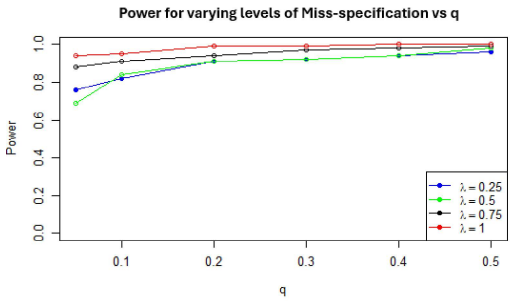

Figure 5.13: T-distribution cross-sectional analysis: how well does the knockoff filter optimise power scores for varying levels of miss-specification?

As before, we obtain approximate FDR control for each of the various degrees of misspecification. Figure 5.12 mirrors the behaviour of 5.10, as we observe no clear relationship between tight control of false discoveries and the level of misspecification within the model. Our results show that the tightest FDR control is associated with $\lambda = 0.75$. Figure 5.11 further consolidates our conclusions about heavier-tailed distributions improving the power of the tests. As we increase the value of λ, more and more of our data is sampled from the heavier-tailed distribution, leading to an increase in power. Additionally, as we raise q (thereby making the relevance threshold T less strict), power approaches 1.

To conclude this experiment, we highlight the practical benefits of improved power scores and maintained FDR control despite misspecification. In real-world scenarios, it's reasonable to expect outliers and inconsistencies. The ability of the knockoff framework to maintain false discovery rate control and improve its ability to detect true covariates, even when data are misclassified, is a significant benefit. This suggests that the methodology remains effective when the underlying data distribution deviates from theoretical assumptions or when the data is noisy. Such robustness to model misspecification makes the knockoff variable selection framework a valuable tool, especially in applied settings where data rarely conform to idealised conditions like normality.

Let us consider the opposite case, when data is uniformly spread. To control confounding effects, we specify a truncated uniform distribution with the same mean and variance as a standard normal distribution. That is, a continuous uniform distribution on $[-\sqrt{3}, \sqrt{3}]$. Consider the results below.

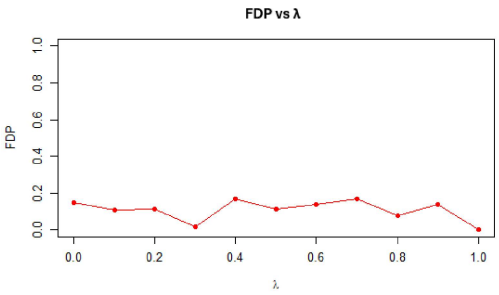

Figure 5.14: A plot of the false discovery proportion as the degree of misspecification is varied under a uniform distribution

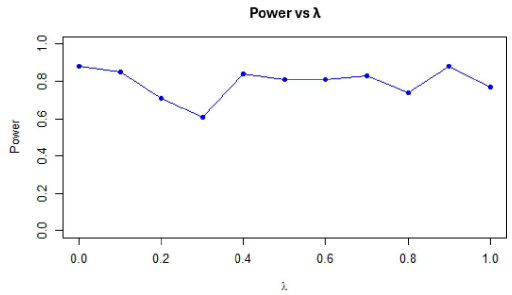

Figure 5.15: A plot of the power score as the degree of misspecification is varied under a uniform distribution

In terms of false discoveries, consider figures 5.14 and 5.16. We observe approximately consistent control across all levels of misspecification and values of q, as seen previously. As discussed in earlier

chapters, the knockoff method relies on the symmetry of knockoff statistics rather than strict distributional assumptions. Consequently, the design of these statistics—and the knockoff filter itself—helps prevent the false discovery proportion from increasing excessively, even in the presence of this misspecification. We also see less of a steady level of power as we raise λ in figure 5.15. The uniform distribution ensures an even spread of values across its range, meaning that increasing λ does not highlight relevant covariates in the same way it would with heavy-tailed distributions. Increasing the level of misspecification simply spreads the covariates more evenly, without concentrating mass in the tails. As a result, for a truly relevant covariate, you are less likely to observe more extreme values that would help differentiate it from noise. Consequently, the coefficient assigned to the relevant covariate may not be as large compared to its knockoff, making it less likely for the associated knockoff statistic to surpass the threshold T.

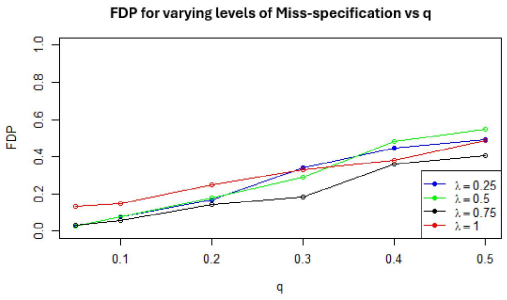

Figure 5.16: Uniform distribution cross-sectional analysis: how well does the knockoff filter control false discoveries below q for varying levels of miss-specification?

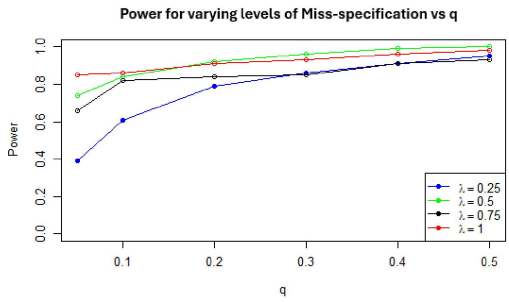

Figure 5.17: Uniform distribution cross-sectional analysis: how well does the knockoff filter optimise power scores for varying levels of miss-specification?

5.4 Biomarker detection simulation study: Experimental design

As discussed in Chapter 1, we now examine the knockoff-based variable selection approach within a causal framework. We want to be able to establish the impact (in terms of our target variable, some key health indicator for example) that a treatment will have on a patient. We previously outlined the difference between predictive (impacts the target as a result of treatment) and prognostic variables (impacts the target irrespective of treatment). Recall the model previously discussed is given by

$$\mathbb{E}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}, T = t] = f(t) + h(\mathbf{x}) + g(\mathbf{x}) = f(t) + h(\mathbf{x}_{prog}) + g(\mathbf{x}_{pred})t,$$

where

- $T = 1$ if the subject is in the treatment group and $T = 0$ if the subject is in the control.
- $h(\mathbf{x}) = h(\mathbf{x}_{prog})$ is the component of the expectation determined by the prognostic covariates.
- $g(\mathbf{x}) = g(\mathbf{x}_{pred})$ is the component that captures the interaction between the predictive covariates and the treatment T, showing how the treatment affects the outcome.
- $f(t)$ is the homogenous treatment effect i.e. the change in outcome due to the treatment, regardless of specific individual covariates. For example, in a study where the treatment is medicine, $f(t)$ would capture the overall improvement in health metrics observed by all individuals taking the medicine, regardless of their biomarker profiles.

If the ultimate aim is to measure the impact that the treatment will have on an individual, this is equivalent to computing the difference in the value of the target variable for $T \in \{0, 1\}$, which is defined in the biomarker paper as $Y(1) - Y(0)$. This value is exactly a measure of the impact of the

interaction between our treatment and the predictive variables. Clearly, we run into a problem here; for any given patient, you only obtain either $Y(0)$ or $Y(1)$ as they are either placed in the control group for which $T = 0$ or the test group for which $T = 1$. Therefore, obtaining $Y(1) - Y(0)$ is not possible in practice. However, we can write the following expected value for the difference.

$$\begin{aligned}\mathbb{E}[Y(1) - Y(0) \mid X = \mathbf{x}] &= \mathbb{E}[Y(1) \mid X = \mathbf{x}] - \mathbb{E}[Y(0) \mid X = \mathbf{x}] \\ &= \mathbb{E}[Y \mid T = 1, X = \mathbf{x}] - \mathbb{E}[Y \mid T = 0, X = \mathbf{x}] \\ &= f(1) + h(\mathbf{x}) + g(\mathbf{x}) - h(\mathbf{x}) - f(0) \\ &= f(1) - f(0) + g(\mathbf{x})\end{aligned}$$

The emphasis here is that we can write the Conditional (on the covariates) Average Treatment Effect (CATE) as a function of solely the predictive variables. We can therefore establish which covariates (biomarkers) influence the treatment outcome.

We need to consider how we would implement this practically so that we can perform some simulated data analysis. We use a *transformed outcome Knockoff filter* [20] to construct a target variable, Y^* in such a way that its conditional expectation matches the CATE. This process of moment matching ensures that the estimates of the treatment effects for individuals aligns with the true conditional expectation of the treatment effect. That way, by identifying the variables that influence Y^*, we are implicitly identifying those that influence the treatment effect, $Y(1) - Y(0)$. Formally, we define Y^* as follows.

$$Y^* := T \cdot \frac{Y}{\pi(X)} + (1 - T) \cdot \frac{-Y}{\pi(X)}$$

where $\pi(X) = P(T = 1|X)$. Since we are performing a randomized control trial, it must be that $\pi(X) = P(T = 1|X) = P(T = 1) = \frac{1}{2}$. Therefore, we have that

$$Y^* := T \cdot \frac{Y}{1/2} + (1 - T) \cdot \frac{-Y}{1/2} = 2TY - 2(1 - T)Y = \begin{cases} 2Y, & \text{if } T = 1 \\ -2Y, & \text{if } T = 0 \end{cases}$$

Using the transformed target variable Y^*, we have that

$$\begin{aligned}\mathbb{E}[Y^* \mid X = \mathbf{x}] &= \mathbb{E}[Y^* \mid X = \mathbf{x}, T = 0] \cdot P(T = 0) + \mathbb{E}[Y^* \mid X = \mathbf{x}, T = 1] \cdot P(T = 1) \\ &= \mathbb{E}[2Y \mid X = \mathbf{x}, T = 0] \cdot \frac{1}{2} + \mathbb{E}[-2Y \mid X = \mathbf{x}, T = 1] \cdot \frac{1}{2} \\ &= \mathbb{E}[Y \mid X = \mathbf{x}, T = 0] - \mathbb{E}[Y \mid X = \mathbf{x}, T = 0] \\ &= \mathbb{E}[Y(1) - Y(0) \mid X = \mathbf{x}]\end{aligned}$$

Essentially, in the causal framework, we are implementing the same methodology with the same goal of variable selection whilst controlling the level of false discoveries but now honing our interest in on solely predictive variables. In the traditional setting, we were interesting in finding any variable that impacted Y i.e. the prognostic variables. Now, we are interesting in finding variables that impact the treatment effect, $Y(1) - Y(0)$. Practically, this could be thought of as the improvement in health an individual receives due to a particular treatment.

5.4.1 Data generation

We have now described the causal framework in which to perform predictive variable selection via knockoffs. As in section 5.2, we begin by outlining the process of generating the data such that we can perform our analysis. The construction of our covariate data matrix $\mathbf{X}$ and the corresponding knockoffs is the same as in the simple simulation study so we will not repeat the details here.

1. Generate the vector $\mathbf{T} = (T_1, T_2, ..., T_n)$
   For row $j \in \{1, 2, .., n\}$ in our dataset, which can be thought patient j, we must assign a treatment indicator according to the RCT. If patient j is in the control group, we set $T_j = 0$. Likewise, $T = 1$ if the patient is in the treatment group. It should be that half of the elements of $\mathbf{T}$ are 0 and the other half of them are 1.

2. Generate Target Data $\mathbf{Y}$
   The generation of the target data for each row in the dataset differs in comparison to the simple simulation study where the target data was constructed as a linear combination of a subset of the covariates. As aforementioned, we have that

$$\mathbb{E}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}, T = t] = f(t) + h(\mathbf{x}) + g(\mathbf{x})t$$

   As in the biomarker paper, we will construct the target data such that

$$\mathbb{E}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}, T = t] = 0.5t + \sum_{j \in S^{prog}} X_j + \theta_{pred} \sum_{j \in S^{pred}} X_j$$

   Where S^{pred} denotes the set of indices associated with predictive covariates, and S^{prog}, the set of indices associated with prognostic covariates. Here, $\theta_{pred} \in [0, 1]$, denotes the strength of heterogeneity of the treatment effect [20]. By this, we mean how impactful the treatment's interaction with the predictive covariates is in influencing the target. If there were no heterogeneous treatment effect, meaning the treatment has the same effect across the entire population, then $\theta_{pred} = 0$. The effect of the treatment would be purely homogenous, as implemented by $f(t) = 0.5t$.

As in the simple simulation study, we will set $n = p = 1000$ to simulate a high dimensional scenario where it is not the case that $n >> p$. Once again, we are performing predictive variable selection meaning we want to locate the covariates that impact the target as a result of taking the treatment. We are not concerned with finding prognostic variables (unless they also happen to be predictive). Therefore, for simplicity, we will assume that all our prognostic variables are indeed predictive. In these experiments, we will have 100 predictive covariates which we randomly select out of the total 1000.

5.5 Biomarker detection simulation study: Key experiments

We begin by outlining the experiments that we will perform within the causal framework.

1. Varying nominal FDR level, q. Once again, this experiment is necessary in order to justify the implementation of the knockoff filter within the causal framework. For this experiment, we generate statistics from a lasso model and a causal forest. The causal forest is an extension of a random forest designed for the causal framework [20] [12].

2. Varying the magnitude of the heterogenous treatment effect, θ_{pred}. In the context of a genetic study, this experiment considers the scenario where subgroups of the population will receive an enhanced level of treatment effect. In other words, this test considers the case where the unique makeup of an individual's genotype leads to differing levels of improvement in terms of Y. For this experiment, since our data only contains linear interactions, we only generate statistics from the lasso model.

5.5.1 Varying the nominal FDR level, q

In this experiment, as in the simple simulation study, we vary the nominal FDR level to assess how well we are still able to control the false discovery rate of predictive variables whilst trying to obtain as many of the truly relevant covariates as possible. As previously discussed, the constraint on false discoveries is prioritized over identifying all truly relevant covariates. In other words, we are more focused on avoiding the false identification of null biomarkers as true ones than on identifying every true biomarker. In the context of treatment effects, we would prefer not to administer treatment to individuals lacking the necessary biomarkers for a substantial treatment effect, rather than risk withholding treatment from those who would benefit. That way, we ensure optimal treatment allocation. Consider the following results from simulations when we vary q and perform repeated experiments for each value of $q \in \{0.01, 0.05, 0.10, 0.20, 0.5\}$. Throughout this experiment, we have that $\theta_{pred} = 1$

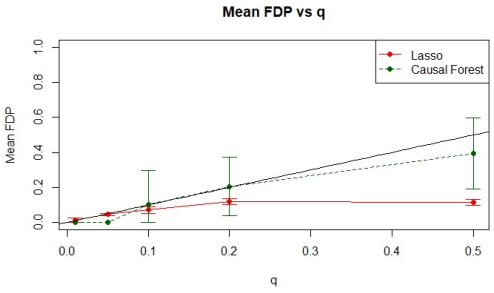

Figure 5.18: Plot of mean proportion of false discoveries as we vary q in the causal framework

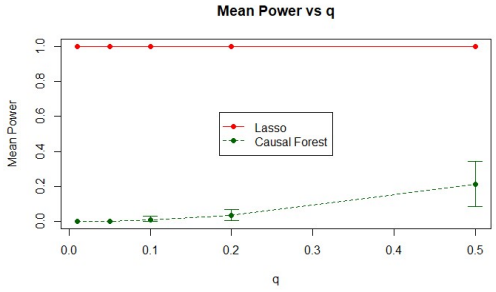

Figure 5.19: Plot of the mean power score as we vary q in the causal framework

In the causal framework, we observe strong FDR control. As before, the mean false discovery proportion for each test remains below the accepted threshold according to q. The diagonal black line in 5.18 is designed to show this. Additionally, the power plot, figure 5.19 displays how each test is successfully and consistently able to acquire all the truly relevant biomarkers that influence the treatment effect. This is a positive result, as it suggests that we can impose strict FDR control whilst still acquiring all the true covariates. As expected, since we only model linear interactions, our lasso model considerably outperforms our causal forest model. A natural follow-up question is what happens when we decrease the predictive signal θ_{pred}, weakening the treatment's heterogeneous effect. While we observed positive results for $\theta_{pred} = 1$, we wish to establish how the model performs when the treatment's effect becomes less pronounced.

5.5.2 Varying the magnitude of the heterogenous treatment effect, θ_{pred}

In this experiment, we vary the magnitude of the heterogeneous treatment effect, which controls how strongly the predictive variables influence the treatment's impact on the outcome. As we increase the value of θ_{pred}, the treatment effect becomes more pronounced, making the predictive covariates more identifiable. Consequently, we would expect power to increase. Consider the results in figures 5.20 and 5.21. Clearly, we see consistent FDR control for the predictive variables when the nominal FDR level is set to the default value of 0.1. We see that as θ_{pred} increases, the ability of the knockoff filter to obtain true covariates improves and in fact, we see that the heterogenous treatment effect can be closer to 0 (see the case when $\theta_{pred} = 0.2$ and we still obtain high power values.

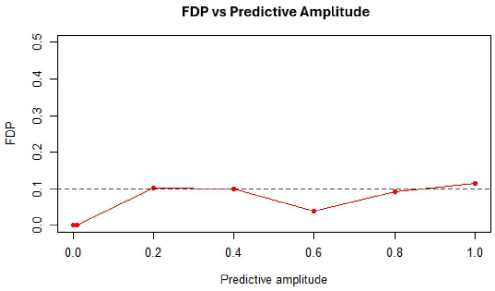

Figure 5.20: Plot of θ_{pred} against FDP

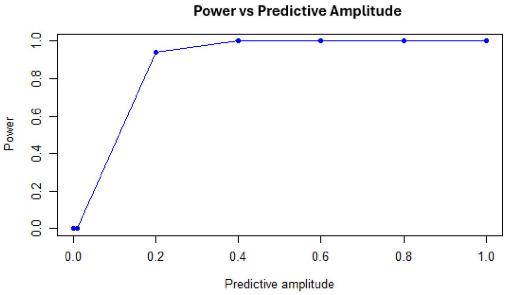

Figure 5.21: Plot of θ_{pred} against power

Now that we have considered the impact of the heterogeneous treatment effect (i.e. the predictive amplitude θ_{pred}, it is also important to consider that different predictive biomarkers may have varying levels of influence on the treatment effect. To explore this, we categorize the predictive biomarkers into three groups based on their strength: highly predictive, moderately predictive, and weakly predictive

biomarkers. This categorisation simulates the real-world scenario where predictive biomarkers will differ in terms of the magnitude of their influence on the treatment effect. To perform this experiment, we will define $\theta^h_{pred}, \theta^m_{pred}, \theta^l_{pred}$ as constants that determine the heterogenous impact of highly, moderate and weakly predictive biomarkers respectively. A percentage of the total predictive covariates will be assigned to each group. We perform the following experiments.

1. Varying the values of $\theta^h_{pred}, \theta^m_{pred}, \theta^l_{pred}$
   Here, we will vary the predictive signal amplitudes of each subgroup whilst keeping the percentage of covariates in each group constant. In particular $\theta^h_{pred} \in \{0.5, 0.6, 0.7\}$, $\theta^m_{pred} \in \{0.2, 0.3, 0.4\}$, $\theta^l_{pred} \in \{0.04, 0.08, 0.12\}$. For this experiment, 50% of the predictive biomarkers will be considered moderately predictive, 25% considered highly predictive and the remaining 25% weakly predictive. Our results given in table 5.3 concur with our previous results; we see that there is a steady control of the level of false discovery across various different permutations of predictive amplitudes for each subgroup. In terms of power, we see the expected result that the framework was very good at selecting those predictive covariates with high predictive amplitude, θ^h_{pred} and less good at selecting those whose corresponding predictive amplitude was lower, θ^l_{pred}. In our table, we see that the model is successful in it's selection of covariates who considerably interact with the treatment, leading to a larger change in treatment effect and likewise, in the case of those predictive covariates who's impact is smaller, a smaller proportion is selected, naturally being the cause of the relatively lower scores.

θ^h_{pred}	θ^m_{pred}	θ^l_{pred}	FDP	Power	Proportion of highs selected	Proportion of moderates selected	Proportion of lows selected
0.50	0.20	0.04	0.11	0.75	1	0.94	0.12
0.5	0.20	0.08	0.078	0.84	1	0.90	0.56
0.5	0.3	0.08	0.024	0.82	1	0.98	0.32
0.6	0.3	0.08	0.078	0.83	1	0.98	0.36
0.6	0.4	0.12	0.04	0.91	1	1	0.64
0.7	0.4	0.12	0.05	0.90	1	1	0.50

Table 5.3: A subset of results for varying θ^h_{pred}, θ^m_{pred}, and θ^l_{pred}

2. Varying the percentage of the predictive covariates assigned to each group.
   We now alter the percentage of covariates placed into each of the three subgroups while keeping the predictive signal amplitudes constant. Specifically, we experiment with allocations such as 10%/80%/10%, 25%/50%/25% and 33%/33%/33%. Our results are presented below.

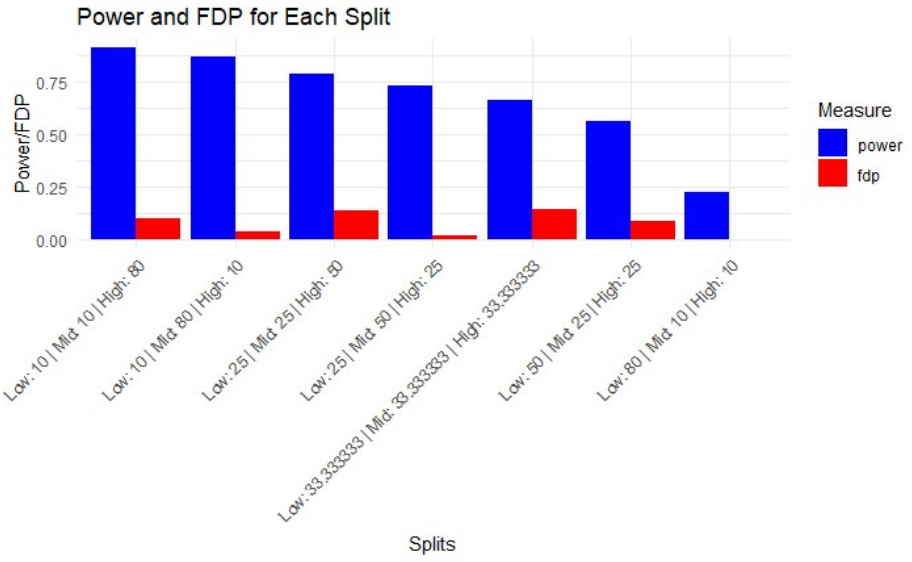

Figure 5.22: Results for varying percentages of covariates assigned to each subgroup.

Once more, we that there is a correlation associated between high predictive amplitudes and the ability of the knockoff filter to select the predictive biomarkers corresponding to such predictive amplitudes whilst still ensuring a control on the false discovery rate (In these experiments, $q = 0.1$). From table 5.3 and figure 5.22, we see that even when the predictive biomarkers whose predictive amplitude is set to the high value of θ^h_{pred} only make up 10% of the total predictive biomarkers, they are all correctly selected. We also find that the model generally struggles to find the predictive biomarkers who have weaker predictive amplitudes which is to be expected.

5.6 Results summary

Our experiments in the traditional knockoff framework have provided additional evidence for the validity of the knockoff framework with various parameter permutations. Our initial experiment which ensured that the framework was indeed able to control the proportion of false discoveries in the traditional framework proved successful with high levels of empirical performance, demonstrating that the statistical foundations of the methodology held in practice. Further tests demonstrated the applicability of the method in cases where traditional variable selection methods inherently struggle, specifically, cases where the volume of data is comparable to the dimensionality of the data. Following this, we saw a potential drawback of knockoffs when we varied the number of truly relevant covariates in our model. The rigidity of the framework in its control of false discoveries means that power will inevitably suffer. In practice, this is unlikely to be a problem due to the nature of the problem we attempt to solve within variable selection. In other words, we would expect that the number of truly relevant covariates is a small fraction of the total number of covariates (for example, in genetic studies [4]). Regardless, a bespoke condition or unique handling of the filter in cases where additional knowledge about the number of relevant covariates could ensure power scores remain high.

Finally, we saw that the knockoff filter performed well under miss-specification, emphasising its robustness in cases where the assumptions of the methodology are violated. In addition, we saw rigorous FDR control in the causal framework, further validating the theoretical underpinnings of the knockoff method, suggesting its potential applicability of high-dimensional biomarker discovery. Varying the heterogeneity of the treatment provided further results that aligned with the intuition that knockoffs saw improved performance when the magnitude of this effect was more apparent.

Chapter 6

Discussion

6.1 Discussion

We begin our discussion by considering potential avenues for further work. Whilst this paper has provided a deep explanation of the methodology of knockoffs in controlling FDR, it is important to recognise alternative approaches to implementing knockoffs. These include varying the models used to relate the target to the covariates and the corresponding feature importance statistics they generate. This is particularly relevant in the Model-X paradigm, where machine learning models can help bridge the knowledge gap between covariates and the target variable.

Additionally, various knockoff statistics may be more beneficial for specific variable selection tasks, with alternatives provided in the reference manual in [19]. In our experiments, we simplified the problem by making modelling assumptions that facilitated simulations, such as assuming covariate normality, which allowed for the straightforward generation of Gaussian Model-X knockoffs. Although this assumption is often well-founded, we did not directly explore the construction of knockoffs from other distributions, though we did assess how Gaussian knockoffs performed under misspecification.A generalized approach to constructing knockoffs within the Model-X framework remains an important area for future research, as noted in [2].

Further parameter variation in the future would allow us to better assess the performance of the knockoff filter in real-world scenarios such as the level of noise in the models, covariance between features and the introduction of non-linearity between covariates as in [20] to compare modelling procedures (in particular, how the performance of lasso fairs against non-parametric approaches). Lastly, while simulated data offers valuable insights, conducting experiments with real-world data will be crucial to assess the practical performance of the knockoff framework.

6.2 Conclusion

In this paper, we have provided alternative rigorous proofs and explanations of the key results that underpin the knockoff methodology. In doing this, we have attempted to build upon the foundational work and methodology provided by Barber and Candés [1] and the further study by Candés which introduces the Model-X methodology [2]. By providing a breakdown of the construction of knockoffs (both Fixed-X and Model-X) as well as describing the process of generating their associated feature importance and knockoff statistics, we have laid out the process of performing variable selection with knockoffs from start to finish. Simulations were able to successfully demonstrate the applicability of the knockoff filter in its statistical guarantee of FDR control, our fixed constraint, whilst also demonstrating good levels of statistical power (although this was parameter-dependent). This displayed the robustness of the framework in cases of high-dimensionality (in which all our experiments were taken under).

In the broader context of biomarker research, the knockoff filter proved successful in extracting predictive variables (those who's impact on an individual's health is seen only as a result of treatment). This is arguably a more important result, original presented by Sechidis et al [20]. This result, combined with the high performance of knockoffs when working with finite volumes or high-dimensional data, suggests its real-world applicability in personalized medicine where a patient's genotype can indicate whether a particular treatment will result in a substantial improvement in their health.

Appendix A

Simulation study code

The code used for this study is available at the following GitHub repository: https://github.com/markmilner21/Variable-Selection-With-Knock-Offs. For additional documentation and code referenced in this study, please refer to sources [20] and [21] in the bibliography.

Appendix B

Bounded increments

In this appendix entry, we find a constant C such that $|M_{k+1} - M_k| \leq C$ for all $k \in \{0, ..., m-1\}$. Recall, $Z_1, \ldots, Z_m$ corresponding the signs of the $W_{(1)}, \ldots, W_{(m)}$, respectively. Furthermore, for each $k \in \{0, \ldots, m-1\}$ and $\omega \in \{-1, +1\}$, we let

$$V_k(\omega) := \frac{m - k + \omega \sum_{j=k+1}^{m} Z_j}{2},$$

We define a filtration $(\mathcal{G}_\ell)_{\ell=0}^m$ by letting $\mathcal{G}_\ell$ denote the smallest σ-algebra such that the random variables $\{|W_{(\ell)}|\}_{\ell=1}^m \cup \{V_0(-1), V_0(1)\} \cup \{Z_j\}_{j \leq \ell}$ are $\mathcal{G}_\ell$-measurable. Applied conditionally on $\{|W_{(\ell)}|\}_{\ell=1}^m$, the process $(M_k)_{k=0}^{m-1}$ defined by

$$M_k := \frac{V_k(1)}{1 + V_k(-1)},$$

for $k \in \{0, \ldots, m-1\}$ is a martingale with respect to $(\mathcal{G}_\ell)_{\ell=0}^m$.

Clearly,

$$\begin{aligned}
|M_{k+1} - M_k| &= \left| \frac{V_{k+1}(1)}{1 + V_{k+1}(-1)} - \frac{V_k(1)}{1 + V_k(-1)} \right| \\
&= \left| \frac{(m-k-1+\sum_{j=k+2}^{m} Z_j)/2}{1 + (m-k-1-\sum_{j=k+2}^{m} Z_j)/2} - \frac{(m-k+\sum_{j=k+1}^{m} Z_j)/2}{1 + (m-k-\sum_{j=k+1}^{m} Z_j)/2} \right| \\
&= \left| \frac{(m-k-1+\sum_{j=k+2}^{m} Z_j)}{2 + (m-k-1-\sum_{j=k+2}^{m} Z_j)} - \frac{(m-k+\sum_{j=k+1}^{m} Z_j)}{2 + (m-k-\sum_{j=k+1}^{m} Z_j)} \right| \\
&= \left| \frac{(m-k-1+\sum_{j=k+1}^{m} Z_j - Z_{k+1})}{2 + (m-k-1-\sum_{j=k+1}^{m} Z_j + Z_{k+1})} - \frac{(m-k+\sum_{j=k+1}^{m} Z_j)}{2 + (m-k-\sum_{j=k+1}^{m} Z_j)} \right|
\end{aligned}$$

Now, if $Z_{k+1} = 1$, then

$$\begin{aligned}
|M_{k+1} - M_k| &= \left| \frac{(m-k-1+\sum_{j=k+1}^{m} Z_j - Z_{k+1})}{2 + (m-k-1-\sum_{j=k+1}^{m} Z_j + Z_{k+1})} - \frac{(m-k+\sum_{j=k+1}^{m} Z_j)}{2 + (m-k-\sum_{j=k+1}^{m} Z_j)} \right| \\
&= \left| \frac{(m-k-1+\sum_{j=k+1}^{m} Z_j - 1)}{2 + (m-k-1-\sum_{j=k+1}^{m} Z_j + 1)} - \frac{(m-k+\sum_{j=k+1}^{m} Z_j)}{2 + (m-k-\sum_{j=k+1}^{m} Z_j)} \right| \\
&= \left| \frac{-2}{2 + (m-k-\sum_{j=k+1}^{m} Z_j)} \right|
\end{aligned}$$

Since $\sum_{j=k+1}^{m} Z_j \leq m-k$, we must have that $2 + (m-k-\sum_{j=k+1}^{m} Z_j) \geq 2$. Therefore $|M_{k+1} - M_k| \leq 1$

Now, if $Z_{k+1} = -1$, then

$$
\begin{aligned}
|M_{k+1} - M_k| &= \left| \frac{(m-k-1+\sum_{j=k+1}^{m} Z_j - Z_{k+1})}{2+(m-k-1-\sum_{j=k+1}^{m} Z_j + Z_{k+1})} - \frac{(m-k+\sum_{j=k+1}^{m} Z_j)}{2+(m-k-\sum_{j=k+1}^{m} Z_j)} \right| \\
&= \left| \frac{(m-k-1+\sum_{j=k+1}^{m} Z_j + 1)}{2+(m-k-1-\sum_{j=k+1}^{m} Z_j - 1)} - \frac{(m-k+\sum_{j=k+1}^{m} Z_j)}{2+(m-k-\sum_{j=k+1}^{m} Z_j)} \right| \\
&= \left| \frac{(m-k+\sum_{j=k+1}^{m} Z_j)}{(m-k-\sum_{j=k+1}^{m} Z_j)} - \frac{(m-k+\sum_{j=k+1}^{m} Z_j)}{2+(m-k-\sum_{j=k+1}^{m} Z_j)} \right|
\end{aligned}
$$

By the triangle inequality, we have that

$$
\begin{aligned}
|M_{k+1} - M_k| &\leq \left| \frac{m-k+\sum_{j=k+1}^{m} Z_j}{m-k-\sum_{j=k+1}^{m} Z_j} \right| + \left| \frac{m-k+\sum_{j=k+1}^{m} Z_j}{2+(m-k-\sum_{j=k+1}^{m} Z_j)} \right| \\
&= \frac{m-k+\sum_{j=k+1}^{m} Z_j}{m-k-\sum_{j=k+1}^{m} Z_j} + \frac{m-k+\sum_{j=k+1}^{m} Z_j}{2+(m-k-\sum_{j=k+1}^{m} Z_j)} \\
&\leq \frac{2(m-k)}{1} + \frac{2(m-k)}{2} \\
&= 3(m-k) \\
&\leq 3m
\end{aligned}
$$

Since we assume $m >= 1$, we have that $|M_{k+1} - M_k| \leq 3m$. It is possible to find stricter bounds but for our purposes, this bound is sufficient.

Bibliography

[1] Rina Foygel Barber and Emmanuel J Candès. "Controlling the false discovery rate via knockoffs". In: *The Annals of statistics* (2015), pp. 2055–2085.

[2] Emmanuel Candes et al. "Panning for gold:'model-X'knockoffs for high dimensional controlled variable selection". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80.3 (2018), pp. 551–577.

[3] Emmanuel Candes et al. "Panning for gold:'model-X'knockoffs for high dimensional controlled variable selection Supplementary material". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* (2018).

[4] Emmanuel Candès. *The Knockoffs Framework: New Statistical Tools for Replicable Selections*. Institute for Pure & Applied Mathematics (IPAM), https://www.youtube.com/watch?v=NuVBHXYBC4k&t=1022s. 2025.

[5] Jinyuan Chang, Xiaohui Chen, and Mingcong Wu. "Central limit theorems for high dimensional dependent data". In: *Bernoulli* 30.1 (Feb. 2024). ISSN: 1350-7265. DOI: 10.3150/23-bej1614.

[6] Wikipedia Contributors. *Optional stopping theorem*. Accessed: 2025-03-21. 2025. URL: https://en.wikipedia.org/wiki/Optional_stopping_theorem.

[7] Kristen Emery and Uri Keich. *Controlling the FDR in variable selection via multiple knockoffs*. 2019. DOI: 10.48550/ARXIV.1911.09442.

[8] Ron Freiwald. *Orthogonal Decomposition Theorem*. Accessed: 2024. 2016. URL: https://www.math.wustl.edu/%7Efreiwald/309FL14L37s.pdf.

[9] Jiti Gao et al. "High Dimensional Correlation Matrices: The Central Limit Theorem and Its Applications". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 79.3 (June 2016), pp. 677–693. ISSN: 1467-9868. DOI: 10.1111/rssb.12189.

[10] Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009. Chap. 15.

[11] Tao Jiang, Yuanyuan Li, and Alison A Motsinger-Reif. "Knockoff boosted tree for model-free variable selection". In: *Bioinformatics* 37.7 (Sept. 2020). Ed. by Jonathan Wren, pp. 976–983. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btaa770.

[12] Michael Lechner. "Modified causal forests for estimating heterogeneous causal effects". In: *arXiv preprint arXiv:1812.09487* (2018).

[13] *Linear Algebra Properties of Transposes*. Accessed: 2024. URL: https://www.web-formulas.com/Math_Formulas/Linear_Algebra_Properties_of_Transposes.aspx.

[14] Matthew Macauley. *Lecture 7.3: Gram Matrices*. Accessed: 2024. URL: https://www.math.clemson.edu/~macaule/classes/f20_math8530/slides/math8530_lecture-7-03_h.pdf.

[15] Péter Medvegyev. *Stochastic Processes: A Very Simple Introduction*. Accessed: 2025-03-13. 2009. URL: https://web.archive.org/web/20150403125546/http://medvegyev.uni-corvinus.hu/St1.pdf.

[16] Amr Essam Mohamed. *High-Dimensional Variable Selection via Knockoffs Using Gradient Boosting*. Western Michigan University, 2023.

[17] Christoph Molnar et al. "Model-agnostic feature importance and effects with dependent features: a conditional subgroup approach". In: *Data Mining and Knowledge Discovery* 38.5 (Jan. 2023), pp. 2903–2941. ISSN: 1573-756X. DOI: 10.1007/s10618-022-00901-9.

[18] R Muthukrishnan and R Rohini. "LASSO: A feature selection technique in predictive modeling for machine learning". In: *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*. IEEE, Oct. 2016, pp. 18–20. DOI: 10.1109/icaca.2016.7887916.

[19] Evan Patterson and Matteo Sesia. *knockoff: The Knockoff Filter for Controlled Variable Selection*. Dec. 2014. DOI: 10.32614/cran.package.knockoff.

[20] Konstantinos Sechidis, Matthias Kormaksson, and David Ohlssen. "Using knockoffs for controlled predictive biomarker identification". In: *Statistics in Medicine* 40.25 (2021), pp. 5453–5473.

[21] M. Sensia. *Variable Selection with Knockoffs*. Retrieved from https://web.stanford.edu/group/candes/knockoffs/index.html (Accessed: 2024).

[22] Asher Spector and William Fithian. *Asymptotically Optimal Knockoff Statistics via the Masked Likelihood Ratio*. 2022. DOI: 10.48550/ARXIV.2212.08766.

[23] Gerhard Tutz and Harald Binder. "Generalized Additive Modeling with Implicit Variable Selection by Likelihood-Based Boosting". In: *Biometrics* 62.4 (June 2006), pp. 961–971. ISSN: 1541-0420. DOI: 10.1111/j.1541-0420.2006.00578.x.

[24] Quang H Vuong. "Likelihood ratio tests for model selection and non-nested hypotheses". In: *Econometrica: journal of the Econometric Society* (1989), pp. 307–333.

[25] Meredith L. Wallace et al. "Use and misuse of random forest variable importance metrics in medicine: demonstrations through incident stroke prediction". In: *BMC Medical Research Methodology* 23.1 (June 2023). ISSN: 1471-2288. DOI: 10.1186/s12874-023-01965-x.

[26] Wikipedia contributors. *Markov blanket — Wikipedia, The Free Encyclopedia*. [Online; accessed 2-March-2025]. 2024. URL: https://en.wikipedia.org/w/index.php?title=Markov_blanket&oldid=1223926674.

[27] G. A. Young and Mark E. Johnson. "Multivariate Statistical Simulation." In: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 151.1 (1988), p. 229. ISSN: 0964-1998. DOI: 10.2307/2982203.