2024

# KNOCKOFF METHODS FOR NONLINEAR FEATURE SELECTION IN DATA WITH CATEGORICAL FEATURES

Behrooz Khalil Loo
*University of Rhode Island*, behroozkhalillo@uri.edu

# KNOCKOFF METHODS FOR NONLINEAR FEATURE SELECTION IN DATA WITH CATEGORICAL FEATURES

BY

BEHROOZ KHALIL LOO

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

STATISTICS

UNIVERSITY OF RHODE ISLAND

2024

MASTER OF SCIENCE THESIS

OF

BEHROOZ KHALIL LOO

APPROVED:

Thesis Committee:

Major Professor Guangyu Zhu
Haihan Yu
Xiaowei Xu

Brenton DeBoef
DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND
2024

## ABSTRACT

This thesis addresses the challenge of nonlinear feature selection in datasets that include categorical features. Conventional feature selection methods often struggle with nonlinear relationships and are ineffective in handling categorical variables. This limitation leads to suboptimal model performance and interpretability issues. Therefore, there is an urgent need to develop methodologies that can robustly handle nonlinearities and categorical features simultaneously.

To tackle this problem, this thesis proposes and explores novel knockoff methods. Knockoff methods have shown promise in feature selection tasks by generating "knockoff" features that mimic the statistical properties of the original features, enabling robust variable selection while controlling the false discovery rate (FDR). In this work, knockoff methods are applied to datasets with categorical features, leveraging advanced statistical techniques to handle the unique challenges posed by categorical variables in nonlinear feature selection.

The findings of this thesis demonstrate the efficacy of the proposed knockoff methods in addressing linear and nonlinear feature selection tasks that involve categorical data. Through comprehensive simulation, we show that the knockoff methods outperform traditional approaches in terms of both FDR and power. Additionally, the methods exhibit robustness across different types of relationships, including linear, nonlinear, and categorical feature distributions,

highlighting their versatility and effectiveness in real-world data analysis scenarios.

## ACKNOWLEDGMENTS

I would like to express my heartfelt gratitude to my major professor, Prof. Guangyu Zhu, for his unwavering support and guidance during my thesis. Professor Zhu continuously encouraged me and was always willing and enthusiastic to assist me throughout my research.

## TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# CHAPTER 1
# INTRODUCTION

## 1.1 Feature Selection in the Era of Big Data

In the era of big data, extracting meaningful insights from complex datasets is paramount across various domains, including healthcare, finance, marketing, and beyond. Central to this endeavor is the feature selection process, where a subset of relevant features is identified from a pool of potential predictors to improve model performance, interpretability, and generalization.

Traditional feature selection methods have predominantly focused on linear relationships between features and the target variable, often assuming Gaussian distributions and continuous variables.

However, in many real-world scenarios, datasets exhibit nonlinear relationships and contain categorical features, posing significant challenges for conventional feature selection techniques. This necessitates the development of novel methodologies capable of simultaneously handling nonlinearities and categorical variables.

The exponential growth of data collection capabilities has led to an unprecedented increase in the dimensionality and complexity of datasets.

While this abundance of data holds great promise for extracting valuable insights, it also presents challenges in computational complexity, model overfitting, and interpretability.

Feature selection is a crucial preprocessing step to mitigate these challenges by identifying the most informative subset of features, thereby reducing dimensionality and improving model efficiency and interpretability.

The knockoff feature selection procedure can be outlined in three main steps as follows:

- Generating Knockoffs, $\tilde{X}$

The first step entails generating knockoff versions of the original data features. These knockoff variables are constructed to emulate the distribution of the original features while maintaining conditional independence from the response variable when given the original features. This generation serves as a basis for a controlled feature selection process, aiming to differentiate between truly influential features and those that are statistically indistinguishable from noise.

- Variable Importance Measures, $W$

Upon the creation of knockoff features, the second step is to compute variable importance measures for both the original and knockoff features. These measures, denoted $W$, assess the impact of each feature on the model's predictive accuracy. The importance is typically quantified by evaluating how the inclusion or exclusion of a feature affects the model performance, such as changes in prediction error or likelihood.

- Feature Selection, $S$

The final step involves selecting features based on the calculated importance measures. The selection criterion involves choosing a threshold that balances the number of selected variables with the goal of controlling the FDR at a pre-specified level. Features are selected if their importance measures exceed the threshold, indicating a stronger influence on the model's performance compared to their knockoff counterparts. This methodologically rigorous approach helps to ensure that the features included in the final model are statistically validated to have a true impact on the response variable, thereby minimizing the risk of model overfitting and enhancing the reliability of inference.

## 1.2 Limitations of Traditional Feature Selection Methods

Traditional feature selection methods have been widely employed in various domains. However, these methods often assume linear relationships between features and the target variable, limiting their effectiveness in capturing nonlinear patterns in many real-world datasets.

Moreover, most traditional techniques must be equipped to handle categorical variables prevalent in healthcare, marketing, and social sciences. Ignoring features' nonlinearities and unconditional nature can lead to suboptimal model performance, biased feature selection, and erroneous conclusions.

# CHAPTER 2

# REVIEW OF LITERATURE

The knockoff filter is a new variable selection procedure controlling the False Discovery Rate (FDR) proposed by [1]. The knockoff filter procedure introduces a set of control variables, known as knockoffs, for each of the original variables in a dataset. These knockoffs are designed to mimic the correlation structure of the original variables but are statistically independent of the response variable. The key idea is that if an original variable is truly associated with the response, it should stand out against its knockoff counterpart. First, a corresponding knockoff is created for each original variable in the dataset. The creation of knockoff variables is a crucial step that involves generating new variables that follow the same correlation pattern as the original variables but are otherwise independent of the response variable. This is done without collecting new data, which makes the process efficient. Once the knockoffs are created, both the original and knockoff variables are included in a regression model to predict the response variable. The model then assigns a score to each variable, reflecting its importance or association with the response. The next step involves comparing the scores of the original variables with their knockoff counterparts. If an original variable has a higher score than its knockoff, it suggests the variable is likely to be truly associated with the response. Conversely, if a knockoff scores higher, it suggests that the original variable's association could be due to chance. The procedure uses these comparisons to

decide which variables to select to control the FDR. It sets a threshold that balances the number of selected variables with the need to keep the rate of false discoveries low. Only variables that score significantly higher than their knockoffs are selected, ensuring that the FDR is controlled. The beauty of the knockoff filter lies in its ability to provide exact control of the FDR in finite samples, regardless of the number of variables or the complexity of the model. It does not require any assumptions about the distribution of the null variables (those not associated with the response) or the proportion of true associations.

An extension of the knockoff framework to a broader context, particularly for non-linear models and high-dimensional settings where the response variable's conditional distribution is unknown, is done by [2]. They outline that traditional methods for building interpretable models that link a large set of covariates to a response, especially in a non-linear fashion (like binary responses), do not effectively control the fraction of false discoveries. This is a significant challenge in high-dimensional logistic regression and other non-linear models. The 'Model-X' knockoff framework is introduced to solve this problem. Unlike the original knockoff procedure limited to homoscedastic linear models with more observations than predictors ($n \geq p$), the model-X knockoffs can provide valid inference from finite samples without any knowledge of the response's conditional distribution, regardless of the number of covariates. This is achieved by constructing knockoff variables probabilistically rather than geometrically, which requires the covariates to be random (independent and identically distributed rows) with a known distribution.

The procedure proposed by [3] for analyzing high-dimensional microbiome data involves a two-step compositional knockoff filter designed to control the FDR while accounting for the compositional nature of the data. The first step of the procedure is a screening process to reduce the dimensionality of the data. In this step, insignificant microbial taxa are filtered out. This is done while maintaining the critical sum-to-zero constraint inherent in compositional data, ensuring that all microbial taxa sum is relative to a constant total. This step is crucial because it simplifies the subsequent analysis by focusing only on the taxa that have a higher likelihood of being truly associated with the response variable. The second step extends the knockoff filter to the selected microbial taxa from the first step. The knockoff filter is a statistical method that creates fake "knockoff" versions of each variable (in this case, microbial taxa) and then compares the importance of the real and knockoff variables in predicting the response variable. If a real variable is consistently more important than its knockoff, it will likely be truly associated with the response. This step is adapted to handle the compositional data by constructing knockoff variables that respect the sum-to-zero constraint. Combining these two steps, the procedure selects a subset of microbial taxa that are likely relevant to the response variable while controlling the FDR. This means that the proportion of false positives among the selected taxa is kept below a pre-specified threshold, which is essential for ensuring the reliability of the findings in high-dimensional data analysis.

A derandomization technique by combining selection outcomes from multiple runs of the knockoff's algorithm was introduced by [4]. This

derandomization step is adaptable and can be applied to any underlying variable selection procedure, ensuring stable decisions without compromising statistical power. They prove that derandomized knockoffs controls both the per-family error rate (PFER) and the k family-wise error rate (k-FWER)

In [5], adaptive knockoff filters build upon the knockoff procedure and utilize both the available data and additional side information to dynamically order the variables of interest and concentrate on those that show the most promise. The key advantage of adaptive knockoffs is their ability to effectively control the finite-sample FDR.

An innovative deep learning framework that uses a combination of feature selection models to find predictive clusters without requiring predefined candidate groups is done by [6]. The framework involves a collection of group selection models and an aggregate predictor to uncover sets of features. It operates similarly to an ensemble of "weak" feature selection models, each attempting to solve the task using a sparse set of features. These models are trained to identify distinct predictive groups, and further details about their combination and training process are discussed.

A new feature selection algorithm for DNNs by integrating the knockoff technique and the distribution information of irrelevant features is proposed by [7]. With the help of knockoff features and central limit theorem, they state that the irrelevant feature's statistic follows a known Gaussian distribution under a few mild conditions. They assume that $S_0$, irrelevant is existing and unique. With

the distribution of irrelevant feature's statistics, they use the hypothesis-testing to decide which feature is informative.

The paper [8] proposes a new method for high-dimensional selective inference called the knockoff filter. The method is designed to control the FDR in high-dimensional linear models, where the number of variables p is much larger than the number of observations n. The key idea behind the knockoff filter is to construct knockoff variables that mimic the correlation structure of the original variables and then use these knockoff variables to compare the coefficients of the original variables and control the FDR. The paper also introduces the concept of the sign-restricted Lasso penalty, which is used to determine the statistics for the knockoff filter. In other words, they are running the same penalized least-squares optimization but with the added restriction that they will only select the $jth$ feature or $jth$ knockoff feature if its estimated effect direction agrees with the sign information gathered at the screening stage. In the screening step, they split the n observations into two disjoint groups of size, $n_0$ and $n_1 = n - n_0$, denoted as $(X_{(0)}, y_{(0)})$ and $(X_{(1)}, y_{(1)})$, respectively. They then use $(X_{(0)}, y_{(0)})$ to identify a subset $\hat{S}_0 \subset [p]$ of potentially relevant features, such that $|\hat{S}_0| < n_1$. In the selection step (splitting), they ignore any features that were discarded in the screening step and run the knockoff procedure on the remaining data, that is, on $(X_{(1)}^{\hat{S}_0}, y_{(1)})$. The purpose of the screening step is to reduce the number of features to be considered in the knockoff procedure, which can improve the power of the procedure.

The contribution of [9] lies in proposing a novel approach to model functionality stealing, where an adversary can transfer the functionality of a victim model into a knockoff via black-box access.

In [10], they used directed acyclic graphs to Identify nonlinear between variables by implementing deep learning for variable selection (DAG-deepVASE). They fit the original features (except $V_1$) and related Knockoff features as input and $V_1$ as output to identify its association with other variables. DAG-deepVASE will run this model with each of the other variables ($V_1$, $V_2$, …, $V_M$) as response and with all the other variables as input. They tackle the issue of threshold selection with a two-step procedure named PC-Knockoff. In the first step, they apply the PC-Screen method to obtain an over-fitted subset of moderate size from the ultrahigh-dimensional features. In the second step, they construct knockoff counterparts for the features which survive in the first step.

The [11] presents a novel model-free feature screening method based on projection correlation, utilizing knockoff features to control the false discovery rate in ultra-high dimensional data without relying on a specific regression model.

The introduction of a machine for sampling approximate model-X knockoffs using deep generative models for arbitrary and unspecified data distributions is done by [12]. The core idea of the work is to iteratively refine a knockoff sampling mechanism until a criterion inspired by the maximum mean discrepancy in machine learning is optimized. This criterion essentially measures the distance to pairwise exchangeability between original and

knockoff features. By leveraging the model-X framework, they have developed a flexible and model-free statistical tool for controlled variable selection.

The [13] focuses on feature selection in high-dimensional data analysis, particularly addressing group structure among features, which is common in various scientific problems. They propose a new Deep Neural Network (DNN) architecture that integrates with the knockoff technique for performing nonlinear group-feature selection. This approach aims to control the group-wise False Discovery Rate (gFDR). Their method is shown to be effective in high-dimensional synthetic data, achieving high power and accurate gFDR control compared to state-of-the-art methods. The paper highlights that the performance of Deep-gKnock is particularly superior in situations involving nonlinearity, high dimensionality (where the number of dimensions p exceeds the sample size n), high between-group correlation, high within-group correlation, and many associated groups. Additionally, Deep-gKnock is robust to feature distribution misspecification and changes in network architecture.

The [14] proposes a practical algorithm for generating knockoffs and presents a heuristic multiple knockoffs approach for assessing the robustness of the selection process. The proposed algorithm can be used to analyze more general data sets involving a mixture of continuous and binary explanatory variables. They validate their methodology through simulations and demonstrate its utility on a large clinical data pool. The paper also discusses the limitations of the knockoff approach and suggests alternative screening methods for handling large numbers of variables.

In [15], they present a novel approach to feature selection by leveraging Generative Adversarial Networks (GANs) to generate knockoff features. This model is designed to work efficiently without assumptions on the feature distribution, making it versatile for various data types.

DeepDRK proposes a deep learning-based approach to feature selection by incorporating dependency regularization into the knockoff framework [16]. Their method aims to balance the control of FDR and the power of feature selection by using novel regularization techniques and a multi-swapper design. This approach enhances the reliability and effectiveness of the knockoff mechanism in selecting significant features, especially in complex datasets where dependencies among features are strong.

A reinforced agent-based method was introduced for feature selection that uses knockoff variables to guide the selection process [17]. The method relies on a single pre-trained reinforced agent, simplifying the feature selection process while ensuring efficiency and accuracy. The reinforced agent makes the method adaptive and capable of handling various types of data distributions, making it a robust choice for feature selection tasks.

The paper proposes an error-based knockoff inference method that integrates knockoff features with error-based feature importance statistics [18]. This approach does not require specifying a regression model and offers theoretical guarantees on controlling the FDR, false discovery proportion (FDP), and k-familywise error rate (k-FWER). The method's adaptivity and flexibility

make it suitable for high-dimensional settings, and it has shown competitive performance in empirical evaluations.

In [19], they present a methodology for derandomizing Model-X knockoffs with guaranteed FDR control. This approach addresses the inherent randomness of the Model-X knockoffs method, which can lead to different sets of selected variables in different runs on the same dataset, a feature considered undesirable in practice. The authors propose using e-values, which are advantageous in multiple testing scenarios due to their dependency only on expected values and not on the dependence structure among tests. By aggregating e-values from multiple realizations of knockoff procedures, the authors derive a derandomized procedure that maintains control over the FDR without requiring additional conditions. This derandomized approach not only retains control over the FDR but also reduces selection variability and maintains power comparable to traditional Model-X knockoffs.

In [20], they delve into the logistic regression model, focusing on sparse high-dimensional settings. They explore a tradeoff between false discovery and true positive rates, particularly in the context of regularized logistic regression models. The authors aim to provide insights into improving variable selection reproducibility through a tradeoff function, which they apply to sample size calculations and calibration of the FDR for enhanced power consideration.

The work [21] develops selective inference methods for group lasso estimators, suitable for a broad range of distributions and loss functions. They introduced a randomized group-regularized optimization problem and

constructs a post-selection likelihood for conditional selective inference, addressing the uncertainty introduced by variable selection methods like the lasso. Their methodology is demonstrated using data from the national health and nutrition examination survey, with simulations highlighting its advantages over other methods.

# CHAPTER 3
# METHODOLOGY

## 3.1 Knockoff Methodology

The knockoff filter was proposed in 2015 [5]. It is a general framework for controlling the FDR when performing variable selection. The idea is to be able to discover the truly associated predictors with the response variable. The knockoff filter generates knockoff variables designed to mimic the correlation structure found within the original data. Creating knockoffs is cheap, and their construction does not require collecting new data. The knockoffs serve as negative controls, and they allow the identification of the important variables related to the response variable while controlling the expected fraction of the false discovery proportion - FDR. The knockoff method selects the original variables that are better than their corresponding knockoff copies based on some measures of feature importance that can be computed with various popular methods. The knockoff filter has been used and shown to ensure accurate FDR control, which traditional methods cannot achieve.

## 3.2 Generating Knockoffs

Generating Model-X knockoffs involves creating synthetic versions of data, known as "knockoffs," that mimic the properties of the original dataset while not carrying any of the original data's actual information. This process is particularly

valuable in scenarios where preserving the privacy of the original data is crucial, such as in sensitive medical, financial, or personal datasets. The primary goal of generating knockoffs is to enable researchers and data scientists to conduct analyses, model testing, and feature selection without risking exposure of the genuine data.

Candes et al. (2018) proposed the Model-X knockoff framework, a more flexible approach valid regardless of the distribution of $Y \mid X$, and we sample knockoffs from the conditional distribution $\tilde{X} \mid X$ . In the Model-X framework, a knockoff $\tilde{X}$ for X satisfies the following properties:

(1) for any subset $S \subset \{1, \ldots, p\}$

$$(X, \tilde{X})_{Swap(S)} \overset{\text{def}}{=} (X, \tilde{X})$$

(2) $\tilde{X} \perp\!\!\!\perp Y \mid X$ if there is a response Y. (2) is guaranteed if $\tilde{X}$ is constructed without looking at Y. The vector $(X; \tilde{X})_{\text{swap(S)}}$ is obtained from $(X; \tilde{X})$ by swapping the entries $X_j$ and $\tilde{X}_j$ for each $j \in S$.

With properties (1) and (2) satisfied, we can use knockoffs to guarantee FDR control at a given level $q \in (0, 1)$.

### 3.3 Variable Importance Measures

The next step is calculating feature statistics $(W_j)$ for each variable $(j)$. These statistics serve as a measure to differentiate between the original variable $(X_j)$ and its corresponding knockoff $(\tilde{X}_j)$. The calculation of $W_j$ is based on a predefined function $W_j$ that takes into account both the original and knockoff

variables $(X, \tilde{X})$ along with the response variable (Y). The choice of function $W_j$ depends on the specific methodological framework but aims to quantify the strength and nature of the association between each variable (and its knockoff) and the response variable.

One common approach to computing $W_j$ is to use the difference in Lasso regression coefficients obtained from an augmented regression model that includes both the original variables and their knockoffs as predictors of Y. Specifically, $W_j$ might be calculated as the absolute difference between the absolute value of the lasso regression coefficient of $X_j$ $(\beta_j)$ and that of its knockoff $(\beta_j)$, i.e., $W_j = |\beta_j| - |\tilde{\beta}_j|$. This difference reflects the relative importance of the original variable over its knockoff in explaining the variation in Y. Large, positive values of $W_j$ indicate a stronger association of the original variable with the response, suggesting that the variable is likely to be a true predictor rather than a false discovery.

It provides a basis for systematically assessing which variables have a genuine link to the response variable. By comparing the magnitudes of $Wj$ across all variables, one can rank variables according to their importance and apply a thresholding procedure to select a subset of variables that are most likely to be truly associated with Y while controlling the FDR.

## 3.4 Feature Selection

The Knockoffs+ procedure is a method designed to control the FDR. The key step in the Knockoffs+ procedure is the selection of variables based on feature statistics ($W_j$) computed for each variable and its knockoff. The selection criterion involves choosing a threshold ($\tau_+$) that balances the number of selected variables and the goal of controlling the FDR at a pre-specified level (q).

The variables to be selected ($S$) are determined as follows:

$\hat{S} = \{j: W_j \geq \tau_+\}$ Where $\tau_+$ is chosen based on a specific criterion aimed at controlling the FDR. It is defined as:

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

where $q$ is the target FDR level. The $\tau_+$ is the smallest positive threshold for which the ratio of the number of negative feature statistics exceeding $\tau_+$ in magnitude (plus one) to the number of positive feature statistics exceeding $\tau_+$ does not exceed the target FDR level $q$.

## 3.5 Variable Selection via Knockoffs using LASSO and XGB.

In the knockoff generation step, we used a sequential conditional independent algorithm, which is as follows. For j = 1,…, p, we sample

$$\tilde{X}_j \sim \mathcal{L}\left(X_j | X_{-j}, \tilde{X}_{1:j-1}\right)$$

where $X_{-j} \coloneqq (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p)$ and $\tilde{X}_{1:j-1} \coloneqq (\tilde{X}_1, \ldots, \tilde{X}_{j-1})$

The implementation in Sequential knockoff algorithm for mixed data types is as the following.

Algorithm 1 (Sequential knockoff generation algorithm for mixed data types)

To estimate the parameters of the sequential conditional distributions, we can use the LASSO linear model or XGB at each step.

To be specific, for $j = 1, \ldots, p$:

- if $X_j$ is continuous, fit a LASSO regression model or XGB with response $X_j$ and covariates $X_{-j}$ and $\tilde{X}_{1:j-1}$.

  Sample $\tilde{X}_j \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ where $\hat{\mu}$ and $\hat{\sigma}^2$ are estimates of the mean and error variance, respectively.

- if $X_j$ is categorical, fit a penalized multinomial logistic regression model or XGB with response $X_j$and covariates $X_{-j}$ and $\tilde{X}_{1:j-1}$.

  Sample $\tilde{X}_j \sim Multinom(\hat{\pi})$, where $\hat{\pi}$ denotes the estimate of the multinomial probabilities.

Return $\tilde{X} = (\tilde{X}_1; \ldots; \tilde{X}_p)$.

# CHAPTER 4
# Simulation and Real Data Analysis

## 4.1 Simulation

In the section detailing the simulation framework for evaluating variable selection methods, a model is described that forms the basis of extensive simulation experiments. This model is meticulously crafted to simulate data sets $(X, y)$ across varied parameter configurations, enabling a rigorous comparison of different variable selection algorithms, including the novel sequential and multiple knockoff methods.

The design matrix X is generated by simulating its rows independently from a multivariate Gaussian distribution with a mean of 0 and a specified covariance matrix, $\Sigma$. The covariance matrix $\Sigma$ is structured to accommodate a range of correlation scenarios including equicorrelated, and AR1 (autoregressive) models, thus allowing the examination of the variable selection methods under diverse correlation structures among features.

To emulate real-world data scenarios where some variables are continuous and others are categorical, a subset of the columns in X is binarized using an indicator function. This binarization process transforms a portion of the continuous variables into binary or multiclass categorical variables, reflecting typical mixed-type data applications. The categorical columns are then scaled to ensure uniform marginal variances across all variables.

The response vector $y$ is simulated from a sparse Gaussian regression model, where only a fraction of the predictors (non-null features) are associated with the response. This sparsity is introduced by setting a predetermined number of regression coefficients (corresponding to the non-null features) to a specified amplitude while the rest are set to zero. The selection of non-null features is randomized for each generated data set, adding an element of variability and robustness to the simulation experiments.

This model is central to the simulation experiments conducted in the study, providing a versatile and realistic framework for assessing the performance of various variable selection methods. By simulating data sets under controlled yet varied conditions, the study aims to offer comprehensive insights into the efficacy, reliability, and applicability of the methods being evaluated. A summary of the underlying simulation parameters can be found in Table 1.

For evaluating the performance of proposed methods, we will use different types of the correlation structures named Equi-correlated and Auto-Regressive 1 (AR1) as the following.

$$\sum_{ij} = \begin{cases} \rho^{1\{i \neq j\}}/n & \text{Equicorrelated} \\ \rho^{|i-j|}/n & AR1 \end{cases}$$

Table 1. The experimental parameters for simulation experiments

| | |
|---|---|
| n | Number of data observations |
| p | Number of covariates |
| Cat_p | Percentage of binarized covariates(%*p) |
| Sparsity | Percentage of non_null covariates(%*p) |
| $\sigma$ | Standard deviation of the random noise |
| $\Sigma$ | covariance matric form either AR1 or Equi |
| $\beta$ | Regression coefficient amplitude |
| $\rho$ | Correlation coefficient |
| Min classes | Minimum number of classes for categorical columns |
| Max classes | Maximum number of classes for categorical columns |

Also, in both a Gaussian linear regression model and a nonlinear Single-Index model.

$$Y_i = X_i^T \beta + \epsilon_i, i = 1, \dots, n$$

$$Y_i = g(X_i^T \beta) + \epsilon_i, i = 1, \dots, n$$

where $Y_i \in R$ is the $ith$ response, $X_i \in R^p$ is the feature vector of the $ith$ observation, $\beta \in R^p$is the coefficient vector, $\epsilon_i \in R^p$ is the noise of $ith$ response, and $g$ is link function, $g(x) = \frac{\sqrt{2}}{2}x^2$ . We explore the effects of different key parameters in different cases and scenarios on the feature selection performance which are as the following. We provide comprehensive

simulation experiments by comparing the performance of proposed methods by state-of-the-art methods.

In our default setting, we set n=2000, p=200, Cat_p=0.25, Sparsity=0.25, σ=1, Σ =AR1, β=2.5, ρ=0.5. In case 2 with binary features, Min classes=Max classes=2 and in case 3 with categorical features, Min classes=2 and Max classes=5.

we vary one setting and keep the other parameters at their default value in each experiment. For each setting, we run each experiment for 20 replications and set the target FDR level at $q = 0.2$.

### 4.1.1 Case 1: Performance Analysis with Numerical Types

In case 1, we compare our methods with basic knockoff, named Model-X and sequential knockoffs. In this case, all features will be numerical without any binary or multiclass categorical features. We evaluate the performance on varying regression coefficient amplitude. The results for FDP and TPP are shown in Figure 1.

The comparative analysis of four distinct methods through the FDR and Power has revealed significant insights into their performance dynamics. Notably, XGB emerged as superior FDR, showcasing enhanced efficacy and reliability. This superiority is particularly pronounced in the mean FDR for XGB is consistently lower in all cases compared to the other models.

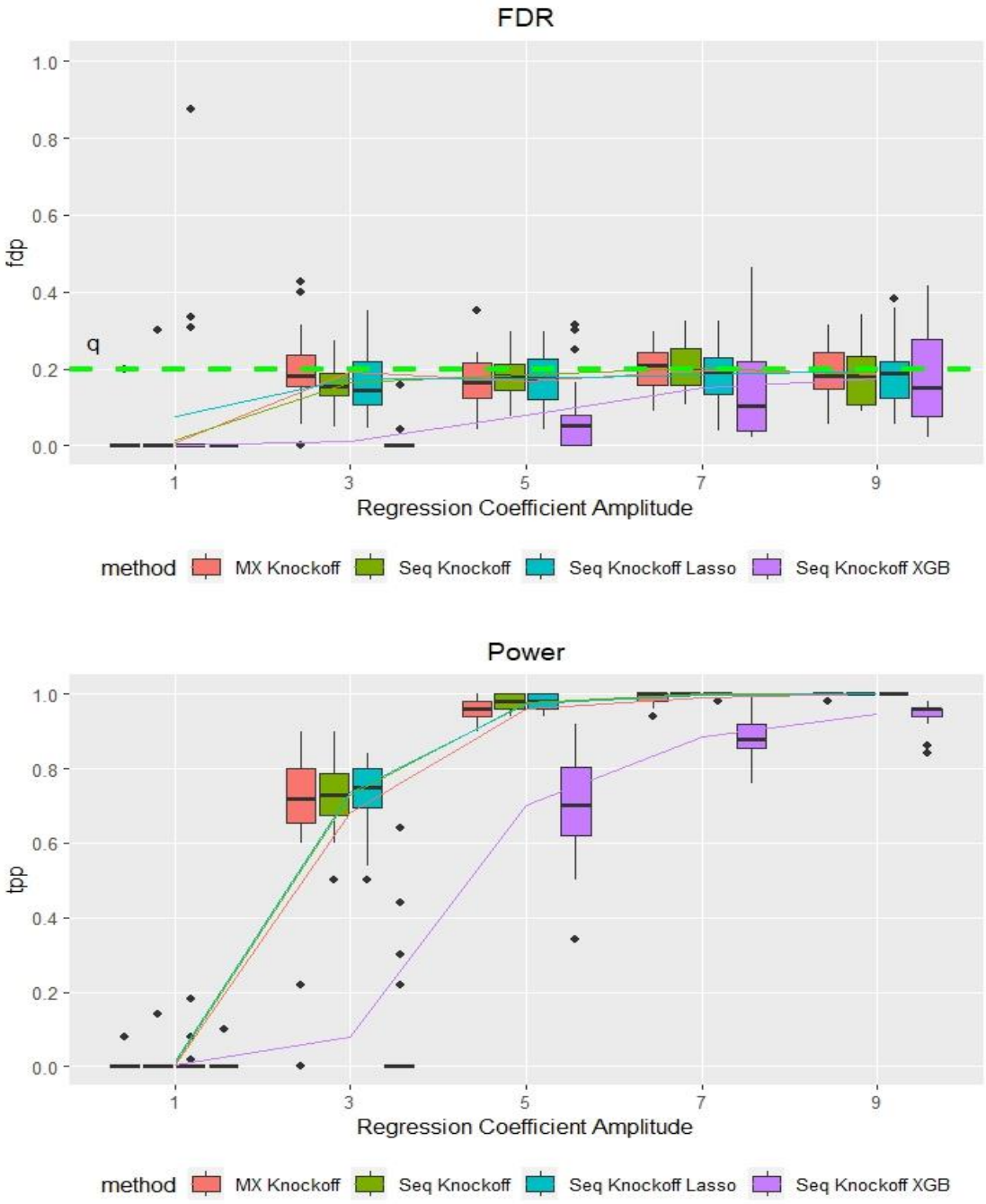


Figure 1: Performance Analysis of FDR and Power Across Varying Regression Coefficient Amplitudes, β. Colored curves represent mean estimates of FDR and Power, highlighting the relationship with coefficient amplitude variations. The horizontal dashed line in the FDR plot indicates the target FDR level.

### 4.1.2 Case 2: Performance Analysis with Mixed Numerical and Binary Data Types

In Case 2, our evaluation extends to include a comparison with the sequential knockoff method, incorporating a dataset characterized by a mix of numerical and binary features. This diverse feature set allows for a nuanced assessment of each method's adaptability and effectiveness across different data types, highlighting their strengths and limitations in a more complex analytical context. The comparison has been done in different scenarios which are follows:

#### *4.1.2.1 Varying Sparsity Equi*

This analysis employed a default setup with equi-correlated correlation while varying the sparsity levels to assess model performance. This approach allowed us to systematically evaluate the impact of sparsity on the FDRs and the power of the models under investigation. The outcomes of this setup are illustrated in Figure 2.

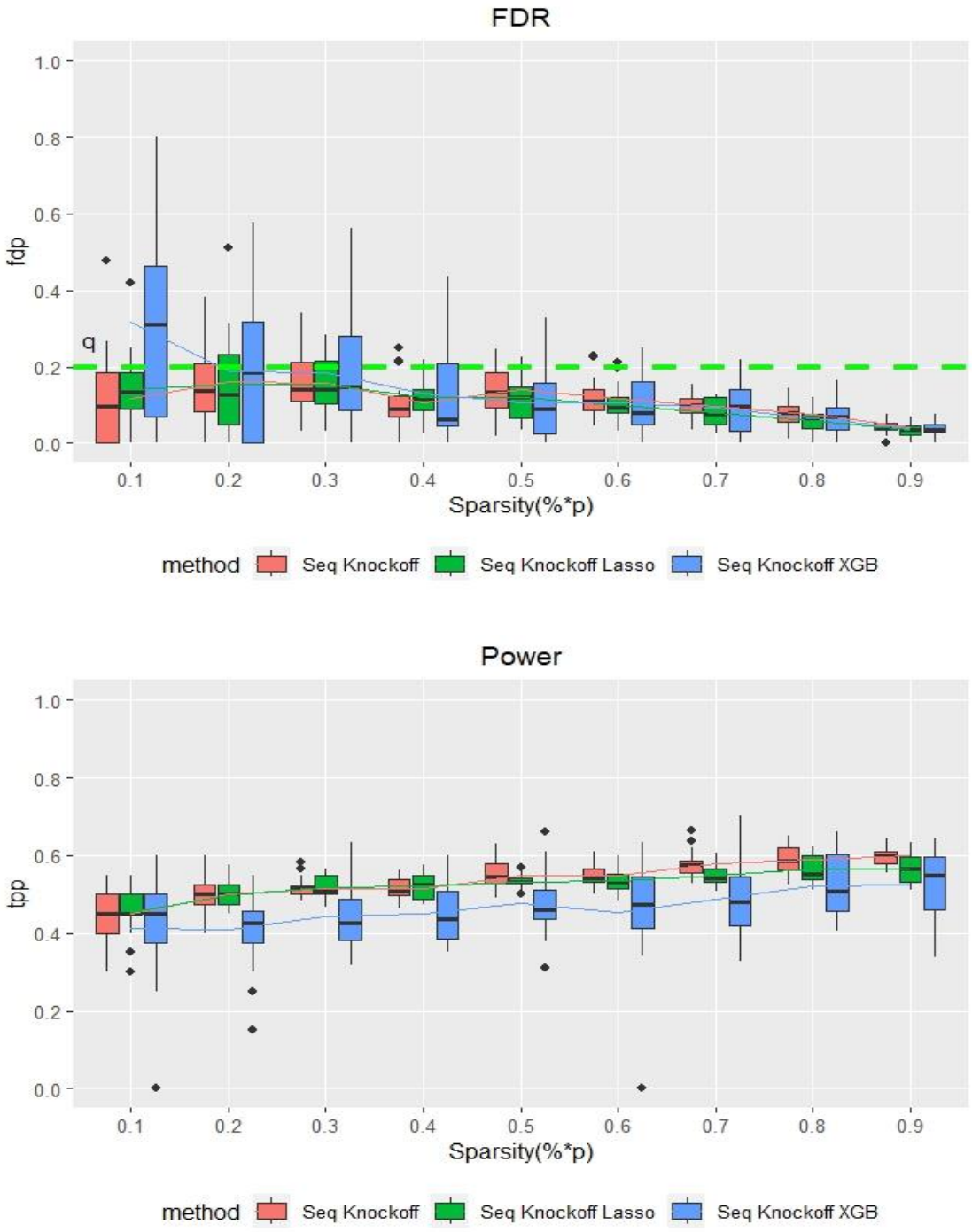


Figure 2: Performance Analysis of FDR and Power Across Varying Sparsity, percentage of non-null covariates with mixed data type including numerical and binary with equi-correlated correlation.

A significant trend was observed regarding the FDR and power in the comparative analysis of varying sparsity levels in models with equi-correlated correlation. Specifically, beyond a sparsity threshold of 0.2, all models exhibited FDR values lower than the target rate of 0.2, which indicates a general improvement in model specificity. Furthermore, the power of these models, which reflects their ability to identify true positives correctly, showed an increasing trend from 0.4 to 0.6 as sparsity increased. This suggests that the models become more effective at detecting true signals within a sparser dataset. Notably, at a sparsity level of 0.1, the XGB model presented a slightly higher mean FDR than the LASSO and Sequential models, indicating a marginally higher rate of incorrectly identifying features as significant. However, as sparsity levels rose, the mean FDR for XGB improved relative to the other models, ultimately resulting in a lower FDR. This indicates that the XGB model, despite its initial lag in performance at lower sparsity levels, adapts more efficiently to increased sparsity, thereby reducing the proportion of false positives more effectively than LASSO and Sequential knockoff. This observation underscores the adaptability and robustness of the XGB model in handling datasets with varying sparsity levels, particularly in scenarios where the preservation of model specificity is critical.

*4.1.2.2 Varying Sparsity AR1*

For this setup, we utilized a standard configuration with AR1 correlation, adjusting the sparsity levels to examine the effects on the models' performance. This method enabled a structured investigation into how changes in sparsity influence the models' FDRs and effectiveness. Figure 3 showcases the result obtained from this configuration.

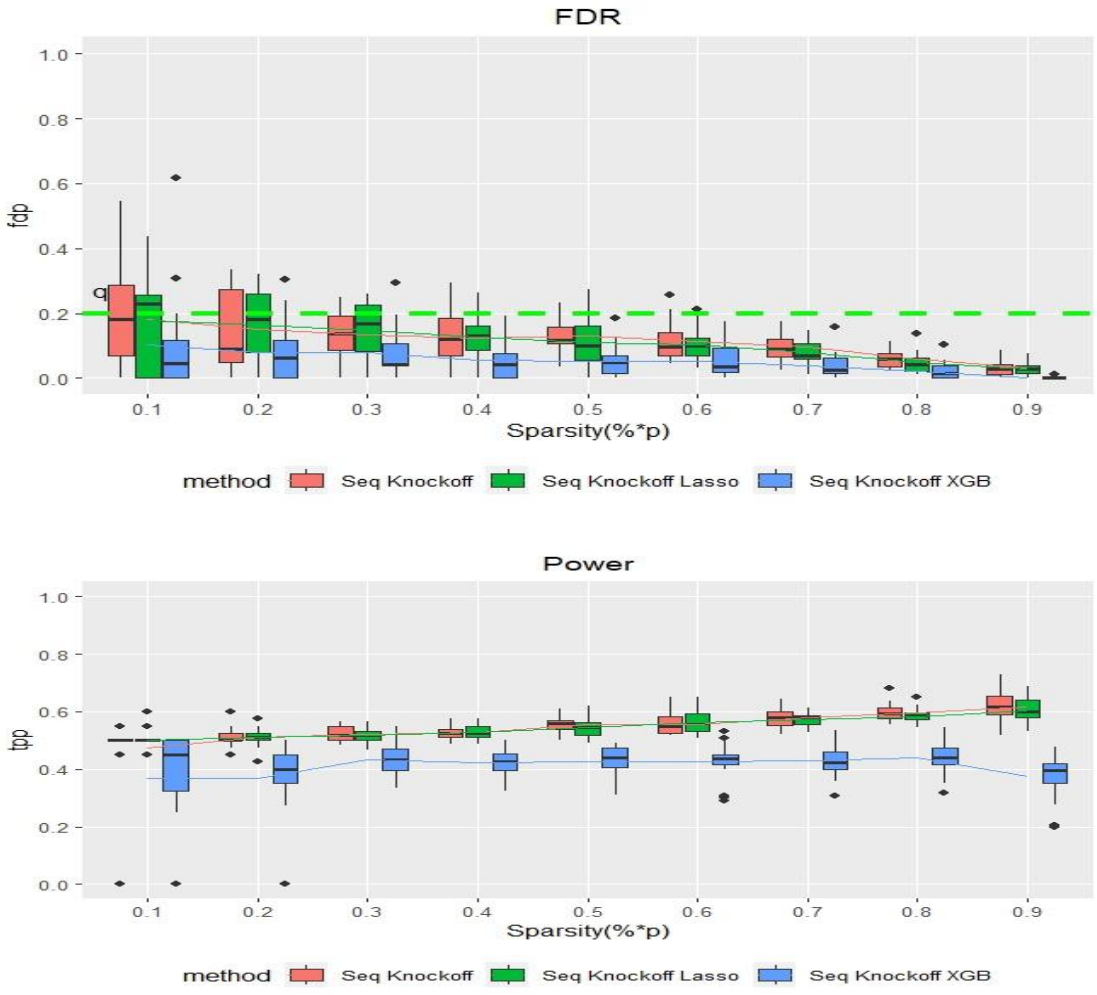


Figure 3: Performance Analysis of FDR and Power Across Varying Sparsity, percentage of non-null covariates with mixed data type including numerical and binary with AR1 correlation.

In an AR1 correlation setting, the XGB model notably outperforms Sequential and LASSO methods in terms of maintaining a lower FDR across varying levels of sparsity, consistently staying below the targeted FDR and showcasing its robustness in minimizing false positives. However, its power, or the ability to accurately identify true positives, remains moderate, ranging from 0.4 to 0.5. On the other hand, while Sequential and LASSO methods exhibit higher FDRs compared to XGB, indicating a greater tendency to incorrectly identify non-significant features as significant, they compensate with a slightly higher power in the range of 0.5 to 0.6. This nuanced performance landscape underscores the critical trade-offs between specificity and sensitivity in model selection, especially in sparse datasets with AR1 correlation, where balancing false discovery minimization and true signal detection becomes paramount for reliable statistical analysis.

### 4.1.3 Case 3: Performance Analysis with Mixed Numerical, Binary, and Multiclass Categorical Data Types

In case 3 of our study, we extend our comparative analysis to encompass models handling mixed data types, including Numerical, Binary, and Multiclass Categorical variables. This comprehensive approach allows for a more realistic assessment of model performance in practical scenarios where data often come in various formats. To thoroughly evaluate the adaptability and efficiency of our methods in this mixed-data environment, we introduce a range of varying

parameters that significantly influence model behavior and outcomes, including non-linearity, to better capture complex relationships within the data.

By altering these parameters, we aim to uncover nuanced insights into how each model copes with the complexity introduced by mixed data types and the specific challenges each parameter presents. For instance, varying the degree of sparsity tests the models' ability to handle sparse data efficiently and their robustness against overfitting and underfitting in different sparsity scenarios. Similarly, by adjusting the noise level, we can assess the models' sensitivity and resilience to irrelevant or misleading information, which is crucial for ensuring the reliability of the model's predictions. The addition of non-linearity allows us to explore each model's capability to model complex, non-linear relationships inherent in many real-world datasets, further testing the limits of their adaptability and predictive power.

This comprehensive evaluation, set against the backdrop of mixed data types and a spectrum of varying parameters, including non-linearity, not only enhances our understanding of each model's strengths and limitations but also guides the development of more robust and adaptable data analysis methods suited to the complex nature of real-world data. This holistic approach ensures that our comparative analysis remains relevant and applicable across various practical scenarios, providing valuable insights into the optimal utilization of various models in diverse data environments.

*4.1.3.1 Varying Noise Level Non-Linear*

Incorporating varying levels of noise and non-linearity into our analysis allows us to examine the models' resilience and adaptability to more complex and realistic data conditions. The findings from this arrangement are represented in Figure 4.

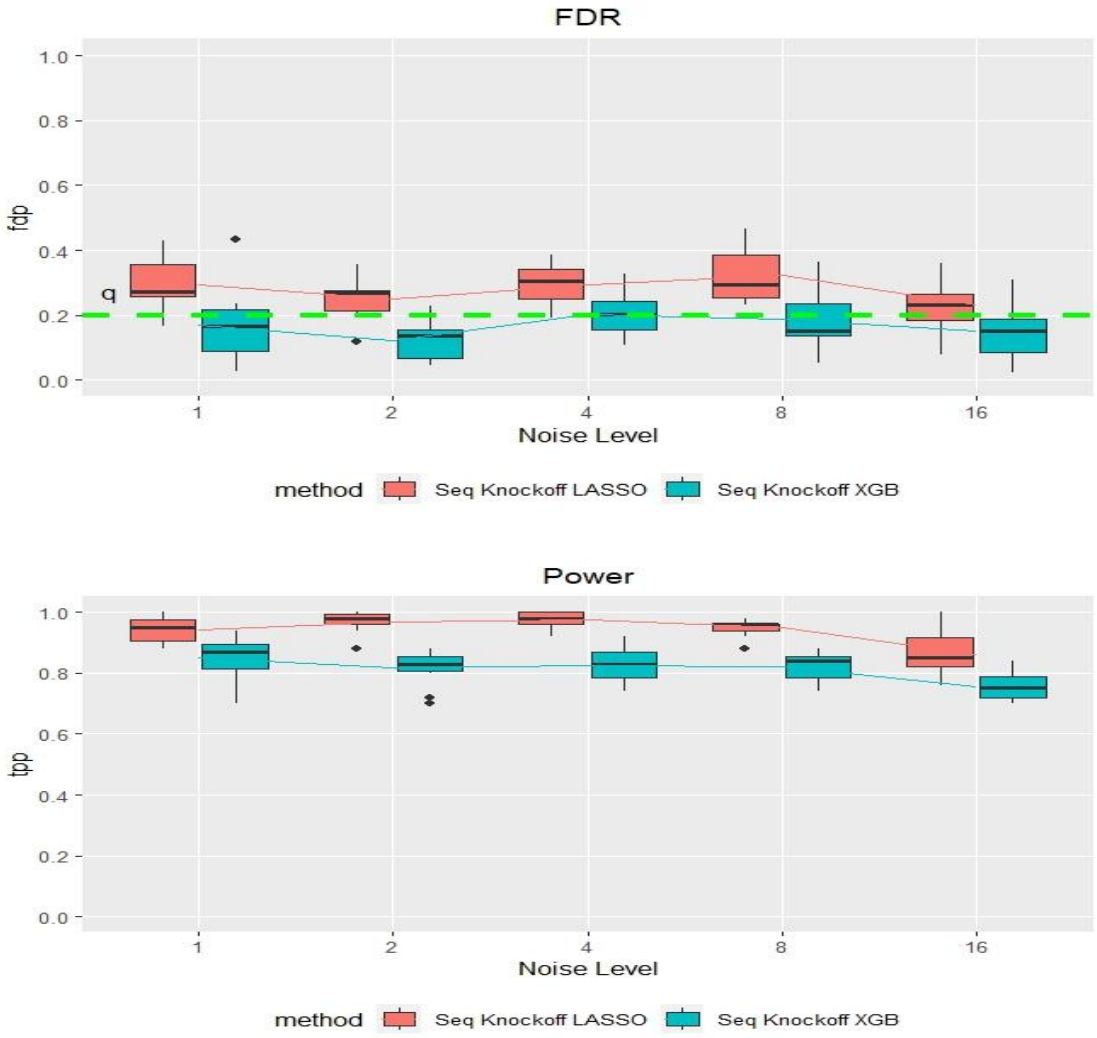


Figure 4: Performance Analysis of FDR and Power Across Varying Noise level with mixed data type including numerical, binary, and categorical with non-linearity.

In analyzing the XGB and LASSO models, a key difference lies in their FDR and power. XGB stands out for its lower FDR, surpassing the target and LASSO, indicating its strength in reducing false positives. However, LASSO exhibits higher power, ranging from 0.85 to 1, compared to XGB's 0.75 to 0.85, suggesting LASSO's slight advantage in detecting true positives. This contrast underscores the importance of model selection based on specific analytical needs, with XGB prioritizing specificity and LASSO leaning towards sensitivity.

#### *4.1.3.2 Varying CatP Non-Linear*

In an analytical setting where we adjust the percentage of categorical features (CatP) within a non-linear framework, the focus is on assessing model performance amidst evolving data complexities. This increment in CatP tests the models' adaptability to a diverse mix of variable types and their capability to handle non-linear associations effectively. The exercise aims to mirror real-world data scenarios, highlighting how well each model can maintain accuracy as the categorical dimension of the dataset intensifies, thereby offering insights into their robustness and applicability in varied analytical situations. The results of this configuration are captured in Figure 5.

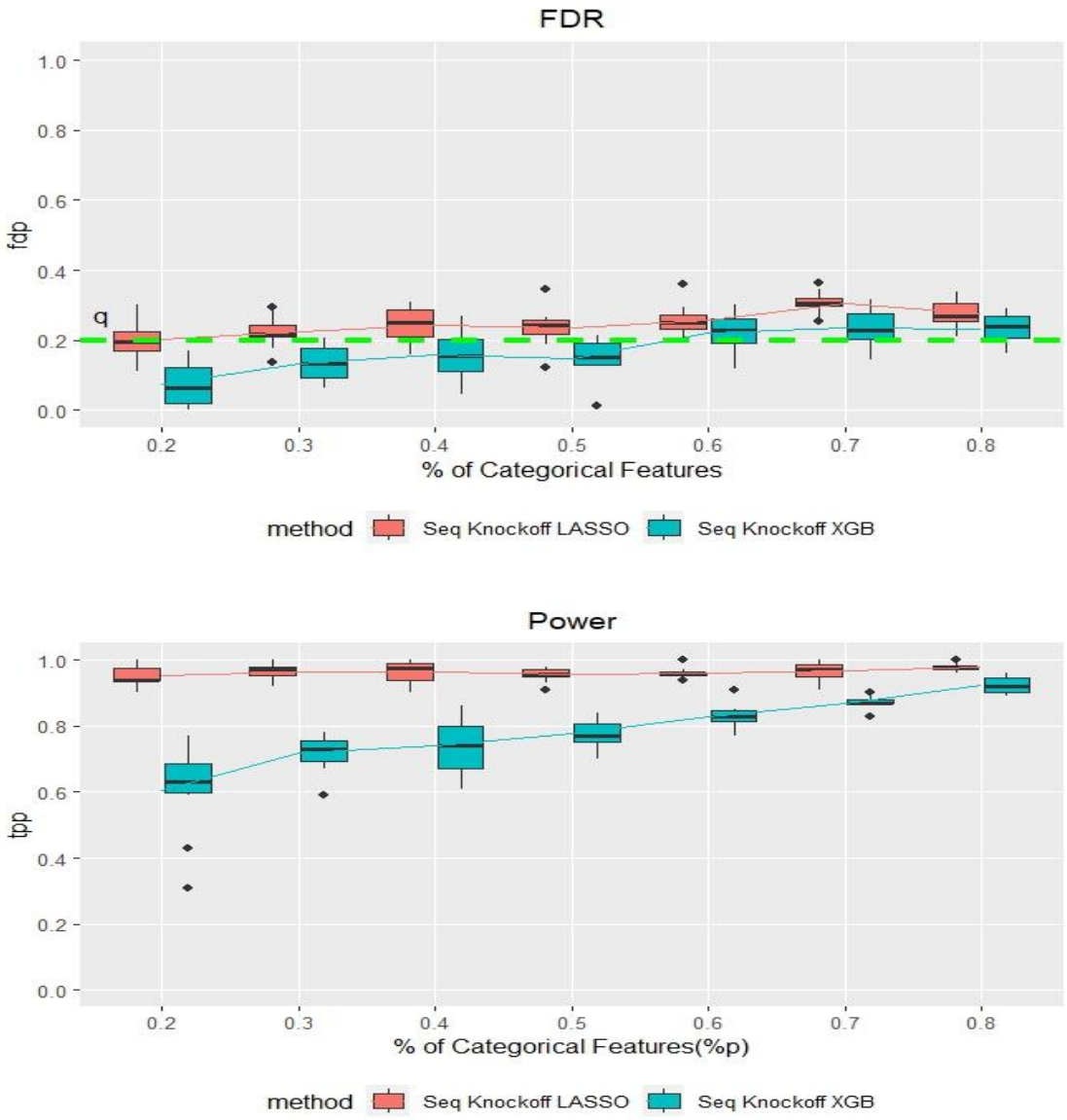


Figure 5: Performance Analysis of FDR and Power Across Varying the percentage of categorical features with mixed data type including numerical, binary, and categorical with non-linearity.

In an analytical setting where we adjust the percentage of categorical features (CatP) within a non-linear framework, the focus is on assessing model performance amidst evolving data complexities. This increment in CatP tests

the models' adaptability to a diverse mix of variable types and their capability to handle non-linear associations effectively. The exercise aims to mirror real-world data scenarios, highlighting how well each model can maintain accuracy as the categorical dimension of the dataset intensifies, thereby offering insights into their robustness and applicability in varied analytical situations.

#### *4.1.3.3 Varying Sparsity Nonlinear*

In our analysis, we utilized a standard configuration with AR1 correlation and incorporated non-linearity, adjusting the sparsity levels to examine their effects on model performance. This setup's outcomes are visualized in Figure 6.

The comparison between XGB and Lasso for feature selection shows that XGB consistently maintains the FDR below the target of 0.2. In contrast, Lasso's FDR slightly exceeds this target but improves with increased sparsity. Regarding power, Lasso exhibits higher performance, approaching a value of one. In contrast, XGB also demonstrates commendable power, with values around 0.8, indicating both methods' effectiveness in identifying relevant features, with Lasso showing a slight edge in power.

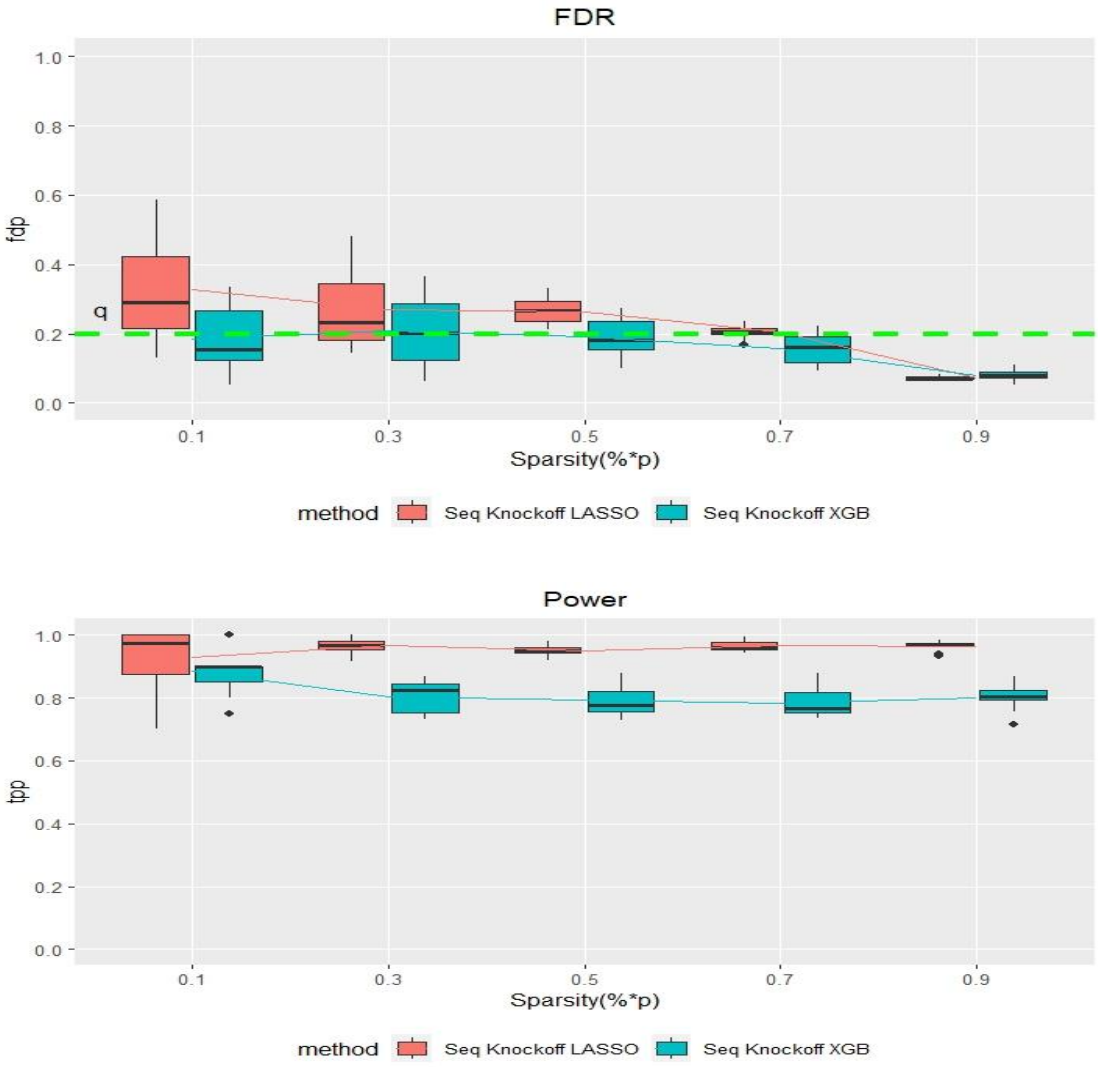


Figure 6: Performance Analysis of FDR and Power Across Varying Sparsity, percentage of non-null covariates with mixed data type including numerical, binary, and categorical with non-linearity.

4.1.3.4 Varying Sample size N Nonlinear

In our study, we employed a conventional setup that included AR1 correlation and integrated non-linearity, while also modifying the sample size to explore their impacts on the model's performance. The results of this arrangement are depicted in Figure 7.

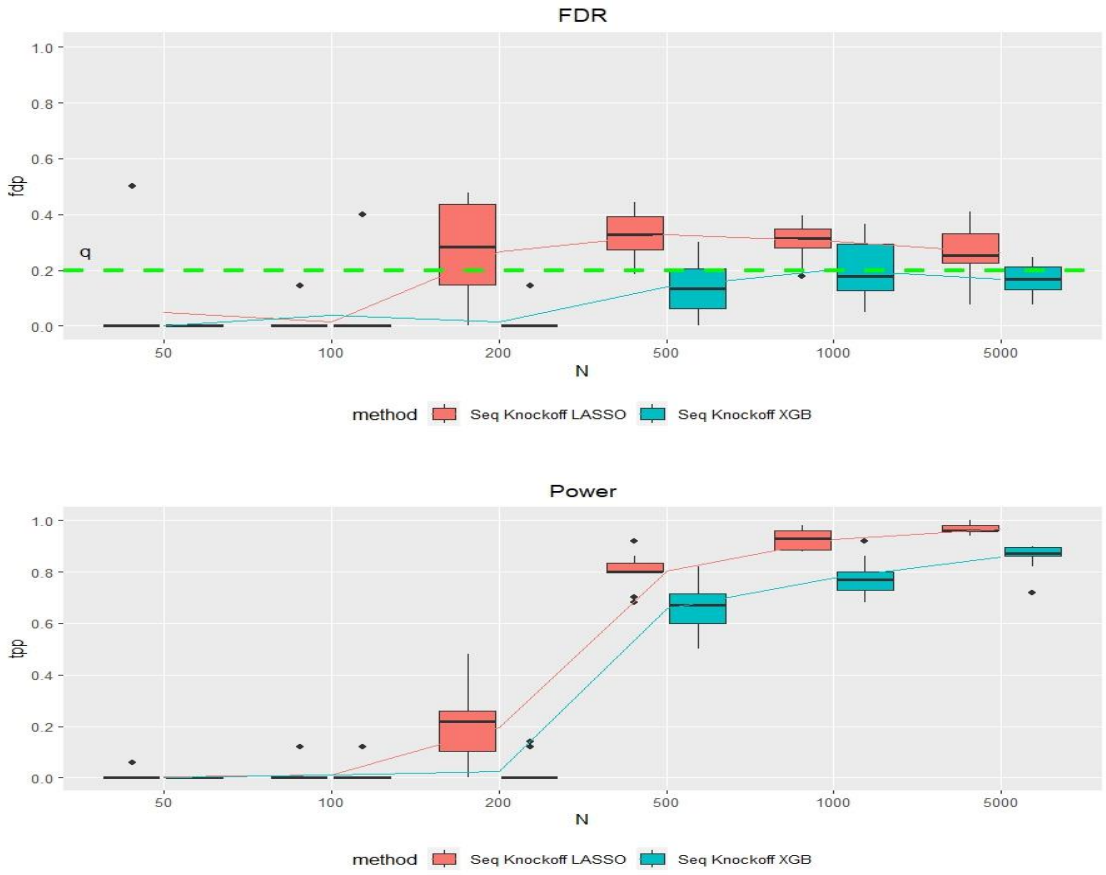


Figure 7: Performance Analysis of FDR and Power Across Varying the sample size with non-linearity.

In our analysis of statistical methods, we observed that at lower sample sizes, there is neither false discovery rate (FDR) nor statistical power, indicating a

lack of effectiveness. Specifically, the XGB method consistently maintains an average FDR below the predetermined target, with its power showing an improvement as the sample size increases. On the other hand, the Lasso method starts with an average FDR slightly above the target, which gradually decreases with larger sample sizes. Importantly, for both methods, an increase in sample size correlates with a significant enhancement in statistical power, underlining the importance of adequate sample sizes in achieving reliable and robust statistical results.

## 4.2 Real Data

In addition to conducting comprehensive simulations, our study also aims to assess the effectiveness of the proposed methodologies through their application to real-world datasets. For this purpose, the Boston Housing Price dataset, which is characterized by a mix of binary and multiclass categorical features, has been selected for detailed analysis. The initial step in our empirical investigation involves the meticulous generation of knockoff features using both the LASSO and XGB methods. Subsequently, we proceed to determine the importance statistics of the features, a process that we undertake with utmost care and precision, which in turn facilitates the process of feature selection.

We employ a fitted regression model that incorporates the selected features to evaluate the performance of the proposed feature selection

methodologies. The efficacy of this model is then benchmarked against a baseline linear regression model that utilizes all available predictors. The comparative analysis is summarized in Table 2, which presents a set of key performance metrics, including the coefficient of determination ($R^2$), mean absolute error (MAE), root mean square error (RMSE), and Akaike Information Criterion (AIC). These metrics collectively provide a comprehensive overview of the model's accuracy, error magnitude, and overall goodness-of-fit.

Table 2: Performance Comparison of Feature Selection methods in Linear Regression Model

| Criteria<br>Method | R2 | MAE | RMSE | AIC |
|---|---|---|---|---|
| Base linear Regression | 0.7498672 | 3.192273 | 4.595225 | 3023.284 |
| LASSO | 0.7498666 | 3.192103 | 4.595231 | 3019.430 |
| XGB | 0.6328907 | 3.653001 | 5.566975 | 3203.420 |

The data in Table 2 explains the subtle differences in model performance across the different feature selection methods. While the LASSO-based model demonstrates a marginally improved AIC value compared to the baseline, indicating a slightly better model fit with fewer predictors, the XGB-based model exhibits a lower $R^2$ and higher error metrics (MAE and RMSE), suggesting a compromise in predictive accuracy when compared to the baseline linear

regression model. This comparative analysis not only highlights the strengths and limitations of each feature selection method but also underscores the importance of carefully considering the choice of methodology in the context of specific real-world datasets.

# CHAPTER 5
# CONCLUSION

In this study, we have extensively evaluated and compared the performance of advanced sequential knockoff methods, specifically, those employing Lasso and XGBoost, against a baseline sequential knockoff model in Data with Categorical Features. Our investigation was structured around three distinct simulation scenarios and an analysis of real-world data, focusing primarily on FDR and the power to assess model performance.

In the first scenario, our methods were juxtaposed with the Model-X and basic sequential knockoff models in an environment where features were exclusively numerical. This foundational comparison established a benchmark for subsequent evaluations. The second scenario introduced a mix of numerical and a limited number of binary features, further challenging the models under study. The third scenario expanded the complexity by including numerical and multiclass features, comparing the proposed methods against their derandomized counterparts. Across these scenarios, we scrutinized model performance under varying conditions of sparsity and sample size, thoroughly examining each method's robustness and adaptability.

The real-world applicability of our methods was demonstrated through the analysis of the Boston Housing Price dataset, characterized by its diverse

feature types. This practical evaluation underscored the relevance of our findings beyond theoretical simulations.

The results of our comprehensive analysis consistently highlight the superior performance of the Lasso and XGBoost-based sequential knockoff methods compared to the baseline model. Both in simulation scenarios and real data application, our methods demonstrated an enhanced ability to control the FDR while maintaining at least the same power, indicating a significant improvement in identifying true feature associations without inflating false discoveries.

This study's findings not only verify the efficacy of incorporating Lasso and XGBoost into sequential knockoff frameworks but also pave the way for further innovations in statistical methodology for feature selection, particularly in complex datasets. The advancements presented here contribute to the broader field of statistical learning, offering robust tools for researchers and practitioners aiming to extract meaningful insights from high-dimensional data.

# BIBLIOGRAPHY

[1] Barber, R.F. and Candès, E.J., “Controlling the false discovery rate via knockoffs”, 2015.

[2] Candes, E., Fan, Y., Janson, L. and Lv, J.,. “Panning for gold:‘model-X’knockoffs for high dimensional controlled variable selection,” Journal of the Royal Statistical Society Series B: Statistical Methodology, 80(3), pp.551-577, 2018.

[3] Srinivasan, A., Xue, L. and Zhan, X. “Compositional knockoff filter for high-dimensional regression analysis of microbiome data. Biometrics,” 77(3), pp.984-995, 2021.

[4] Ren, Z., Wei, Y. and Candès, E., “Derandomizing knockoffs. Journal of the American Statistical Association,” 118(542), pp.948-958, 2023.

[5] Ren, Z. and Candès, E., “Knockoffs with side information. The Annals of Applied Statistics,” 17(2), pp.1152-1174, 2023.

[6] Imrie, F., Norcliffe, A., Liò, P. and van der Schaar, M., “Composite Feature Selection Using Deep Ensembles. Advances in Neural Information Processing Systems,” 35, pp.36142-36160, 2022.

[7] Zhao, X., Li, W., Chen, H., Wang, Y., Chen, Y. and John, V., “Distribution-dependent feature selection for deep neural networks,” Applied Intelligence, pp.1-11, 2022.

[8] Barber, R.F. and Candès, E.J., “A knockoff filter for high-dimensional selective inference,” 2019.

[9] Orekondy, T., Schiele, B. and Fritz, M., “Knockoff nets: Stealing functionality of black-box models,” In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4954-4963), 2019.

[10] Fan, Z., Kernan, K.F., Sriram, A., Benos, P.V., Canna, S.W., Carcillo, J.A., Kim, S. and Park, H.J., “Deep neural networks with knockoff features identify nonlinear causal relations and estimate effect sizes in complex biological systems,” 2023.

[11] Liu, W., Ke, Y., Liu, J. and Li, R., “Model-free feature screening and FDR control with knockoff features,” Journal of the American Statistical Association, 117(537), pp.428-443, 2022.

[12] Romano, Y., Sesia, M. and Candès, E., “Deep knockoffs,” Journal of the American Statistical Association, 115(532), pp.1861-1872, 2020.

[13] Zhu, Guangyu, and Tingting Zhao. "Deep-gknock: Nonlinear group-feature selection with deep neural networks," Neural Networks 135 (2021): 139-147.

[14] Kormaksson, Matthias, Luke J. Kelly, Xuan Zhu, Sibylle Haemmerle, Luminita Pricop, and David Ohlssen. "Sequential knockoffs for continuous and categorical predictors: With application to a large psoriatic arthritis clinical trial pool," Statistics in Medicine 40, no. 14 (2021): 3313-3328.

[15] J. Jordon, J. James, and J. Yoon, M. van der Schaar, "KnockoffGAN: Generating knockoffs for feature selection using generative adversarial networks," in Proceedings of the International Conference on Learning Representations, 2018.

[16] H. Shen, Y. Yan, and Z. Zhao, "DeepDRK: Deep Dependency Regularized Knockoff for Feature Selection," arXiv preprint arXiv:2402.17176, 2024.

[17] X. Wang, D. Wang, W. Ying, R. Xie, H. Chen, and Y. Fu, "Knockoff-Guided Feature Selection via A Single Pre-trained Reinforced Agent," arXiv preprint arXiv:2403.04015, 2024.

[18] X. Zhao, H. Chen, Y. Wang, W. Li, T. Gong, Y. Wang, and F. Zheng, "Error-based knockoffs inference for controlled feature selection," in Proc. of the AAAI Conference on Artificial Intelligence, vol. 36, no. 8, pp. 9190-9198, 2022.

[19] Z. Ren and R. Foygel Barber, "Derandomised knockoffs: leveraging e-values for false discovery rate control," in Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 86, no. 1, pp. 122-154, 2024.

[20] J. Zhou and G. Claeskens, "A tradeoff between false discovery and true positive proportions for sparse high-dimensional logistic regression," Electronic Journal of Statistics, vol. 18, no. 1, pp. 395-428, 2024.

[21] Y. Huang, S. Pirenne, S. Panigrahi, and G. Claeskens, "Selective inference using randomized group lasso estimators for general models," arXiv preprint arXiv:2306.13829, 2023.