KU LEUVEN

FACULTY OF SCIENCES

# Knockoffs for Variable Selection in Models with a Discrete Response

**Juli DEMA**

Supervisor: Prof. Dr., G. Claeskens
KU Leuven

Cosupervisor: *Dr. J. Zhou*
KU Leuven

Master thesis submitted in fulfillment
of the requirements for the degree in
Master of Science in Theoretical Statistics and Data Science

Academic year 2021-2022

# Preface

Practitioners are faced with the difficult variable selection problem in virtually every application. To further complicate matters, there exists a vast sea of variable selection procedures to choose from. Oftentimes, these procedures offer some asymptotic performance guarantees however, when applied in a practical, finite sample setting, it can be unclear whether one has encountered good or poor performance. Recently, the knockoff procedure for variable selection has been introduced. It is an interesting procedure because it guarantees to the user that, on average, the proportion of wrongfully selected variables, will not exceed a given threshold, which is determined by the user themselves. In a further extension, termed 'model-X knockoffs', this claim holds regardless of the sample size and the underlying relationships in the data. This thesis aims to provide a summary on the existing dimensions of research on knockoffs. Further, through numerical simulations, which investigate several aspects of the procedure, and a real data application, an attempt is made to contribute insights into the behavior and application of knockoffs, while motivating future research in this direction.

I would like to express my deep gratitude to my supervisor, Prof. Dr. Gerda Claeskens, and co-supervisor, Dr. Jing Zhou, for their valuable suggestions and tireless support throughout my research. I would also like to acknowledge and express my appreciation towards every one of my professors during the past two years, who have made my graduate studies the most valuable. Finally, I would like to thank my family for their wholehearted support throughout my journey, and my close friends who have encouraged me, and been there with me in every step of the way.

# Contribution

The broad topic and structure of this thesis, along with some ideas for initial directions of research, were proposed by my thesis supervisor, Prof. Dr. G. Claeskens. The lasso, adaptive lasso and elastic net, to which the knockoff procedure is compared to in the simulations of Chapter 4, were also suggested to me by Prof. Dr. Claeskens.

I was guided to most of the references in Section 1.2 by my thesis co-supervisor, Prof. J. Zhou, from whom I also got the idea to pose the data application of Chapter 5 as a classification problem.

The rest of this thesis, along with the associated code, is my personal contribution.

# Summary

The central topic of this dissertation is the model-X knockoff procedure and specifically, its application to the case of a discrete response. This is a newer variable selection method which has a fairly unique positioning among the existing variable selection techniques. In particular, it can guarantee to the user that on average, the proportion of wrongfully selected variables will not exceed a given threshold. This guarantee is non-asymptotic and holds regardless of the underlying, generally unknown, relationship between the response variable and the measured covariates. As such, the procedure can be applied without the assumption of any model and is considered a 'model-free' selection technique.

An overview of the variable selection problem is given, and the choice of the model-X knockoff procedure as a solution to this problem is motivated. A deeper dive is taken on the knockoff procedure itself and the literature surrounding it, to familiarise the reader with its development and the inherently important research directions for this method.

Next, controlled simulations are carried out to examine several aspects of the knockoff procedure. The effect of covariate correlation on the procedures' performance, in terms of false discovery proportion and power, is studied in the discrete and continuous settings. It is shown that the knockoff procedure loses significant power with the increase of covariate correlation when the response follows a binary distribution, but is not significantly impacted, at the same sample size, when the response follows Poisson or Gaussian distributions. Further, its performance is compared to that of the lasso, adaptive lasso and elastic net variable selection techniques. Knockoffs comparatively exhibit higher power and lower false discovery proportions. Additionally, a comparison is drawn between the knockoff procedure and its more conservative version, knockoff+, showing that the difference between the two diminishes as the threshold for the false discovery rate control is made more liberal. The coefficient amplitude is also demonstrated to have a large impact on the power of these two versions of the procedure. Lastly, considering the effect of the datasets' number of observations and number of covariates jointly, it appears that the knockoff procedures' power is more impacted by the former.

To conclude, an application to the BC-TCGA dataset is illustrated. This high-dimensional dataset contains gene expression information for over 17 000 genes in 590 tissue samples that either belong to the class of normal tissue samples or breast cancer tissue samples. The knockoff procedure is applied for selection after a principal component analysis. A classification of the two types of tissue is then performed, on the basis of this selection, using a logistic regression classifier. Compared to the literature, a lesser performance is achieved on this classification task. This signals the need for more research and guidelines on the application of knockoffs in practical settings.

# Glossary

$n$ the number of observations.

$p$ the number of covariates.

$k$ the number of true covariates.

$q$ the level at which the FDR is controlled.

$A$ the coefficient amplitude.

$R$ the total number of rejected null hypotheses.

$V$ the number of wrongfully rejected null hypotheses.

$\boldsymbol{X}$ a matrix of covariates.

$X_j$ a covariate vector.

$\tilde{\boldsymbol{X}}$ a matrix of knockoff variables.

$\tilde{X}_j$ a knockoff variable vector.

$W_j$ an importance statistic.

$Y$ a vector of responses.

$diag\{s\}$ a nonnegative diagonal matrix.

$\beta_j$ a covariate's coefficient.

$\lambda$ a regularization parameter.

$\rho$ the parameter of an AR(1) correlation structure.

$\tau_q$ the threshold for the knockoff procedure.

$\tau_{+,q}$ the threshold for the knockoff+ procedure.

$\hat{\mathcal{S}}$ a subset of selected variables, or an estimate of the Markov blanket.

$\mathcal{S}$ the Markov blanket.

# Acronyms

**AIC** Akaike information criterion.

**BH** Benjamini-Hochberg procedure.

**BIC** Bayesian information criterion.

**CRT** conditional randomization test.

**FDP** false discovery proportion.

**FDR** false discovery rate.

**FWER** familywise error rate.

**GAN** generative adversarial network.

**LASSO** least absolute shrinkage and selection operator.

**PCA** principal component analysis.

**SCIP** sequential conditional independent pairs.

**SDP** semidefinite program.

**SIS** sure independence screening.

# List of Figures

# List of Tables

# Contents

# Chapter 1

# Introduction

## 1.1 An Overview of the Variable Selection Problem

Driven by the information revolution, handling high dimensional datasets is becoming a necessity in many modern data analysis applications. In such high dimensional settings, it can be expected that many of the covariates are either irrelevant, meaning that they are not related to the response, or redundant, if they are highly correlated with other covariates. Reducing the dimensionality of the problem by screening out such variables may be a necessary step to identify and estimate certain models (i.e. models that require the number of observations $n$ to be larger than the number of covariates $p$) and regardless, a desirable step in order to build stable, parsimonious models, and aid interpretability. Thus, variable selection is an important part of any data analysis problem.

Halinski and Feldt (1970) mention the two main goals in statistical learning to be achieving good predictions in terms of some metric, and discovering relevant explanatory variables. As suggested in Cox and Snell (1974), in the first case, we may be impartial between two models that have different covariates but fit the data equally well and make a choice between the two based on other aspects, like model simplicity. Here, the variable selection step is inherently linked to the model selection step. We expect models built with this goal in mind to have good predictive power when applied to future data generated in a similar setting. On the other hand, in the second case, we seek to estimate the true sparsity pattern and discover true 'explanatory' variables, viewing variable selection as a separate and preceding step to model selection. By building models in this way, we expect them to generalize fairly well, to even vastly different settings, as they are estimating the underlying relationship. In other words, we expect our results to be replicable in greater generality.

More precisely, given a response variable $Y$ and a set of $p$ covariates $\{X_1, X_2, ..., X_p\}$, the variable selection problem is posed as selecting a subset, $\hat{\mathcal{S}} \subset \{1, 2, ..., p\}$, of these variables. This selection process may be driven by different goals. Historically, many popular variable selection methods use model-based metrics to evaluate the 'goodness' of a subset relative to others. Examples of metrics include Mallows' $C_p$ (Mallows, 1973), information criteria such as Akaike information criterion (AIC) (Akaike, 1973) and Bayesian information criterion (BIC) (Schwarz, 1978), estimates of the model's prediction error using for example bootstrap, cross-validation, the mean square error of prediction (Efron, 1992; Shao, 1996; Zhang, 1993; Kohavi,

1995; Allen, 1971) etc. As the number of non trivial subsets equals $2^p - 1$ (excluding the empty set), the evaluation of the entire search space very quickly becomes computationally unwieldy. This has encouraged the development of methods for the reduction of the search space, such as 'best' subset selection (Hocking and Leslie, 1967), sequential selection or elimination (Draper et al., 1966; Efroymson, 1960; Miller, 2002), highest-posterior probability selection (George and McCulloch, 1993), penalization methods such as the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996, 1997), adaptive LASSO (Zou, 2006), the Dantzig selector (Candes and Tao, 2007) etc.

In the application of these methods, an underlying model is assumed or must be chosen by the practitioner, and only then can selection be performed. Much of the research has been focused on applying the class of Gaussian linear models due to its simplicity and the possibility of deriving analytical results in this setting (George, 2000) or on the class of generalized linear models. Even in practice it is common that the same model class is specified for each subset when evaluating and ranking them. This may be done for simplicity or in an attempt to 'eliminate' the influence of the model on the variable selection, by artificially separating variable selection from model selection. However, if the specified model is not a good approximation of the underlying model, poor results can be expected.

Selection may also be based on a significance criterion like the likelihood-ratio test (Vuong, 1989). In this case, if we are comparing several models, we are testing a family of hypotheses and an additional complication that arises, is the need to control for multiple testing. This is important to ensure that our procedure does not result in an unexpected amount of erroneous discoveries. Depending on the problem at hand, different error rates may be the subject of control. A common choice is to ensure that the familywise error rate (FWER), which is the probability of making at least one false discovery, is under control. Benjamini and Hochberg (1995) argued that controlling the FWER is sensible when the overall conclusion of a study could change if one hypothesis is wrongfully rejected, but otherwise it is too stringent and can lead to significant losses in power. Instead, they propose to control the false discovery rate (FDR), which is defined as the expectation of the false discovery proportion (FDP),

$$FDR := \mathbb{E}(FDP) = \mathbb{E}\left(\frac{V}{R}\right), \quad R \neq 0, \tag{1.1}$$

where $V$ denotes the number of wrongfully rejected null hypotheses and $R$ denotes the total number of rejected null hypothesis.

The methods mentioned here are not by any means exhaustive as there exists a vast amount of literature on variable selection, see Miller (2002); Fan and Lv (2010); George (2000); Heinze et al. (2018) for more extensive reviews. See also O'Hara and Sillanpää (2009) for a review on Bayesian methods for variable selection, Huang et al. (2010) for methods applicable to nonparametric models, and Dash and Liu (1997) for the case of a discrete response specifically.

## 1.2 Model-Free Variable Selection

In high dimensional settings, it is difficult and sometimes not even feasible to specify a model. Thus, the classical variable selection methods that also require the specification of a model may not be applicable. Further, if we have no knowledge about the underlying relationship between

the response variable and the covariates, the chances of encountering poor performance of the selection procedure are significant. For these reasons, model-free variable selection methods have been gaining traction.

Model-X knockoffs were introduced by Candès et al. (2018) as a model-free extension to the original knockoffs procedure (Barber and Candès, 2015). Model-X knockoffs are a powerful variable selection tool which can be applied in great generality, without assuming knowledge of the underlying relationship between the response and the covariates, or the noise level, and while maintaining finite sample control of the FDR. Another approach to model-free selection is sure independence screening (SIS), first introduced by Fan and Lv (2008) in the context of linear models. As the name suggests, SIS is a screening procedure which can be applied in high dimensional settings to select a subset containing the true covariates with high probability. Its application in the case of generalized linear models has been discussed in Fan and Song (2010), and the more general case in (Zhu et al., 2011). Other works in the direction of model-free selection include Cui et al. (2015); Mai and Zou (2013); Li et al. (2012); He et al. (2013).

# Chapter 2

# Literature Review

## 2.1 From Fixed-X Knockoffs to Model-X Knockoffs

The 'knockoff filter' was originally introduced by Barber and Candès (2015) as a variable selection procedure that keeps the proportion of false discoveries under control at a level $q$, specified by the user. In simple terms, it relies on the generation of knockoff copies of the covariates, which are independent from the response and thus serve as negative controls to help discern the true covariates from the noise. This result provably holds only for the case of the homogeneous Gaussian linear model

$$Y = \boldsymbol{X}\beta + \varepsilon, \tag{2.1}$$

where $Y \in \mathbb{R}^n$ is the response vector, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is a known and *fixed* design matrix with the number of observations $n$ being at least as large as the number of covariates $p$, and $\varepsilon \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$ with $\sigma^2$ unknown. More specifically, the procedure outputs a subset of selected covariates $\hat{\mathcal{S}} \subset \{1, 2, ..., p\}$ for which the inequality,

$$\mathbb{E}\left[\frac{\#\{j : j \in \hat{\mathcal{S}} \text{ and } \beta_j = 0\}}{\#\{j : j \in \hat{\mathcal{S}}\} + q^{-1}}\right] \leq q, \tag{2.2}$$

holds even in the finite sample setting. The left-hand side quantity in Equation $(2.2)$ is closely related to the FDR given in Section 1.1. Indeed, in the context of variable selection, $V$ can be replaced by the number of null variables wrongfully selected, and $R$ can be replaced by the total number of selected variables. For a Gaussian linear model, a wrongfully selected variable can be defined as a variable $X_j$, for which $\beta_j = 0$, as Candès et al. (2018) have shown that in this case, the response is independent from $X_j$, conditional on the other covariates. In other words, such an $X_j$ is unimportant for the response. The FDR can thus be rewritten as

$$FDR = \mathbb{E}\left[\frac{\#\{j : j \in \hat{\mathcal{S}} \text{ and } \beta_j = 0\}}{\#\{j : j \in \hat{\mathcal{S}}\}}\right], \quad \hat{\mathcal{S}} \neq \emptyset. \tag{2.3}$$

It is apparent now that the two quantities in Equation $(2.2)$ and Equation $(2.3)$ are closely related, and the only difference between them, is the term $q^{-1}$ in the denominator. When a large number of variables is selected, the importance of the additional term diminishes, and the expressions evaluate to a similar quantity. However, considering the relationship between the two quantities,

$$\mathbb{E}\left[\frac{\#\{j : j \in \hat{\mathcal{S}} \text{ and } \beta_j = 0\}}{\#\{j : j \in \hat{\mathcal{S}}\} + q^{-1}}\right] \leq \mathbb{E}\left[\frac{\#\{j : j \in \hat{\mathcal{S}} \text{ and } \beta_j = 0\}}{\#\{j : j \in \hat{\mathcal{S}}\}}\right], \tag{2.4}$$

controlling the FDR at the level $q$ will lead to a slightly more conservative procedure, compared to controlling the left-hand side at the level $q$. Barber and Candès (2015) give a modification to the knockoff procedure, termed knockoff+, which allows the control of the FDR in Equation $(2.3)$ exactly.

A severe limitation to the original knockoff framework is its applicability to only the low dimensional Gaussian setting. In further work, Barber and Candès (2019) propose an extension, which allows the application of the knockoff framework to a high dimensional ($n < p$), Gaussian setting. The procedure is a two step procedure which involves first liberally screening the covariates, using a method of choice, to select a subset $\hat{\mathcal{S}}_0 \subset \{1, 2, ..., p\}$, such that $|\hat{\mathcal{S}}_0| < n$. Only after that is the knockoff procedure applied to select yet a smaller subset $\hat{\mathcal{S}} \subset \hat{\mathcal{S}}_0$. Additionally, instead of controlling the FDR, they suggest controlling the directional FDR ($\text{FDR}_{dir}$) defined as

$$FDR_{dir} = \mathbb{E}[FDP_{dir}] = \frac{|\{j \in \hat{\mathcal{S}} : \widehat{sign}_j \neq sign(\beta_j)\}|}{|\hat{\mathcal{S}}|}, \tag{2.5}$$

where it is given that $|\hat{\mathcal{S}}| \neq 0$ (otherwise $FDR_{dir} \equiv 0$). $\widehat{sign}_j$ denotes the estimated sign of the coefficient of $X_j$ in the true model and $sign(\beta_j)$ denotes its sign in the true model. The directional FDR counts non-null variables selected with the wrong sign as additional false discoveries. The authors argue that controlling $\text{FDR}_{dir}$ is more sensible in such a two step procedure. In fact, Barber and Candès (2019) show that the original knockoff procedure also controls the directional false discovery rate and suggest that it can alternatively be used to check that the signs of the estimated coefficients are reliable. Aside from the FDR and its variants, other type I error rate measures have been proposed as targets to be controlled in the multiple testing setting. One such measure is the $k$-FWER, the probability of making at least $k$ false discoveries. The knockoff procedure has been adapted in Janson and Su (2016) to control the $k$-FWER.

Knockoffs were greatly generalized in Candès et al. (2018) to not only be applicable to a high dimensional setting, but to additionally relax the assumption of an underlying Gaussian linear model for the conditional distribution $F_{Y|X}$. The reimagined framework is named 'model-X knockoffs' to necessarily distinguish it from the original procedure, which is referred to as 'fixed-X knockoffs'. The generality of the theoretical guarantees is achieved by treating the covariates as *stochastic*, and assuming instead full knowledge of their joint distribution $F_X$. This gives hint to the names 'model-X knockoffs' and 'fixed-X knockoffs'. While assuming full knowledge of $F_X$ may be unreasonable in many cases, Candès et al. (2018) show through simulations that the procedure is fairly robust when the joint distribution of the covariates is

Gaussian but with an unknown variance-covariance matrix that must be estimated from the data. In further research, Barber et al. (2020) derive how the misspecification of the joint covariate distribution inflates the false discovery rate, and show that the procedure performs well as long as $F_X$ is reasonably estimated.

Thus, the model-X knockoff procedure is a unique, completely model-free variable selection framework. It performs selection by estimating the Markov blanket $\mathcal{S}$ (Pearl, 1988), which can be thought of as the smallest subset to capture all of the information that the $p$ covariates provide about the response. In other words, conditional on the Markov blanket, the response is probabilistically independent from the other measured covariates:

$$Y \perp\!\!\!\perp \boldsymbol{X}_{\mathcal{S}^C} | \boldsymbol{X}_{\mathcal{S}}. \tag{2.6}$$

Estimating the Markov blanket has been shown in Koller and Sahami (1996) to be a good goal for variable selection as it should rid both irrelevant and redundant variables. Hereinafter, 'knockoffs' refer to 'model-X knockoffs' for simplicity, unless it is made explicit otherwise.

## 2.2 Comparison to Other Methods

In their paper, Benjamini and Hochberg (1995) propose a simple Bonferroni type step-up procedure that acts on p-values and that can control the false discovery rate under independence of these p-values. Benjamini and Yekutieli (2001) later prove that the same procedure controls the FDR under certain forms of dependency as well, and derive a modified version that can be applied to cases of arbitrary dependence. It seems that with valid p-values, performing variable selection while controlling the FDR is possible. In this context, we are particularly interested in obtaining p-values associated to testing the conditional independence of covariates from the response. Obtaining conditional p-values is often difficult or not possible and, as such, it remains a massive obstacle for the application of this procedure to many settings. In their paper, Candès et al. (2018) introduce the conditional randomization test (CRT), similar in spirit to the traditional randomization test (Edgington, 1964), as a method to generate valid conditional p-values but at a heavy computational cost. Alternatively, one may consider working with marginal p-values, associated to marginal tests of independence between the covariates and the response. In this case, the procedure is likely misguided and may lead to the selection of redundant variables and inflated FDR.

For a binary response generated from a binomial linear model with a logit link function, Candès et al. (2018) compare in simulations, the performance of knockoffs to the Benjamini-Hochberg procedure (BH) above, applied to marginal p-values ($\text{BH}_m$) and conditional p-values obtained by the CRT ($\text{BH}_{CRT}$). Under a setting with no correlation between the covariates, they found that $\text{BH}_m$ controls the FDR at a nearly equal level as knockoffs, but achieves lesser power. The ability of the BH procedure to control the FDR in this case can be explained by noting that when no covariate correlation is present, the marginal p-values are equivalent to the conditional p-values. Indeed, once correlation is introduced to the simulation, $\text{BH}_m$ fails to control the FDR. On the other hand, $\text{BH}_{CRT}$ achieves slightly better performance compared to knockoffs, in the sense that it exhibits higher power for a lower FDR. However, the massive computational time required to implement the procedure makes it unfeasible in practice. To

the author's knowledge, there are no other comparable variable selection methods that control the FDR in such generality as model-X knockoffs.

## 2.3 Other Topics

### 2.3.1 Construction of Knockoffs

By definition, knockoffs $\tilde{\boldsymbol{X}}$ are variables that are constructed to satisfy the following two properties.

**Property 1.** *Conditional independence*. (Candès et al., 2018) Knockoffs should be conditionally independent from the response, meaning

$$\tilde{\boldsymbol{X}} \perp\!\!\!\perp Y | \boldsymbol{X}. \tag{2.7}$$

**Property 2.** *Pairwise Exchangebility*. (Candès et al., 2018) For any subset $H \subset \{1, ..., p\}$,

$$(\boldsymbol{X}, \tilde{\boldsymbol{X}})_{swap(H)} \stackrel{d}{=} (\boldsymbol{X}, \tilde{\boldsymbol{X}}), \tag{2.8}$$

where $swap(H)$ denotes the swapping operation applied to covariate-knockoff pairs, and $\stackrel{d}{=}$ denotes equality in distribtuion. In other words, **Property 2** means that swapping a covariate with its knockoff counterpart leaves the joint distribution invariant. For example,

$$(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)_{swap(\{1,2\})} \stackrel{d}{=} (\tilde{X}_1, \tilde{X}_2, X_3, X_1, X_2, \tilde{X}_3).$$

Property 1 is satisfied easily, as it suffices that the knockoffs are constructed only using the available covariate data and without looking at the response. On the other hand, the second property necessarily guides the solution to the problem of knockoff construction. In practice, satisfying this property can be difficult and can become a limiting factor to the generality at which the procedure can be applied. For this reason, the construction of knockoffs is an active area of research.

In the case of Gaussian covariates, Candès et al. (2018) give an exact construction that involves sampling the knockoffs from the conditional Gaussian distribution

$$\tilde{\boldsymbol{X}} | \boldsymbol{X} \stackrel{d}{=} \mathcal{N}(\mu, \boldsymbol{V}), \tag{2.9}$$

with

$$\begin{aligned} \mu &= \boldsymbol{X} - \boldsymbol{X}\boldsymbol{\Sigma}^{-1} diag\{s\}, \\ \boldsymbol{V} &= 2diag\{s\} - diag\{s\}\boldsymbol{\Sigma}^{-1} diag\{s\}, \end{aligned} \tag{2.10}$$

where $\boldsymbol{\Sigma}$ is the *population* covariance matrix and $diag\{s\}$ is a diagonal matrix with nonnegative entries. As the population covariance matrix is assumed to be known, only $diag\{s\}$ remains

to be chosen. For this choice, the equivariant and semidefinite program (SDP) constructions, given in Barber and Candès (2015), can be applied.

Now considering the case of non-Gaussian covariates, Candès et al. (2018) propose the sequential conditional independent pairs (SCIP) algorithm, for the exact construction of knockoffs, which relies on sequentially sampling $\tilde{X}_j$ from the conditional distributions

$$F(X_j|X_{-j}, \tilde{X}_{1:j-1}), \tag{2.11}$$

where $X_{-j}$ denotes $\{X_1, ..., X_{j-1}, X_{j+1}, ..., X_p\}$, and $\tilde{X}_{1:j-1}$ denotes $\{\tilde{X}_1, \tilde{X}_2, ..., \tilde{X}_{j-1}\}$. Bates et al. (2021) argue that this SCIP algorithm is difficult to implement, not only because of the computational complexity, but also because in some cases it may be difficult to determine the conditionals and generate a valid sample from them. Further, they argue that not all valid distributions can be generated with SCIP, which means the procedure may miss out on possible higher power constructions. As an alternative, Bates et al. (2021) propose an accept-reject algorithm to sample exact knockoffs for arbitrarily distributed covariates. This algorithm is unfortunately also quite computationally expensive. In fact, Bates et al. (2021) derive the lower bound for the time-complexity of a completely general knockoff sampler, and show that maintaining generality requires high time-complexity in any case.

Facing difficulties with exact construction, Candès et al. (2018) propose to sample approximate knockoffs, termed 'second-order knockoffs', instead. Second-order knockoffs are approximate in the sense that $(\boldsymbol{X}, \tilde{\boldsymbol{X}})_{swap(H)}$ leaves only the first two moments of the joint distribution invariant. Note that in the case of Gaussian covariates, the approximate construction is equivalent to the exact construction, as the joint distribution is completely determined by its first two moments. It is only when covariates are non-Gaussian that this construction is approximate. Various other constructions for approximate knockoffs have been developed. Romano et al. (2020) give a method that utilizes deep learning to generate approximate knockoffs that approximate higher moments instead of only the first two. Nguyen et al. (2020) introduce a generative adversarial network (GAN) approach for constructing approximate knockoffs. Other works in this direction include Spector and Janson (2022); Jiang et al. (2021); Liu and Zheng (2018); Sudarshan et al. (2020).

### 2.3.2 Aggregation of Knockoffs

The knockoff procedure is a stochastic procedure that is not only characterized by variance that stems from the sampling variability of $\boldsymbol{X}$, but also from sampling the knockoffs themselves. When applied in practice, the procedure can possibly suffer from high fluctuations in terms of FDR and power. Thus, researchers seek to find ways to stabilize knockoffs and achieve more consistent performance. Ren et al. (2021) propose applying the knockoff procedure several times by constructing conditionally independent knockoffs, and afterwards selecting the variables that surpass a certain selection frequency threshold. In this way, control of the k-FWER can be maintained. Alternatively, Nguyen et al. (2020) propose a method for computing intermediate p-values, on top of which the BH procedure can be applied to maintain control of the FDR.

# Chapter 3

# Methodology

## 3.1 Knockoff Procedure

The knockoff procedure is implemented using the library `knockoff` in the software package `R`. The three steps to the knockoff procedure, as carried out by this library, are presented in some detail below.

**Step 1. Constructing Knockoffs.** In order to guarantee FDR control, knockoffs must be constructed in a way that satisfies the properties of *conditional independence* and *pairwise exchangeability* given in Section 2.3.1.

As the numerical experiments in Chapter 4 deal with Gaussian covariates, the exact construction of knockoffs, as shown in Equation $(2.9)$ and Equation $(2.10)$, is used. Additionally, $diag\{s\}$ is determined by the SDP construction (Barber and Candès, 2015) which involves solving the optimization problem

$$\text{maximize} \quad \Sigma_j s_j \quad \text{subject to} \quad 0 \leq s_j \leq 1, \quad 2\Sigma - diag\{s\} \succeq 0, \tag{3.1}$$

with $j \in \{1, 2, ..., p\}$.

**Step 2. Compute importance statistics.** To perform selection, an importance statistic $W_j$ must be computed for each pair $(X_j, \tilde{X}_j)$. The choice of the importance statistic is very flexible. It suffices that the statistic can be expressed as

$$W_j = w_j([\boldsymbol{X}, \tilde{\boldsymbol{X}}], Y), \tag{3.2}$$

for a function $w_j$ that satisfies the *flip-sign property*, defined below. This property is crucial to ensure control of the FDR.

**Property 3.** *Flip-Sign Property.* Candès et al. (2018) If $X_j$ is swapped with its knockoff counterpart $\tilde{X}_j$, the sign of $W_j$ changes, meaning

$$W_j = w_j([\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(H)}, Y) = \begin{cases} w_j([\boldsymbol{X}, \tilde{\boldsymbol{X}}], Y), & j \notin H, \\ -w_j([\boldsymbol{X}, \tilde{\boldsymbol{X}}], Y], & j \in H. \end{cases} \tag{3.3}$$

Constructed to satisfy **Property 3**, the importance statistic $W_j$ has a symmetric distribution under the null hypothesis, which means that a positive or negative sign is equally likely for an unimportant covariate. On the other hand, for an important covariate, a large positive value for $W_j$ is expected. Here, the importance statistic of choice is the coefficient difference

$$W_j = |\hat{\beta}_j(\lambda)| - |\hat{\beta}_{j+p}(\lambda)|, \tag{3.4}$$

where $\hat{\beta}_j(\lambda)$ is the estimated coefficient of the $j$th variable, and similarly $\hat{\beta}_{j+p}(\lambda)$ is the estimated coefficient of its knockoff counterpart, in the fitted model. The fitted model here is a generalized linear model, estimated on the augmented design matrix $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]$, using penalized maximum likelihood, with regularization parameter $\lambda$. More specifically, the coefficients are estimated by

$$(\hat{\beta}_0, \hat{\beta}_j) = \underset{\beta_0, \beta_j}{\arg\min} \left( -\ell(\beta_0, \beta_j) + \lambda \sum_{j=1}^{p} |\beta_j| \right), \tag{3.5}$$

where $\ell(\beta_0, \beta_j)$ denotes the log-likelihood. The penalty is the lasso penalty, and $\lambda$ is determined by 10-fold cross-validation. It should be emphasized that this model is merely used to compute the statistics $W_j$ and not meant to represent the relationship between $\boldsymbol{X}$ and $Y$.

**Step 3. Compare statistics to a data dependent threshold.** Finally, to make a decision on which variables are selected and which are not, the previously computed statistics are compared to a data-dependent threshold. To ensure control of the FDR, the knockoff+ version is applied, which differs from the knockoff procedure, only in this step due to the way that the threshold is defined. For knockoffs+ the threshold $\tau_{+,q}$ is specified as

$$\tau_{+,q} = \min \left\{ t > 0 : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}} \leq q \right\}. \tag{3.6}$$

On the other hand, the threshold of the knockoff procedure $\tau_q$, is

$$\tau_q = \min \left\{ t > 0 : \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}} \leq q \right\}. \tag{3.7}$$

Comparing the statistics $W_j$ to the threshold $\tau_{+,q}$ ($\tau_q$) yields a set of selected covariates $\hat{\mathcal{S}}$. That is, if $W_j \geq \tau_{+,q}$ ($W_j \geq \tau_q$), $X_j$ is selected. To give some insight into why the knockoff procedure manages to maintain FDR control, consider the false discovery proportion for a given $t$,

$$FDP(t) = \frac{\#\{j : X_j \notin \mathcal{S} \text{ and } W_j \geq t\}}{\#\{j : W_j \geq t\}}. \tag{3.8}$$

In practice, the numerator is an unknown quantity since $\mathcal{S}$ is unknown. Using the flip-sign property, it can be shown that given a $W_j$, conditional on its magnitude $|W_j|$, it holds that

$$\#\{j : X_j \notin \mathcal{S} \text{ and } W_j \geq t\} \stackrel{d}{=} \#\{j : X_j \notin \mathcal{S} \text{ and } -W_j \geq t\}, \tag{3.9}$$

or equivalently,

$$\#\{j : X_j \notin \mathcal{S} \text{ and } W_j \geq t\} \stackrel{d}{=} \#\{j : X_j \notin \mathcal{S} \text{ and } W_j \leq -t\}. \tag{3.10}$$

Further, it trivially holds that

$$\#\{j : W_j \leq -t\} \geq \#\{j : X_j \notin \mathcal{S} \text{ and } W_j \leq -t\}. \tag{3.11}$$

Using Equation $(3.10)$ and Equation $(3.11)$ together, it is implied that

$$\#\{j : W_j \leq -t\} \geq \#\{j : X_j \notin \mathcal{S} \text{ and } W_j \geq t\}, \tag{3.12}$$

meaning $\#\{j : W_j \leq -t\}$ is an upwardly biased estimator of $\#\{j : X_j \notin \mathcal{S} \text{ and } W_j \geq t\}$, the numerator of $FDP(t)$. Finally, because of this relationship, $(3.13)$ can be estimated by

$$\widehat{FDP(t)} = \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}}, \tag{3.13}$$

where the numerator is now known. This is precisely the quantity appearing in Equation $(3.7)$. Since more variables will be selected as $t$ (and consequently $\tau_q$ or $\tau_{+,q}$) gets smaller, it can be understood that the knockoff procedure seeks to select as many variables as possible without exceeding the FDR at level $q$ (see Barber and Candès (2015); Candès et al. (2018) for more details and proof).

As the data in Chapter 4 is self-generated, the true non-null variables, or in other words, the Markov blanket $\mathcal{S}$, is known exactly. Considering this, the reported FDP is computed as

$$FDP = \frac{\#\{j : j \in \hat{\mathcal{S}} \text{ and } j \notin \mathcal{S}\}}{\#\{j : j \in \hat{\mathcal{S}}\}}, \tag{3.14}$$

given that $\hat{\mathcal{S}} \neq \emptyset$ otherwise $FDP \equiv 0$. On the other hand, the power of the procedure is computed as the proportion of the true covariates that have been selected,

$$Power = \frac{\#\{j : j \in \hat{\mathcal{S}} \text{ and } j \in \mathcal{S}\}}{\#\{j : j \in \mathcal{S}\}}. \tag{3.15}$$

## 3.2 Other Methods

In this section, the variable selection methods that are compared to the knockoff procedure in the numerical simulations of Chapter 4 are shortly presented.

### 3.2.1 LASSO

The LASSO (Tibshirani, 1996) is a shrinkage technique yielding estimates for the model coefficients as

$$\hat{\beta}_{0,L}, \hat{\beta}_{j,L} = \underset{\beta_0,\beta_j}{\arg\min} \left( \frac{1}{n} \sum_{j=1}^{p} (Y - \beta_0 - X\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right), \tag{3.16}$$

where $\lambda \geq 0$ is the shrinkage parameter to be chosen. A method such as cross-validation based on the prediction error, may be used for this purpose. Because the LASSO uses an $\ell_1$ penalty term, it forces some coefficients to shrink to exactly zero, meaning that it can be used to perform variable selection as well. Under certain conditions the LASSO is consistent for variable selection (Zhao and Yu, 2006). Consistency means here that, as the sample size $n$ approaches infinity, the probability of selecting the true model approaches one.

### 3.2.2 Adaptive LASSO

The adaptive LASSO (Zou, 2006) is essentially a weighted LASSO, and similar to the LASSO it tends to shrink some coefficients to zero. Due to this fact, it can also be used to select a subset of variables. The adaptive LASSO estimates are given by

$$\hat{\beta}_{0,AL}, \hat{\beta}_{j,AL} = \underset{\beta_0,\beta_j}{\arg\min} \left( \frac{1}{n} \sum_{j=1}^{p} (y - \beta_0 - X\beta_j)^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right). \tag{3.17}$$

The weights can be chosen as $w_j = \frac{1}{|\tilde{\beta}_j|^\gamma}$, where $\tilde{\beta}_j$ is an unbiased estimator of the true coefficient of $X_j$. In this way, the procedure is consistent for selection as the weights for the unimportant variables will tend to 0, while for the important covariates, they will tend away from 0.

### 3.2.3 Elastic Net

The elastic net is a shrinkage and variable selection method proposed by Zou and Hastie (2005). It was developed with the intention of improving upon a few of the limitations of the LASSO. Namely, the LASSO cannot select more than $n$ variables before it becomes saturated and further, in highly correlated settings, it tends to select a single variable among a cluster of highly correlated covariates, instead of selecting a group. The elastic net improves upon these points while also retaining the positive aspects of the LASSO. Namely, it performs continuous shrinkage and results in a similar sparsity representation. The elastic net estimates $(\hat{\beta}_{0,EN}, \hat{\beta}_{j,EN})$, are obtained by

$$\hat{\beta}_{0,EN}, \hat{\beta}_{j,EN} = \underset{\beta_0,\beta_j}{\arg\min} \left( \frac{1}{2n} \sum_{j=1}^{p} (y - \beta_0 - X\beta_j)^2 + P(\alpha, \lambda) \right), \tag{3.18}$$

where the penalty term

$$P(\alpha, \lambda) = \lambda \sum_{j=1}^{p} \left( \frac{(1-\alpha)}{2} \beta_j^2 + \alpha |\beta_j| \right), \tag{3.19}$$

is a convex combination of the lasso penalty and ridge penalty (Hoerl and Kennard, 1970). The two regularization parameters, $(\lambda, \alpha)$, need to be specified by the user. Once again, a technique like cross-validation may be used for this purpose.

# Chapter 4

# Numerical Simulations

## 4.1 Data Generation

For all of the experimental simulation studies in this chapter, first data was sampled for the covariates and afterwards, the response was sampled to satisfy a particular model of these covariates. In this way, the true underlying model that generated the response variable is exactly known. The parameters of the simulations are: the number of observations $n$, the number of covariates $p$, the number of non-null covariates $k$, the correlation coefficient $\rho$, the coefficient amplitude $A$, and $q$, the level at which the FDR is controlled. The data generation process is explained in generality below as the abovementioned parameters are varied throughout the different experiments.

To begin, the covariates were sampled from a joint Gaussian distribution with mean $\mu = 0_p$ and variance-covarince matrix $\mathbf{\Sigma}_{\boldsymbol{p}\times\boldsymbol{p}}$, following an AR(1) correlation structure with autoregressive coefficient $\rho$. More precisely, the variance-covariance matrix has the following structure,

$$\mathbf{\Sigma}_{\boldsymbol{p}\times\boldsymbol{p}} = \begin{bmatrix} 1 & \rho & \rho^2 & ... & \rho^p \\ \rho & 1 & \rho^3 & ... & \rho^{p-1} \\ \rho^2 & \rho^3 & 1 & ... & \rho^{p-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^p & \rho^{p-1} & \rho^{p-2} & ... & 1 \end{bmatrix}. \tag{4.1}$$

Among the $p$ covariates, by uniform sampling (without replacement), $k$ were designated as the true covariates, composing the Markov blanket $\mathcal{S}$. The coefficient magnitudes of these non-null variables were set to be equal as

$$|\beta_j| = \frac{A}{\sqrt{n}}, \quad j \in \mathcal{S}, \tag{4.2}$$

and their sign was assigned randomly with equal probability.

Finally the response is generated by sampling from either a Bernoulli, Poisson or Gaussian distribution. The binary response is obtained by sampling $n$ times from a Bernoulli

distribution such that $Y_i \sim Bernoulli(\pi_i)$, with $\pi_i$ the probability of success determined by $\pi_i = \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}$. Thus, the underlying conditional model for $Y$ is a logistic regression model

$$\text{logit}(\mathbb{E}[Y|\boldsymbol{X}]) = \text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{X}\beta. \tag{4.3}$$

The Poisson response is sampled from the Poisson distribution such that $Y_i \sim Pois(\lambda_i)$, with the rate parameter determined as $\lambda_i = \exp(X_i\beta)$. The underlying conditional model in this case is

$$\log(\mathbb{E}[Y|\boldsymbol{X}]) = \log(\lambda) = \boldsymbol{X}\beta. \tag{4.4}$$

Lastly, the Gaussian response was obtained by sampling from a Gaussian distribution with mean $\mu_i = X_i^T\beta$ and variance $\sigma^2 \sim \mathcal{N}(0,1)$. The true conditional model for the Gaussian response is given by

$$\mathbb{E}[Y|\boldsymbol{X}] = \boldsymbol{X}\beta. \tag{4.5}$$

## 4.2 The Effect of Covariate Correlation

For this simulation study, the covariate data generation parameters are $n = 1000$, $p = 500$, $k = 50$, $A = 10$ and $\rho$ varied from 0 to 0.8. The effect of correlation on the false discovery rate and power is studied by averaging the simulation results obtained on 50 datasets. The case of a discrete response is examined by looking at both Bernoulli and Poisson distributed responses. A Gaussian response is also considered to draw a comparison between the discrete and continuous cases. Additionally, the knockoff+ procedure is compared to the LASSO, adaptive LASSO, and elastic net variable selection methods, as these are popular choices among practitioners.

The knockoff+ procedure was applied, as specified in Chapter 3 with $q = 10\%$. For the other procedures, referring to Equations $(3.16)$, $(3.17)$ and $(3.19)$, $\lambda$ is chosen by 10-fold cross-validation, whereas $\alpha$ for the elastic net is fixed to be 0.5, as a balance between the lasso and ridge penalties. Lastly, the weights for the adaptive LASSO were computed as

$$w_j = \frac{1}{|\tilde{\beta}_j(\lambda)|},$$

where $\tilde{\beta}_j$ is the coefficient of $X_j$ in the appropriate generalized linear model fitted with penalized maximum likelihood $(3.5)$.

### 4.2.1 Binary Response

To begin, the binary response case is considered. The FDR is estimated by the average FDP over 50 datasets, generated with the same parameters but different seeds, and is shown in Figure 4.1, for all procedures. The knockoff+ procedure is the only one to maintain an average

FDP under the level $q = 10\%$, regardless of the amount of correlation in the data. This result is expected from the theory as it is the only one among the compared procedures to offer such a guarantee. The other side of the story can be seen in Figure 4.2, which shows the average number of selected variables (left) and average power (right) in each case. It becomes clear that the LASSO and elastic net procedures select, on average, many more variables, making simultaneously more true and false discoveries. Hence, these two procedures have both higher power and a higher FDP. However, as the correlation increases, the number of selected variables by the LASSO and elastic net decreases. The opposite occurs in the case of the adaptive LASSO, which tends to select more variables in higher correlations, resulting in an increase of its average FDP but not so much of its power.

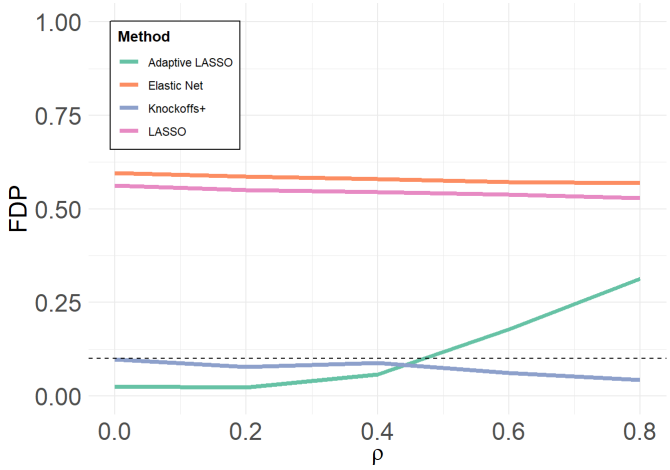


Figure 4.1: Average FDP in the binary response setting. The dashed line marks the level $q$ of FDR control ($q = 10\%$).

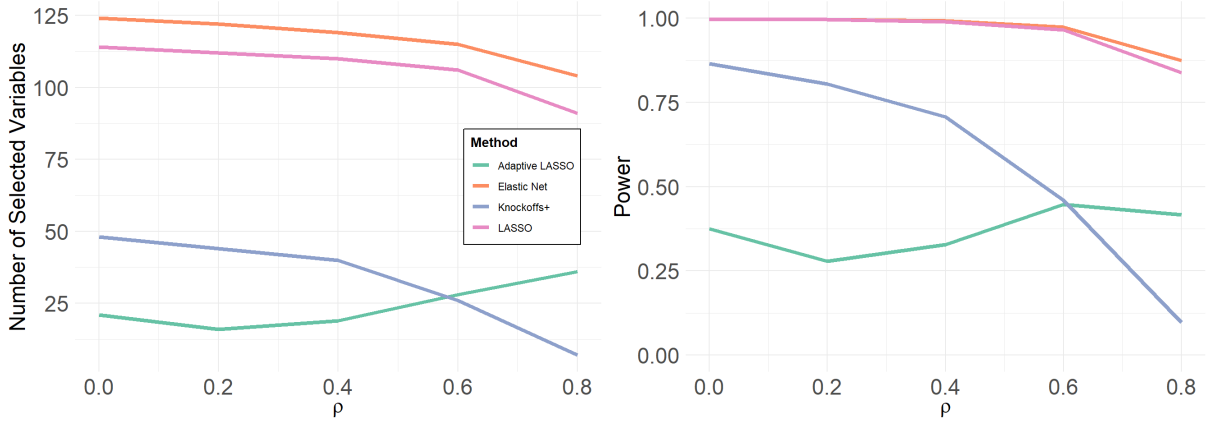


Figure 4.2: Average number of selected variables (left) and average power (right), in the binary response setting.

Shifting the focus to the knockoff+ procedure, it is seems that as the correlation increases, it tends to select fewer variables to maintain control of the FDR, and loses a significant amount of power. Further, shown in Figure 4.3 (left), the distribution of the procedure's FDP remains fairly stable as the correlation is varied, except when $\rho = 0.8$. In this case, the distribution becomes peaked at 0 because, for some datasets, no selection is made, resulting in a relatively

high occurrence of zero FDP and power selections. On the other hand, the distribution of the power, displayed in Figure 4.3 (right), shifts steadily towards zero with the increasing correlation, and similarly becomes peaked at 0 when $\rho = 0.8$ for the abovementioned reason.

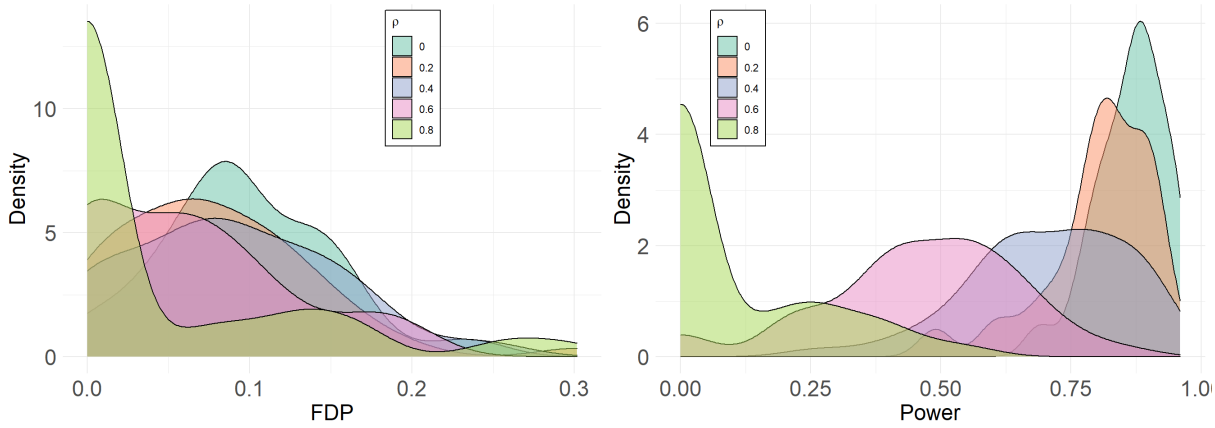


Figure 4.3: Distribution of the knockoff+ procedure's FDP (left) and power (right), in the binary response setting.

### 4.2.2 Poisson Response

In a similar fashion, the case of a Poisson response is examined. The average FDP and power, under the different levels of correlation, are shown in Figure 4.4. It seems that all procedures maintain an average FDP around 10% and, aside from knockoffs, exhibit fairly low power. However, it is important to note that in this experiment, oftentimes, the adaptive LASSO, but especially the LASSO and elastic net, select exactly zero variables. Meanwhile, the knockoff+ procedure makes a selection in every case. Table 4.1 shows a summary of the times that these procedures fail to select a subset of variables.

| | $\rho$ | | | | |
|---|---|---|---|---|---|
| **Selection Procedure** | **0** | **0.2** | **0.4** | **0.6** | **0.8** |
| **Adaptive Lasso** | 58% (29) | 60% (30) | 53% (27) | 46% (23) | 58% (29) |
| **Elastic Net** | 82% (41) | 86% (43) | 86% (43) | 80% (40) | 78% (39) |
| **Lasso** | 82% (41) | 86% (43) | 86% (43) | 82% (41) | 78% (39) |

Table 4.1: Percentage (count) of times where each selection procedure does not select any variables.

Taking this fact into account, the average FDP and power for the adaptive LASSO, LASSO, and elastic net, considering only cases where a selection was made, are shown in Figure 4.5. The difference is especially large for the LASSO and elastic net. It seems that, when applying these methods, it can either occur that no selection is made, resulting in zero power and FDP, or that many variables are selected (Figure 4.6), resulting in simultaneously high power and FDP. The average in this case, does not portray a very accurate picture of what one could expect in a particular realization, when applying the two procedures.

Notably, the LASSO and elastic net perform worse on average in the Poisson response case than the binary response case. Meanwhile, the adaptive LASSO achieves similar power, but at

a lower FDP. The knockoff+ procedure demonstrates an impressive improvement compared to the binary case, achieving (near) perfect power, even in the cases of high correlation, while of course maintaining FDR control.

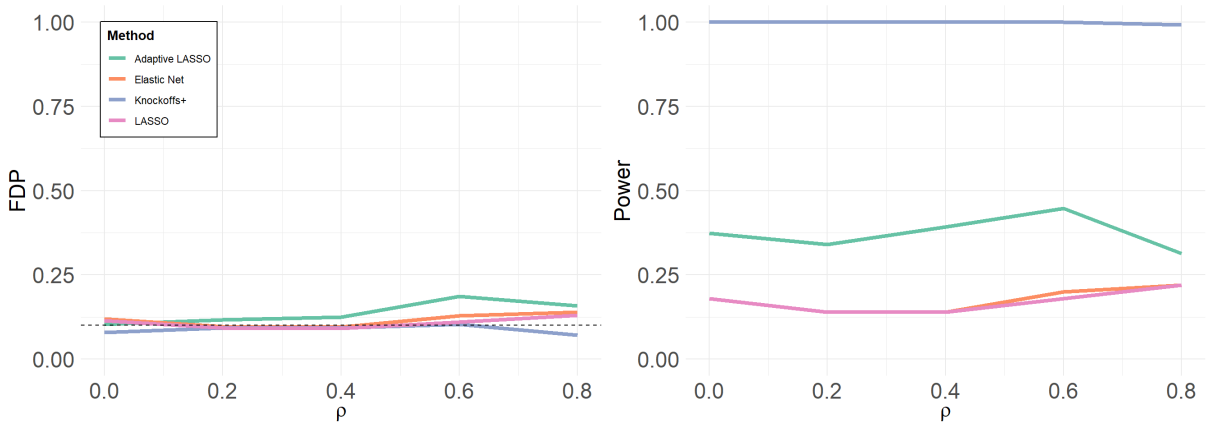


Figure 4.4: Average FDP (left) and power (right) in the Poisson response setting. The dashed line marks the level $q$ of FDR control ($q = 10\%$).

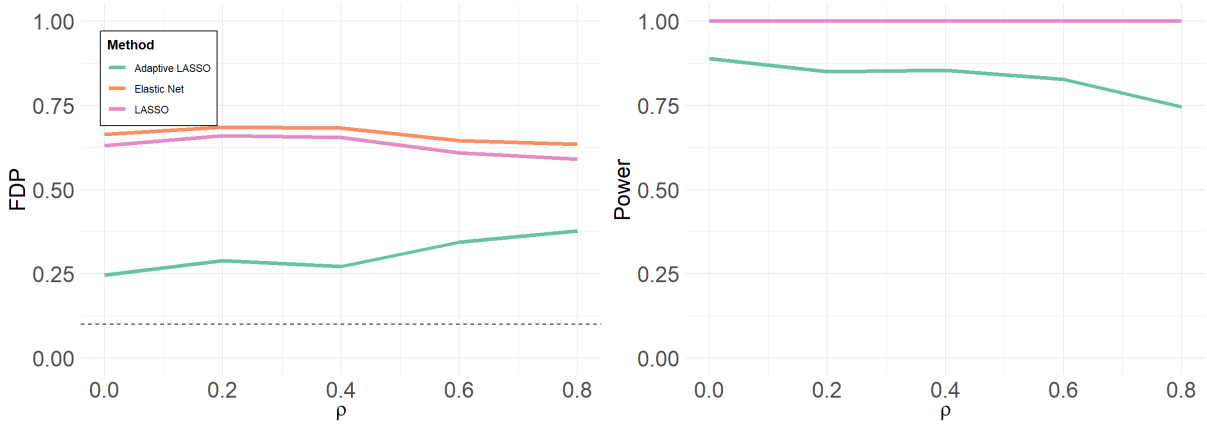


Figure 4.5: Average FDP (left) and power (right) in the Poisson response setting, considering only cases where a selection was made. The dashed line marks the level $q$ of FDR control ($q = 10\%$). Note that the elastic net and LASSO curves are overlapping in the right plot.

### 4.2.3 Gaussian Response

Finally, a continuous, Gaussian response is under consideration. When $\rho = 0$, the LASSO and elastic net fail to make a selection around 80% of the time once again. This explains their exceptionally low FDP and power, shown in Figure 4.7 and Figure 4.8 respectively. The adaptive LASSO also tends to select too few variables, resulting in comparatively poor power. On the other hand, the knockoff+ procedure demonstrates very high power, accompanied by a low average FDP. A small drop off in the number of selected variables and power is noticed when $\rho$ increases to 0.8.

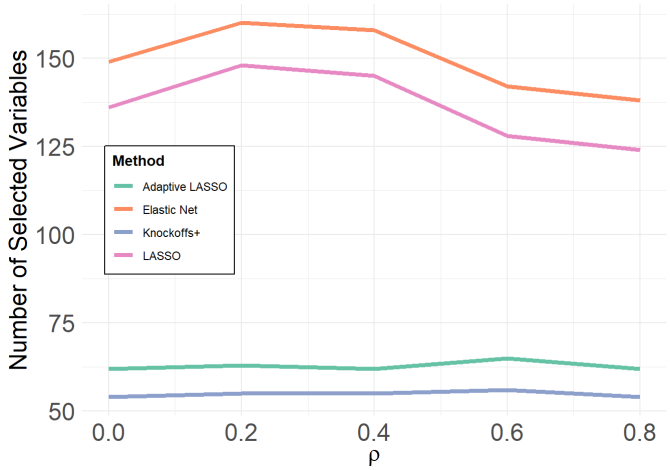


Figure 4.6: Average number of selected variables, among cases where a selection was made, in the Poisson response setting.

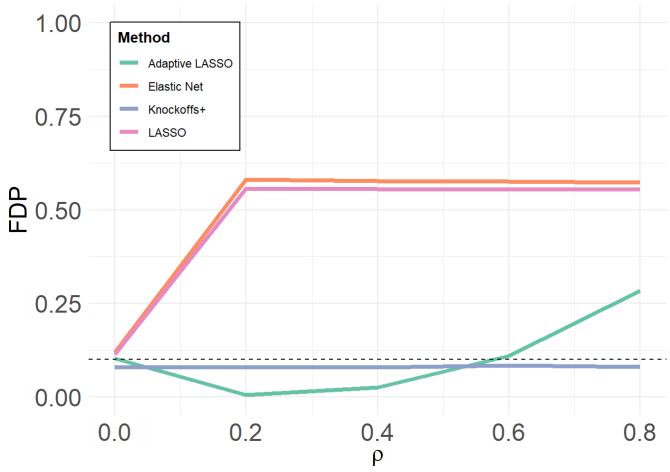


Figure 4.7: Average FDP in the Gaussian response setting. The dashed line marks the level $q$ of FDR control ($q = 10\%$).

## 4.3 Knockoffs vs. Knockoffs+

In this simulation study, the knockoff and knockoff+ procedures are compared at several levels of FDR control $q$, to investigate the difference between the two. It is advantageous that the knockoffs and importance statistics only need to be computed once for this purpose, as the thresholds $\tau_q$, $\tau_{+,q}$ can be calculated easily to re-perform the selection, on the basis of the same importance statistics, at each level of $q$. The following results are averaged over 20 datasets, generated with $n = 1000$, $p = 200$, $k = 50$ and $\rho = 0$. Only a binary response is under consideration. To add another dimension to this experiment, the coefficient amplitudes $A$ were varied jointly from 1 to 50. The corresponding coefficient magnitudes on the logit scale, as determined by Equation $(4.2)$, are shown below in Table 4.2.

Firstly, the effect of the coefficients' amplitude is examined. Referring to Figure 4.10, which shows the average power, it can be seen that at $A = 1$, both procedures demonstrate exactly zero power. However, already at $A = 5$, some power is gained for the knockoff procedure at

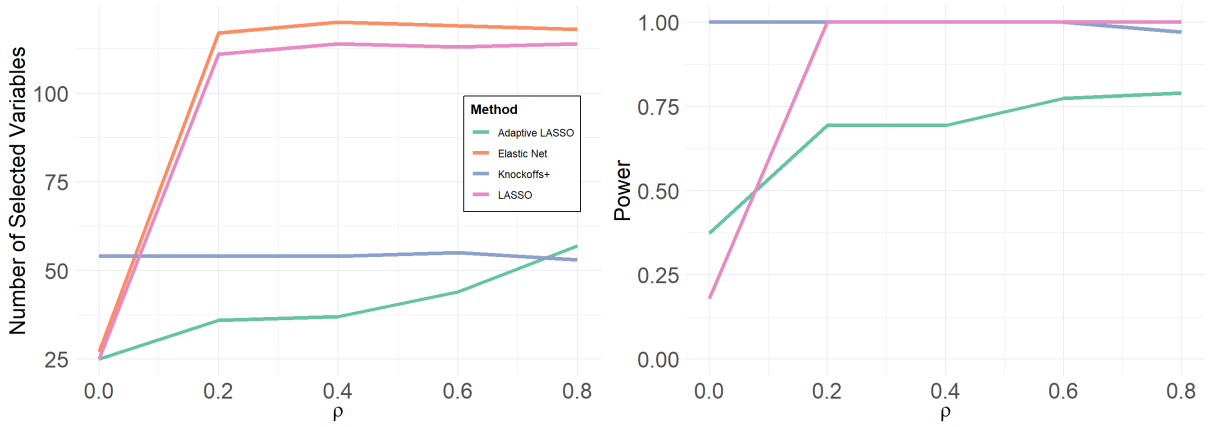


Figure 4.8: The average number of selected variables (left) and average power (right) in the Gaussian response setting. Note that the elastic net and lasso curves are overlapping in the right plot.

| **Amplitude** | **Coefficient Magnitude** |
|---:|---:|
| 1 | 0.01 |
| 5 | 0.05 |
| 10 | 0.1 |
| 20 | 0.2 |
| 30 | 0.3 |
| 40 | 0.4 |
| 50 | 0.5 |

Table 4.2: True model coefficient magnitudes on the logit scale at each amplitude level.

all levels of $q$ and for the knockoff+ procedure at $q \geq 10\%$. The power increases rapidly, until $A = 20$, where a near perfect power is achieved and maintained onward. A further increase in amplitude has the effect of decreasing the average FDP, as fewer variables, and in particular null variables, are selected. To illustrate this, the average number of selected variables is shown in Figure 4.11. Understandably, it is much easier to distinguish the true variables from the noise when their effects are large in magnitude.

Now shifting the focus to the comparison of the two procedures, a difference in behavior is observed, especially in terms of FDP, in the settings of lower $q$ and lower $A$. This can be seen in Figure 4.9. In particular, when $A = 1$, the knockoff+ procedure does not result in any selection until $q = 50\%$, while the knockoff procedure sometimes makes a selection, even at $q = 5\%$. Aside from these 'extreme' cases, in terms of FDP, both procedures remain fairly close to each other and to the threshold $q$, indicating that the knockoff+ procedure is not overly conservative. Although using the threshold $\tau_q$ rather than $\tau_{+,q}$ generally results in more selected variables, in this case, it is usually translated into a higher FDP rather than a higher power for the procedure. As $q$ is made more liberal, the two procedures seem to approach a convergence. Since more variables are generally selected in that case, the impact of the extra term in the numerator of $\tau_{+,q}$, as given in Equation $(3.6)$, compared to $\tau_q$ given in Equation $(3.7)$, diminishes.

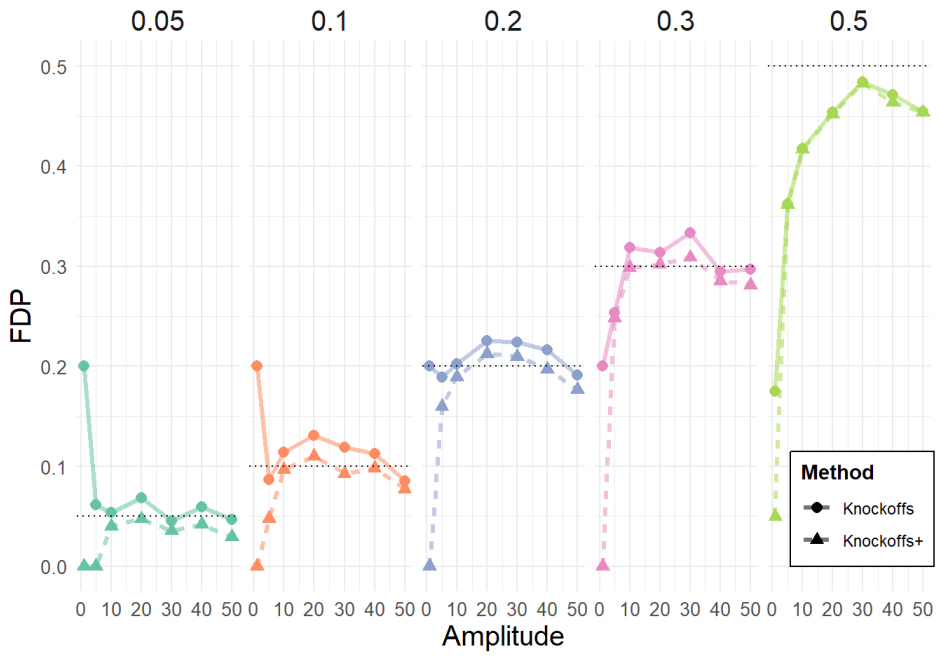


Figure 4.9: Average FDP for the knockoff and knockoff+ procedures. The top axis, and black dashed lines, indicates the level $q$ at which the FDR is controlled.

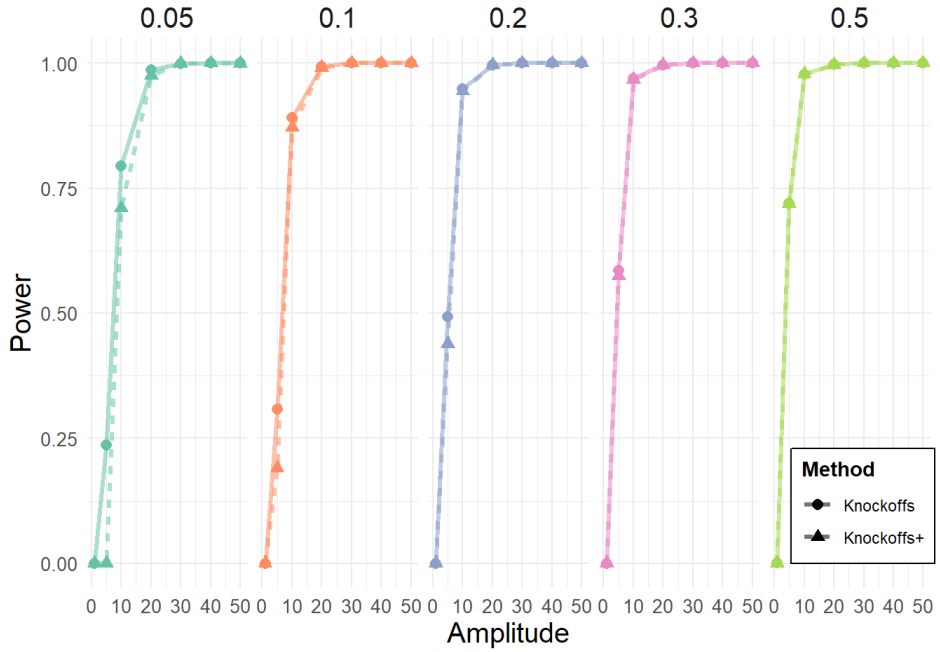


Figure 4.10: Average power for the knockoff and knockoff+ procedures. The top axis indicates the level $q$ at which the FDR is controlled.

## 4.4 The Joint Effect of $n$ and $p$ on Power

Here, covariate data was generated with $k = 20$, $A = 10$ and $\rho = 0$ while a grid of values ranging from 50 to 1000, in increments of 50, were used for $n$ and $p$. Once again, a binary

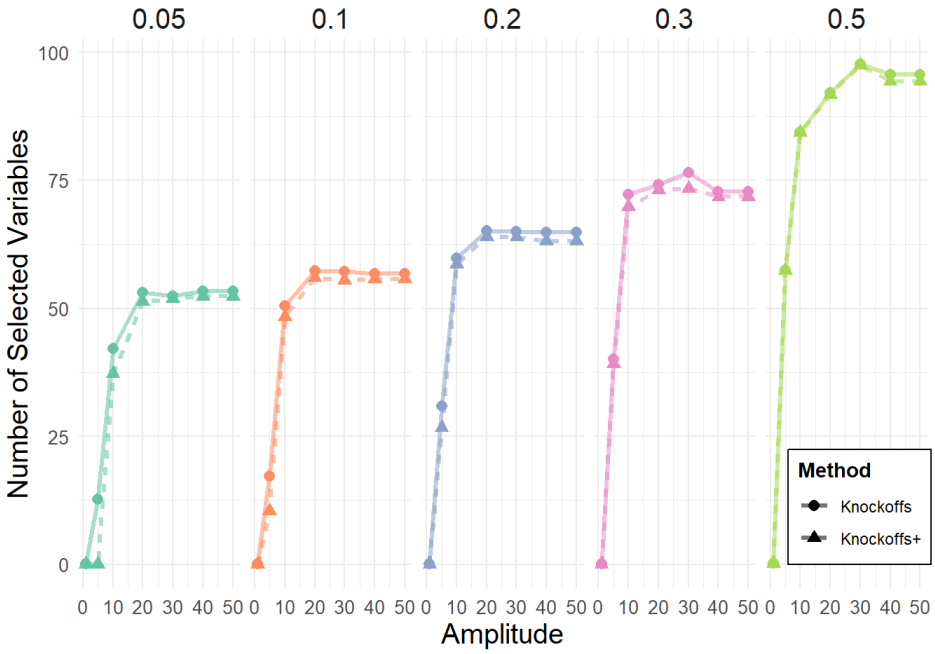


Figure 4.11: Average number of selected variables for the knockoff and knockoff+ procedures. The top axis indicates the level $q$ at which the FDR is controlled.

response was considered. The knockoff+ procedure was applied with $q = 10\%$. The purpose of this simulation is to examine whether there is a joint dependency between the procedures' power on the one hand, and the number of observations $n$ and number of covariates $p$, on the other. The following results were obtained by averaging over 20 datsets.

Figure 4.12 displays the average power for each combination of $n$ and $p$. It appears that generally, for a particular sample size $n$, the power is not greatly impacted by the size of $p$, relatively speaking. On the other hand, for a given $p$, the power is more significantly impacted by the sample size $n$. Comparing the high dimensional settings located under the diagonal, to the low dimensional settings located over the diagonal, a considerable difference in the power is observed, with the high dimensional settings suffering from considerably lower power. This is however, not a surprising result.

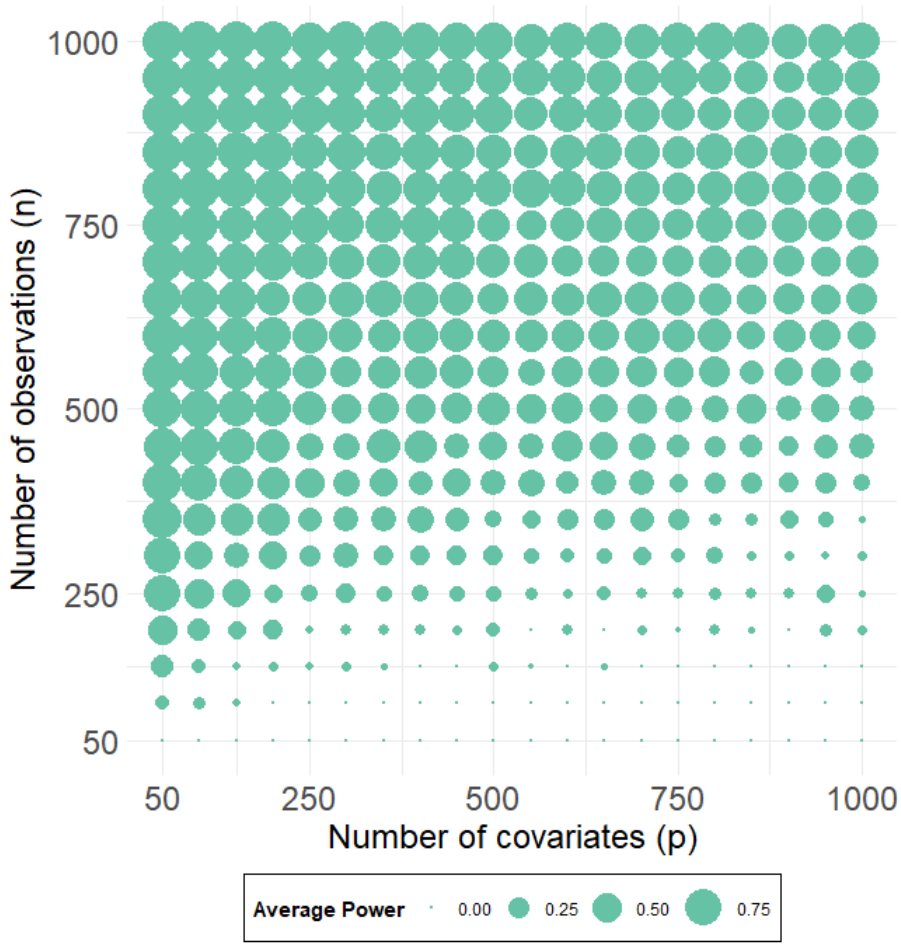


Figure 4.12: Average power for combinations of $n$ and $p$ between 50 and 1000.

# Chapter 5

# Data Application

## 5.1 Data Description

In this chapter, an application of the knockoff procedure to a real dataset is demonstrated. For this purpose, the BC-TCGA dataset (Koboldt et al., 2012) (retrieved from `https://data.mendeley.com/datasets/v3cc2p38hb/1`), was used. This dataset contains the expressions of $p = 17\,814$ genes for two different kinds of tissue samples. It includes $n = 590$ tissue samples, among which 61 are healthy tissue and 528 are breast cancer tissue. As such the data is high dimensional with $p \gg n$. The goal is to use this dataset to perform a classification of the two types of tissues, with the knockoff procedure first applied to perform variable selection.

## 5.2 Data Preprocessing

The dataset at hand presents two main challenges. Firstly, the number of genes is very large, making the application of the knockoff procedure difficult computationally. To add on, gene expression datasets contain very high, sometimes near perfect correlations, and the BC-TCGA dataset is no exception to this. It is known that the presence of correlation complicates the variable selection process. This fact was also apparent in Section 4.2.1, where particularly in the binary response setting, the knockoff loses a significant amount of power in high correlations. Thus, to make the application of the knockoff procedure feasible, some preprocessing was performed.

To begin, the data was examined for missing values. No missing values are present, but upon further inspection, 1 695 'null' values, distributed across 534 genes, were found. A decision was made to remove these genes from the dataset for simplicity. The data was then standardized, to have mean 0 and variance 1, in order to account for possible differing magnitudes in the gene expressions. A training and test set were sampled in a stratified fashion to contain respectively 75% and 25% of the observations. The stratification ensures that the normal tissue and cancer tissue samples are represented in equal proportion in both sets. This is important for evaluating the model, considering that the dataset is highly unbalanced. Indeed, almost 90% of the observations are of the breast cancer type.

The next step is to perform a dimensionality reduction, for which principal component analysis (PCA) is used. The transformation is applied to the training set, resulting in a

442×441 dataset. Note that these dimensions are obtained because if $p$ exceeds $n$, as in this case, the number of nontrivial (nonzero) principal components is at most $n-1$. The same transformation is then applied to the test set. The principal components yielded by PCA, are essentially a new set of covariates which are orthogonal, or in other words, uncorrelated. It seems that PCA solves the two issues at once: it rids the correlation and reduces the dimensionality of the data. All of this comes at a cost however, as it is much more difficult to interpret the solution in terms of the original variables. In this case, since the purpose is to perform classification, this fact does not pose a major issue for concern.

## 5.3 Knockoff Procedure

To apply the knockoff procedure, second-order knockoffs are constructed by first estimating a multivariate Gaussian distribution to fit the data, and afterwards applying the Gaussian construction shown previously in Section 2.3.1. The choice of $diag\{s\}$ is made by the approximate SDP construction which speeds up computation (see Candès et al. (2018) for more details on these methods). Further, the FDR control level is specified as $q = 10\%$. Once the knockoff procedure is applied to the set of principal components, only 11 are selected. Although the less conservative threshold $\tau_q$ (Equation 3.7) was used, it seems that few variables are selected. Significantly increasing $q$, up to 90%, does not yield a different selection. Thus, after selection, the training dataset at hand is reduced to dimensions 442×11.

## 5.4 Classification

Finally, the classification is carried out using a standard logistic regression classifier, implemented in the popular `Python` library `scikit-learn`. The parameters are first tuned on the training set, afterwards, the model is evaluated on the test set. The confusion matrices for the training and test sets are shown in Table 5.1 and Table 5.2 respectively.

| | | **Predicted** | |
|---|---|---|---|
| | | **0** | **1** |
| **Actual** | **0** | 8 | 35 |
| | **1** | 2 | 397 |

Table 5.1: Training set confusion matrix for logistic regression classifier (0 represents normal tissue and 1 represents cancer tissue).

| | | **Predicted** | |
|---|---|---|---|
| | | **0** | **1** |
| **Actual** | **0** | 2 | 16 |
| | **1** | 0 | 130 |

Table 5.2: Test set confusion matrix for logistic regression classifier (0 represents normal tissue and 1 represents cancer tissue).

The logistic regression classifier has an accuracy of 91.6% on the training set and 89.2% on the test set. While the performance appears good at first, looking at Table 5.1 and Table 5.2, it quickly becomes obvious that the classifier is almost always classifying observations as 'tumor' (1) and is achieving a high accuracy due to the predominance of such tissue samples in the data. To further illustrate this point, while the true positive rate is 99.4% in the training set and 100% in the test set, the true negative rate is only 18.6% and 11.1%. Alternatively, if classification is performed using the entire set of principal components (i.e. without additionally applying the knockoff procedure on top of PCA), perfect prediction accuracy is achieved on both the training and test set.

Looking towards the literature, a classification of the BC-TCGA dataset was previously performed by Xie et al. (2016), using a support vector classifier. Various methods were initially applied to reduce the dimensionality of the data. The best performance on the test set (98.97%) was achieved by retaining genes that had a statistically significant mean difference in the two groups, as determined by $t$-tests. Afterwards, random projection was used to project the data onto a lower dimensional subspace.

Thus, using PCA + knockoffs, resulted in a worse performance. It is difficult to say what the issue in the pipeline is. However, the results could possibly be improved by tackling the sampling imbalance, combining knockoffs with a different classification model or dimensionality reduction technique, or modifying the knockoff procedure in itself (e.g using a different construction for the knockoffs or a different importance statistic).

# Chapter 6

# Conclusion

Several dimensions of performing variable selection using model-X knockoffs were investigated through numerical simulations and a real data application. The knockoff+ selection was compared to that of the LASSO, adaptive LASSO and elastic net, under increasing correlation and in both discrete and continuous settings. The results confirm the theoretical guarantees of FDR control for the knockoff procedure, which are not impacted by the level of correlation or the response type. To add on, knockoffs often achieve comparatively higher power at a similar or lower FDP. Generally the knockoff procedure exhibits poorer performance in the binary response setting, compared to the Poisson and Gaussian settings. Further, the effect of correlation can be severe in this case, where a power drop from around 85% to less than 15% is observed. Meanwhile in the Poisson and Gaussian settings, the power drop is nearly negligible. The binary response encodes very coarse information, which makes it much more difficult to distinguish the true effects from the noise. Further combining this with a high correlation, appears to make performing a selection while maintaining FDR control very difficult. While a sample size of $n = 1000$ suffices to achieve near perfect power in the Poisson and Gaussian response case, a larger sample size is likely needed to achieve higher power when the response is binary to account for the difficulty of this setting.

Comparing the knockoff+ selection to the less conservative knockoff selection, it is observed that while the latter typically results in a larger selected subset, the power difference of the two selections is usually small and not justifying the larger FDP. As the threshold $q$ increases, both methods select more variables, making knockoffs and knockoffs+ nearly indistinguishable, except in the cases of very low amplitude $A = 1$, where very few selections are still made. This result is expected from the equations defining the selection thresholds. The extra '+1' term in the threshold for the knockoff+ procedure $(3.6)$ becomes increasingly unimportant as the numerator increases with the number of selected variables. A comparison of the average power obtained for a grid of values for $n$ and $p$, shows that the power of the procedure depends more so on the sample size than the number of covariates but also that increasingly larger samples are needed to maintain a similar power when $p$ increases. Furthermore, performing selection in the high dimensional $(n < p)$ setting while maintaining FDP control results in considerably less power as compared to the low dimensional $(n > p)$ setting.

Finally, an application of the knockoff procedure to the BC-TCGA dataset, containing the expressions of 17 814 genes in 61 normal tissue samples and 528 breast cancer tissue samples, was demonstrated. Selection using knockoffs was conducted on the set of 441 principal com-

ponents obtained after initially performing a PCA. The application of the procedure resulted in 11 principal components being selected. Then, using a logistic regression classifier, tissue samples were classified into two types, normal tissue and cancer tissue. A classification accuracy 91.6% and 89.2%, which can be largely attributed to true positives, was achieved on the training and test sets respectively. This is a lesser accuracy than achieved in the comparable literature. Two notable issues were faced in the application of knockoffs to this dataset, which created the need for a significant amount of preprocessing. Firstly, the presence of large correlations in the data poses an issue for solving the optimization problem necessary to construct knockoffs, and in general, complicates the variable selection problem. Secondly, with the number of covariates being larger than 17 000, the computation of knockoffs can take several hours. These two aspects become prohibitive to the 'out-of-the-box' application of the knockoff procedure. The main conclusion that can be drawn from this demonstration is that it has not yet become well established how these issues should be tackled to make the knockoff procedure an accessible variable selection technique to the average practitioner, so advantage can be taken of the unique theoretical guarantees of this method. As the knockoff procedure appears very promising in simulation studies, further research in this direction is well motivated.

# Bibliography


Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, pages 267–281. Akadémiai Kiadó Location Budapest, Hungary.

Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics*, 13(3):469–475.

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

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

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

Bates, S., Candès, E., Janson, L., and Wang, W. (2021). Metropolized knockoff sampling. *Journal of the American Statistical Association*, 116(535):1413–1427.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57(1):289–300.

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

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351.

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

Cox, D. R. and Snell, E. J. (1974). The choice of variables in observational studies. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 23(1):51–59.

Cui, H., Li, R., and Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association*, 110(510):630–641.

Dash, M. and Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, 1(1-4):131–156.

Draper, N., Draper, N., Smith, H., and Smith, N. (1966). *Applied Regression Analysis*. Number v. 1 in Applied Regression Analysis. Wiley.

Edgington, E. S. (1964). Randomization tests. *The Journal of Psychology*, 57(2):445–449.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in Statistics*, pages 569–593. Springer.

Efroymson, M. A. (1960). Multiple regression analysis. *Mathematical methods for digital computers*, pages 191–203.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, 38(6):3567–3604.

George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association*, 95(452):1304–1308.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Halinski, R. S. and Feldt, L. S. (1970). The selection of variables in multiple regression analysis. *Journal of Educational Measurement*, 7(3):151–157.

He, X., Wang, L., and Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics*, 41(1):342–369.

Heinze, G., Wallisch, C., and Dunkler, D. (2018). Variable selection–a review and recommendations for the practicing statistician. *Biometrical Journal*, 60(3):431–449.

Hocking, R. R. and Leslie, R. (1967). Selection of the best subset in regression analysis. *Technometrics*, 9(4):531–540.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. *The Annals of Statistics*, 38(4):2282.

Janson, L. and Su, W. (2016). Familywise error rate control via knockoffs. *Electronic Journal of Statistics*, 10(1):960–975.

Jiang, T., Li, Y., and Motsinger-Reif, A. A. (2021). Knockoff boosted tree for model-free variable selection. *Bioinformatics*, 37(7):976–983.

Koboldt, D., Fulton, R., McLellan, M., Schmidt, H., Kalicki-Veizer, J., McMichael, J., Fulton, L., Dooling, D., Ding, L., Mardis, E., et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International joint conference on artificial intelligence*, volume 14, pages 1137–1145. Montreal, Canada.

Koller, D. and Sahami, M. (1996). Toward optimal feature selection. Technical report, Stanford InfoLab.

Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139.

Liu, Y. and Zheng, C. (2018). Auto-encoding knockoff generator for FDR controlled variable selection. *arXiv preprint arXiv:1809.10765*.

Mai, Q. and Zou, H. (2013). The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*, 100(1):229–234.

Mallows, C. (1973). Some comments on $C_p$. *Technometrics*, 15(4):661–675.

Miller, A. (2002). *Subset selection in regression*. CRC Press.

Nguyen, T.-B., Chevalier, J.-A., Thirion, B., and Arlot, S. (2020). Aggregation of multiple knockoffs. In *International Conference on Machine Learning*, pages 7283–7293. Proceedings of Machine Learning Research.

O'Hara, R. B. and Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, 4(1):85–117.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Ren, Z., Wei, Y., and Candès, E. (2021). Derandomizing knockoffs. *Journal of the American Statistical Association*, pages 1–11.

Romano, Y., Sesia, M., and Candès, E. (2020). Deep knockoffs. *Journal of the American Statistical Association*, 115(532):1861–1872.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, pages 461–464.

Shao, J. (1996). Bootstrap model selection. *Journal of the American statistical Association*, 91(434):655–665.

Spector, A. and Janson, L. (2022). Powerful knockoffs via minimizing reconstructability. *The Annals of Statistics*, 50(1):252–276.

Sudarshan, M., Tansey, W., and Ranganath, R. (2020). Deep direct likelihood knockoffs. *Advances in Neural Information Processing Systems*, 33:5036–5046.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):385–395.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, pages 307–333.

Xie, H., Li, J., Zhang, Q., and Wang, Y. (2016). Comparison among dimensionality reduction techniques based on random projection for cancer classification. *Computational Biology and Chemistry*, 65:165–172.

Zhang, P. (1993). Model selection via multifold cross validation. *The Annals of Statistics*, pages 299–313.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

Zhu, L.-P., Li, L., Li, R., and Zhu, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496):1464–1475.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

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

**AFDELING**
Straat nr bus 0000
3000 LEUVEN, BELGIË
tel. + 32 16 00 00 00
fax + 32 16 00 00 00
www.kuleuven.be