MACHINE LEARNING AIDED FEATURE SELECTION FOR ULTRAHIGH DIMENSIONAL DATA: THEORY AND APPLICATIONS

By

Arkaprabha Ganguli

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Statistics—Doctor of Philosophy

2023

ABSTRACT

Feature selection methods for ultra-high dimensional datasets have gained significant popularity in the field of statistical machine learning due to their wide applicability across various scientific domains. These methods aim to uncover the true sparsity pattern by identifying a small subset of features that are truly associated with the response variable. However, traditional feature selection algorithms may suffer from high false discovery rates, limiting their ability to provide meaningful insights into the underlying relationships. To address this issue, this thesis focuses on the development and study of two novel feature selection methods that incorporate False Discovery Rate (FDR) control. These methods are specifically applied to real-world diffusion magnetic resonance imaging (DMRI) tractography data, demonstrating their effectiveness in addressing several challenging issues in ultrahigh dimensional datasets.

In the first chapter, we propose a p-value-free FDR controlling method for feature selection. Most of the state-of-the-art methods in the literature for controlling FDR rely on p-value, which depends on specific assumptions on the data distribution and may be questionable in some high-dimensional settings. To surpass this problem, we propose a 'screening & cleaning' strategy consisting of assigning importance scores to the predictors, followed by constructing an estimate of the FDR. We study the theoretical properties of the method and demonstrate its superior performance compared to existing methods in an extensive simulation study. Finally, we apply the method to a gene expression dataset and identify important genes associated with drug sensitivity.

In the second chapter, We extend the feature selection method from a linear model to a non-linear and non-parametric setting by utilizing the Deep Learning (DL) framework. The DL has been at the center of analytics in recent years due to its impressive empirical success in analyzing complex data objects. Despite this success, most existing tools behave like black-box machines, thus the increasing interest in interpretable, reliable, and robust deep learning models applicable to a broad class of applications. Feature-selected deep learning has emerged as a promising tool in this realm. However, the recent developments do not accommodate ultra-high dimensional and highly correlated features or high noise levels. In this article, we propose a novel screening and cleaning method with the aid of deep learning for a data-adaptive multi-resolutional discovery of highly correlated predictors with a controlled FDR. Extensive empirical evaluations over a wide range of simulated scenarios and several real datasets demonstrate the effectiveness of the proposed method in achieving high power while keeping the false discovery rate at a minimum.

In the third and final chapter, we apply the proposed feature selection methods to the brain imaging tractography dataset. Our motivation comes from the evidence from studies of dementia which shows that some older adults continue to maintain their cognitive abilities despite signs of ongoing neuropathological diseases. Commonly referred to as cognitive reserve, this phenomenon has unclear neurobiological substrates and a current understanding of corresponding markers is lacking. This study aims at investigating the immense system of structural connections between brain regions constituting subcortical white matter (WM) as potential markers of cognitive reserve. Diffusion MRI tractography is an established computational neuroimaging method to model WM fiber organization throughout the brain. Standard statistical analyses capable of leveraging the high dimensionality of tractography data face additional methodological complications beyond those encountered in typical feature selection problems. Our proposed methodology is specifically tailored for addressing these concerns. Extensive simulation studies on synthetic datasets mimicking the real tractography dataset demonstrate a substantial gain in power with minimal false discoveries, compared with state-of-the-art methods for feature selection. Our application to predicting cognitive reserve in a clinical aging neuroimaging tractography dataset produces anatomically meaningful discoveries in brain regions associated with risk and resilience to neurodegeneration.

Overall, this thesis presents novel and effective methods for feature selection in ultrahigh dimensional settings. Our proposed framework would benefit the researchers and professionals who encounter the difficulty of choosing pertinent variables from correlated and vast datasets in diverse fields, ranging from finance and social sciences to biology.

I dedicate this thesis to my parents, Rita Ganguli and Malay Ganguli, whose unwavering love, support, and encouragement have been the foundation of my academic journey.

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my advisors, Dr. Tapabrata Maiti and Dr. David Todem, for their invaluable guidance, continuous support, and encouragement throughout my graduate studies. Their expertise, wisdom, and generosity have been instrumental in shaping my research and shaping me as a statistician.

I am also grateful to my thesis committee members, Dr. Haolei Weng, Dr. Lyudmila Sakhanenko, and Dr. Andrew Bender, for their time, feedback, and valuable insights. Their immense help is a critical part of the theory and application part of this thesis. Their constructive criticism and thoughtful suggestions have greatly enriched my research and improved the quality of my thesis.

I am grateful to the faculty and staff in the Department of Statistics and Probability at Michigan State University for providing a stimulating academic environment throughout my studies. The coursework I undertook in the department expanded my knowledge and equipped me with valuable skills for my future endeavors. I will always cherish the interactions with fellow students and faculty, and I am thankful for the unwavering support from the staff.

I would like to extend my deepest gratitude to my family, friends, and loved ones for their constant support, love, and inspiration that has been instrumental throughout my academic journey. I am especially grateful to my wife, Sikta, for her presence, solace, and joy that she has brought into my life. I am profoundly grateful to my parents and sister for their unwavering support, unconditional love, and countless sacrifices they have made throughout my academic journey. They have been my pillars of strength. I am also thankful to my friends at MSU, including Anirban, Sampriti, Mookyong, Tathagata, Sumegha, Hema, and many others, for their companionship and support. Their presence has made my journey even more memorable. I extend my heartfelt wishes to all of you for a bright and prosperous future filled with happiness and success.

To all of you who have been an integral part of my journey, I am deeply thankful for making it a truly remarkable and fulfilling experience.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

High-dimensional data analysis has become increasingly popular in various fields, such as biology, finance, social sciences, and engineering. In these fields, it is common to have datasets with a large number of features or variables, but a relatively small sample size. The main challenge in these situations is to identify the relevant features that are associated with the response variable while discarding the irrelevant ones, which is called variable selection.

1.1 Mathematical formulation

Within the framework of supervised learning, we denote a continuous response variable Y and a set of p continuous covariates $X = (X_1, \ldots, X_p)$. The cumulative distribution function (CDF) of the response variable Y is denoted by $F_y(\cdot)$, while the CDF of the k^{th} predictor X_k is denoted by $F_k(\cdot)$. We consider an ultrahigh-dimensional setting with a sample size of n and $p = O(exp(n^\tau))$ where $\tau > 0$. Now, to induce the sparsity, we assume the existence of a subset $S_0 \subset 1, 2, \ldots, p$ where $|S_0| = O(1)$, such that conditional on features in S_0, the response Y is independent of features in S_0^c. In other words, S_0 can be defined as $k : f(y|X)$ depends on X_k, where $f(y|X)$ is the conditional density of y given X. For the rest of this thesis, we call the relevant features in S_0 important or nonnull features; and the irrelevant features in S_0^c as unimportant or null features. Our objective is to identify the sparsity structure by estimating S_0.

1.2 The basic model selection methods

Variable selection has a long history in statistics, and various methods have been proposed for this purpose. The basic Model selection approaches aim to select the best parsimonious model among a set of candidate models based on a criterion such as the Akaike Information Criterion (AIC) Bozdogan (1987) or Bayesian Information Criterion (BIC) Chen and Chen (2008). These basic model selection methods are computationally feasible and work well for low-dimensional features. However, for ultra-high dimensional feature space, they become computationally intractable. As a solution, one might use regularization methods, which impose a penalty on the model complexity to avoid overfitting. A detailed review of this literature can be found in Fan and Lv (2010).

Under the linear model framework, suppose, the response y is associated with X through a linear model:

$$y_i = \mu + \sum_{j=1}^{p} X_{ij}\beta_j + \epsilon_i, \forall i = 1,2,\ldots,n$$

where due to the sparsity in the model, $\exists$ a subset $S_0 \subset \{1,2,\ldots,p\}$ for which $\beta_j \neq 0, \forall j \in S_0$ and $\beta_j = 0, \forall j \notin S_0$. The most popular penalized regression methods such as Lasso Tibshirani (1996) and MCP Zhang (2010) minimize the following objective function:

$$\hat{\beta} \in_{\beta \in \mathscr{R}^p} \frac{1}{2n}||Y - X\beta||_2^2 + p_\lambda(\beta)$$

The first part of this objective function minimizes the Mean Square Error (MSE) and the second part imposes a penalty on the number of features in the model, thus increasing the model parsimony. For example, for the lasso, $p_\lambda(\beta) = ||\beta||_1$, for the MCP $p_\lambda(\beta) = \lambda \sum_{j=1}^{p} \left(|\beta_j| - \frac{\lambda}{a}\right)_+$, for SCAD, $p_\lambda(\beta) = \lambda \sum_{j=1}^{p} \begin{cases} |\beta_j| & \text{if } |\beta_j| \leq \lambda \\ \frac{(a\lambda - |\beta_j|)\mathbf{1}_{\{\lambda < |\beta_j| \leq a\lambda\}}}{(a-1)} & \text{if } \lambda < |\beta_j| \leq a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\beta_j| > a\lambda \end{cases}$

These sparsity-inducing regularized methods typically result in a set of features for which the estimated coefficients are non-zero; i.e. the selected set of relevant features is $\hat{S}_n = \{j \in \{1,2,\ldots,p\} \ni |\hat{\beta}_j| \neq 0\}$. To assess the performance of a feature selection method, one would

- maximize the Power= $E\left(\frac{|\hat{S}_n \cap S_0|}{|S_0|}\right)$, the expected proportion of relevant features that are correctly identified and
- minimizes the FDR=$E\left(\frac{|\hat{S}_n \cap S_0^c|}{|\hat{S}_n|}\right)$, the expected proportion of falsely identified features among all the identified features.

Similar to the scenario in multiple testing, a trade-off exists between power and FDR in high-dimensional variable selection. Among the various methods available, the model selection consistent algorithms are the most desirable as they asymptotically uncover the true sparsity pattern, and have an asymptotic power of 1 with FDR decreasing to zero. However, these algorithms are based on stringent assumptions on the design matrix, which are typically not satisfied by modern high-dimensional datasets.

For instance, Lasso is model selection consistent only under irrepresentable conditions, as described in Zhao and Yu (2006), which requires the noise variables to be weakly correlated with signal variables. As a result, Lasso becomes inconsistent for variable selection in most modern studies. To relax the assumptions, we can first control the associated error and then try to maximize the power given the controlled error. For example, Lasso can attain asymptotic power 1 under Restricted Eigenvalue (RE) and beta-min conditions. However, such an approach does not offer any control over the associated error. Next, we discuss these two approaches: Sure screening property and FDR control.

1.2.1 Feature screening methods with sure screening property

A feature selection method enjoys the sure screening property if its output $\hat{S}_n$ satisfies the condition:

$$P(S_0 \subset \hat{S}_n) \to 1 \text{ as } n \to \infty$$

This implies all the relevant features are retained in the selected set of features. Hence, the asymptotic power converges to one; however, due to the lack of error control, these methods may result in higher FDR. The sure screening property was first introduced by Fan and Lv (2008) in the context of variable selection for linear regression models. They proposed the sure independence screening (SIS) method, which selects a subset of features based on marginal correlation with the response variable. Under certain conditions, SIS is able to identify all relevant features with high probability. Consequently, it has been shown to have good empirical performance in many applications.

Relaxing the linearity assumption, several model-free feature screening methods have developed over the years. For example, Xue and Liang (2017) developed a screening procedure based on a two-step procedure: (1) transforming the data (Y, X) to a Gaussian distributed array by norparanormal transformation Liu et al. (2009), and then (2) performing pairwise marginal independence test for a bivariate gaussian distribution on the transformed variables (Y, X_j), for each $j = 1, 2, \ldots, p$. This method enjoys the sure screening property under mild regularity conditions on the dimension of the feature space and the minimum signal strength. However, due to the inherent structure of the testing procedure, this method only considers the continuous response variable, which is applicable for regression tasks only. In a broader context, the feature screening method proposed by Zhou et al. (2018), can be employed for classification tasks. We discuss this in detail in chapters 2 and 3 later in this thesis.

1.2.2 Feature screening methods with FDR control

In addition to variable selection, controlling the False Discovery Rate (FDR) has also become an important issue in high-dimensional data analysis. As mentioned above, the FDR is the proportion of false positives among all the rejected null hypotheses. FDR-controlled methods aim to restrict the FDR at a pre-specified level while allowing some false positives. FDR control has been extensively studied in the multiple testing framework, and various methods have been proposed for this purpose, starting from the famous Benjamini-Hochberg (BH) method Benjamini and Hochberg (1995). Despite being widely used across scientific domains, the BH procedure relies on stringent assumptions regarding p-values, which can be challenging to satisfy in real-world datasets. As a solution, many other methods have been proposed along this line, see Tansey et al. (2018); Xia et al. (2017); Li and Barber (2019); Lei and Fithian (2018) for a more detailed overview. However, generating interpretable p-values for nonlinear models on high-dimensional data remains an unresolved research problem.

To overcome this limitation, the knockoff framework was proposed by Candès et al. (2018). Essentially, this is a model-free variable selection algorithm with provable FDR control, assuming prior knowledge on the predictors' distribution is available. In the next section, we discuss in detail the knockoff-based approaches.

1.3 Model free methods - Knockoffs and Deep Learning based approaches

Knockoff-based methods have emerged as an alternative approach to address the challenges of variable selection in high-dimensional data. The Knockoff framework, introduced by Candès et al. (2018), constructs a set of knockoff variables to mimic the correlation structure of the original variables while setting the knockoff features unassociated with the response y. The key idea is to use these knockoff variables as a reference to estimate the false discovery rate (FDR) of the selected variables, allowing for a controlled selection of features. The Knockoff approach has been extended to various regression settings and has been shown to outperform many other popular variable selection methods, especially in settings with complex nonlinear relationships between the response and the features. One advantage of the Knockoff framework is that it does not rely on any specific distributional assumptions, and thus is applicable to a wide range of data types. However, to generate the knockoff variables, one needs to know or estimate the distribution of the features which might be a daunting task in practice, especially for ultra-high dimensional data. In some cases, it may be possible to have prior knowledge of the correlation pattern among the features. For example, in genetics studies, there is a common notion of linkage disequilibrium, which helps to specify the dependency pattern among the alleles at polymorphisms (Sesia et al., 2018). However, this information is typically unavailable in many other domains.

Recently, Barber et al. (2020) demonstrated that the knockoff framework can yield inflation in false discoveries, consistent with the error incurred in estimating the predictor's distribution. This problem is further exacerbated by highly correlated features. An empirical illustration is provided in Figure 1.1, demonstrating how the model-X knockoff (Candès et al., 2018) typically fails to control FDR under a simplistic setting with high multicollinearity.

In recent years, deep learning (DL) based methods have also gained popularity in high-dimensional data analysis due to their ability to automatically extract relevant features from raw data. As a consequence, DL-based flexible knockoff generating algorithms have been proposed (Liu and Zheng, 2019; Jordon et al., 2019; Romano et al., 2020); however, they are trained in a typical big-n-small-p setting, and it is unclear how they will perform when the sample size n is significantly smaller than the dimension of the covariates p, and the predictors are highly correlated. We discuss this issue in detail in Chapter 3.

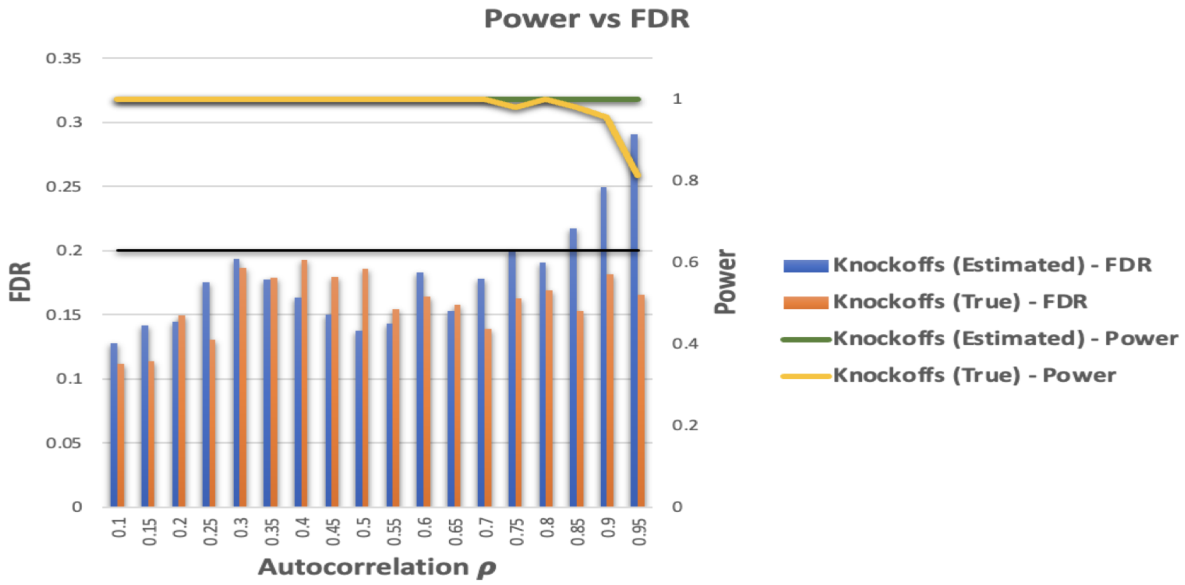

Figure 1.1 How multicollinearity affects Knockoffs - A demonstration using simplistic simulation setting: We simulate $n = 400$ iid copies of $(y \in \mathscr{R}, X \in \mathscr{R}^{100})$, where the outcome y is generated from a linear model: $y = 0.5(X_{20} + X_{40} + X_{60} + X_{80} + X_{100}) + \epsilon, \epsilon \sim N(0,20)$ and the features $X \sim N_{100}(0,\Sigma), (\Sigma)_{ij} = \rho^{|i-j|}$. We implement Model-X knockoff in two ways: (1) Model-X estimated: method proposed in Candès et al. (2018), generating knockoffs from the estimated distribution of the features from the data, and (2) Model-X True: generating the knockoff from the true distribution of X. For a higher autocorrelation ρ, which is a well-known difficult setting for traditional feature selection methods, the knockoffs-estimated loses its FDR control. This is because of the error in estimating the distribution of the features as knockoffs-true successfully maintains the power-FDR balance even for higher correlation.

On the other hand, Deep learning-based methods have gained significant attention in recent years for feature selection and prediction tasks in high-dimensional data. Deep learning models are designed to automatically learn feature representations from the data, without requiring explicit distributional assumptions. In the context of feature selection, deep learning models can be used to extract relevant features from high-dimensional data, which can then be used as inputs to downstream statistical models for prediction or classification tasks. Chen et al. (2021) proposed an L_0-norm based penalized neural network, called Deep Feature Selection (DFS). It enjoys exact model recovery under some mild assumptions on the underlying functional relationship between the response and the features. However, one challenge with deep learning-based feature selection is the lack of sufficient training data; which is quite common in many modern biological datasets. Recent efforts have been made to address this challenge, by developing methods to interpret the learned features and to relate them to known biological or physical processes. To address this issue, Liu et al. (2017) proposed an ensemble-based method utilizing random dropouts, especially for high dimensional low sample size data.

In this thesis, we propose an ensemble-based feature selection method for ultrahigh dimensional data with FDR control. Our method combines the strengths of regularization and model selection methods and achieves better performance than existing methods in terms of both variable selection and FDR control. We also provide a theoretical analysis of our method and show that it has desirable statistical properties.

BIBLIOGRAPHY

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

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

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370.

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(3):pp. 551–577.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Chen, Y., Gao, Q., Liang, F., and Wang, X. (2021). Nonlinear variable selection via deep neural networks. *Journal of Computational and Graphical Statistics*, 30(2):484–492.

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.

Jordon, J., Yoon, J., and van der Schaar, M. (2019). KnockoffGAN: Generating knockoffs for feature selection using generative adversarial networks. In *International Conference on Learning Representations*.

Lei, L. and Fithian, W. (2018). Adapt: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):649–679.

Li, A. and Barber, R. F. (2019). Multiple testing with the structure-adaptive benjamini–hochberg algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(1):45–74.

Liu, B., Wei, Y., Zhang, Y., and Yang, Q. (2017). Deep neural networks for high dimension, low sample size data. In *IJCAI*, pages 2287–2293.

Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(10).

Liu, Y. and Zheng, C. (2019). Deep latent variable models for generating knockoffs. *Stat*, 8(1):e260.

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

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

Tansey, W., Wang, Y., Blei, D., and Rabadan, R. (2018). Black box FDR. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4867–4876. PMLR.

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

Xia, F., Zhang, M. J., Zou, J. Y., and Tse, D. (2017). Neuralfdr: Learning discovery thresholds from hypothesis features. In *NIPS*.

Xue, J. and Liang, F. (2017). A robust model-free feature screening method for ultrahigh-dimensional data. *Journal of Computational and Graphical Statistics*, 26(4):803–813.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty.

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

CHAPTER 2

A P-VALUE-FREE FDR CONTROL METHOD FOR HIGH DIMENSIONAL VARIABLE SELECTION

2.1 Introduction

Variable selection is a fundamental problem in high-dimensional statistical analysis, where the goal is to select a subset of relevant variables from a large pool of potential predictors to build a parsimonious model. In modern applications in almost all scientific domains, such as genomics, brain imaging, and finance, the number of potential predictors can be much larger than the sample size, which makes traditional variable selection methods such as stepwise regression or Akaike's information criterion Bozdogan (1987) inefficient or impractical. Therefore, recent years have witnessed the development of various ultrahigh-dimensional variable selection methods that can handle a large number of variables relative to the sample size. These methods play a crucial role in improving the statistical accuracy and model interpretability while reducing the computational complexity at the same time. Penalized regression, which applies a penalty term to the likelihood function or objective function, is one of the most popular methods for variable selection in high-dimensional data analysis. Examples of these methods include Lasso and relevant algorithms Tibshirani (1996), SCAD Xie and Huang (2009), the elastic net Zou and Hastie (2003), adaptive Lasso Zou (2006) , and many more. Theoretical findings concerning parameter estimation, model selection, prediction, and oracle properties have been formulated across various model contexts. A comprehensive review of this literature is available in Fan and Lv (2010), and therefore, it is excluded from this discussion.

One of the main challenges in ultrahigh-dimensional variable selection is controlling the false discovery rate (FDR), which is the proportion of falsely selected variables among all selected variables. FDR control is essential for ensuring the validity and reproducibility of the results, especially in large-scale studies involving thousands or millions of variables. For example, in genome-wide association studies (GWAS), researchers need to consider hundreds of thousands of genetic markers to identify the variants associated with a particular trait or disease. Since the cost of false discoveries is high, as each selected variant requires a costly follow-up experiment, it is crucial to limit the number of false discoveries. Hence, researchers are interested in developing methods that can model the dependence structure of the data while ensuring an upper bound on the false discovery rate (FDR).

The False Discovery Proportion (FDP) can be represented as a random variable denoted by FDP, where

$$FDP = \frac{e_0}{N_+ \wedge 1}$$

Here, e_0 denotes the number of falsely selected variables, N_+ denotes the total discoveries, and $a \wedge b = max(a,b)$ tackles the situation where there is no discovery, i.e. $N_+ = 0$. The FDR is defined as $FDR = E(FDP)$. Estimating this expectation is a challenging task for the variable selection problem, and researchers have attempted to tackle this issue from multiple perspectives. Traditional FDR controlling methods, such as the Benjamini-Hochberg procedure Benjamini and Hochberg (1995), are based on p-values, which require an assumption of normality and are often not robust to non-normality or heavy-tailed distributions. Moreover, the validity of p-values depends on several assumptions such as independence or positive regression dependency on a subset (PRDS), which may be difficult to justify or examine in practice. Additionally, as noted in Candès et al. (2018), while maximum-likelihood theory can derive asymptotic p-values for low-dimensional generalized linear models (GLMs), it is unclear how to obtain p-values for high-dimensional models where the number of predictors exceeds the number of observations (n < p). Given these challenges, there is a need to develop a new FDR controlling method that does not rely on p-values and is suitable for both linear models (LMs) and GLMs.

As a p-value-free FDR controlling method, The Model-X knockoffs Candès et al. (2018) are widely used offering guaranteed control of the FDR and the flexibility to employ arbitrary predictive models. To create the distinction between the null and nonnull features, it generates the 'knockoff' features mimicking the dependence structure of the feature space while being independent of the response. To generate these knockoffs one needs to know or estimate the predictor's distribution, which is a daunting task in practice, especially for ultra-high dimensional feature space. However, even with knowledge of the underlying feature distribution, this method is infeasible unless the feature distribution is either a finite mixture of Gaussians Gimenez et al. (2019) or has a known Markov structure Bates et al. (2020). To address some of these complexities, Dai et al. (2022) proposed an FDR control method based on multiple data splitting (MDS) and combining different test statistics from multiple splits to construct the estimate of the FDR. Due to the name, we further call this method "MDS".MDS can theoretically control the FDR while avoiding the limitations of traditional approaches that rely on assumptions about the distribution of the data. It has been shown to perform well in simulations and real data analyses, and it represents a promising avenue for future research in the field of high-dimensional statistical inference. The method works well in general, however, is computationally intensive due to fitting a prediction model multiple splits of an ultra-high dimensional dataset. Also, from our experience, this method is too conservative to discover any features in many practical settings, thus reducing the power of the method.

To address these challenges, we present a novel p-value-free FDR controlling method for ultra-high dimensional datasets. The proposed method consists of two steps: screening and cleaning. In the screening step, we simply reduce the dimension of the feature space by eliminating some of the null features by utilizing a feature screening method with the sure screening property. Then in the cleaning step, we further clean out the null features from the selected set of screened features by constructing a p-value-free estimate of the FDR. We distinguish the null features from the set of relevant features by effectively utilizing the adaptive penalization on the repeatedly perturbed lasso. The idea of screening and cleaning is not new in statistics literature and was first used in Wasserman and Roeder (2009). Our method is easy to implement and does not require any tuning parameters, making it suitable for a wide range of applications. Under some mild regularity assumptions, we study the theoretical properties of the proposed method. For an empirical evaluation, we demonstrate that our method outperforms existing methods in terms of both FDR control and variable selection accuracy. We also provide insights into the behavior of our method under different scenarios and show that it can handle various types of data and noise structures. Additionally, we apply our method to a real-world dataset from gene expression analysis for several drugs where we demonstrate its practical usefulness and interpretability.

In summary, our proposed method provides a promising solution for controlled variable selection in ultrahigh dimensional problems. We believe that our method has significant practical applications in fields such as genomics, imaging, and natural language processing, where the number of features can be orders of magnitude larger than the number of samples. The remainder of this paper is organized as follows. In Section 2.2, we introduce our p-value-free FDR controlling method and describe its implementation details. Next, in Section 2.3, we study the asymptotic properties of the proposed method and showed its FDR control guarantee. In Section 2.4, we present simulation studies to evaluate the performance of our method and compare it with existing methods. In Section 2.5, we apply our method to two real-world datasets and demonstrate its practical usefulness. Finally, we conclude the paper with a discussion and future research directions in Section 2.6.

2.2 Methodology

2.2.1 Model setup and assumptions

In the context of a supervised learning regression framework, we have n independent and identically distributed (i.i.d.) copies of $(Y_i, X_i), i = 1, 2, \dots, n$, where Y is the continuous response variable and X, is the set of p continuous covariates, denoted by $X = (X^{(1)}, X^{(2)}, \dots, X^{(p)})$. We consider here the ultrahigh dimensional setting, allowing $p = p_n \to \infty$, as $n \to \infty$. For any square matrix C, let $\phi(C)$ and $\Phi(C)$ denote the smallest and largest eigenvalues of C. Also, if k is an integer, define $\phi_n(k) = \min_{M:|M|=k} \phi(\frac{1}{n} X'_M X_M)$ and $\Phi_n(k) = \max_{M:|M|=k} \Phi(\frac{1}{n} X'_M X_M)$. In the following sections, we will assume the following linear model and some basic key assumptions:

A1 $Y_i = X_i^{'} \beta^* + \epsilon_i$, where ϵ_i are independent and following $N(0, \sigma^2)$.

A2 The design matrix $X \in \mathbf{R}^{n \times p_n}$ allowing $p_n \to \infty$ as $n \to \infty$ with $p_n \leq c_1 e^{n^{c_2}}$, for some $c_1 > 0$ and $0 \leq c_2 < 1$.

A3 $S = \{j : \beta_j^* \neq 0\}$ with $s = |S| = s = O(1)$ and $\psi = min\{|\beta_j^*| : j \in S\} = \psi > 0$, $\{\beta_j^*, j \in S\}$ is assumed to be fixed.

A4 For any weight vector $w \in (0,1]^p$, there exists $\kappa > 0$ with

$$P(\liminf_{n\to\infty} min_{\Delta\in\mathscr{C}} \frac{\Delta'\left(\frac{X'X}{n}\right)\Delta}{||\Delta||_2^2} \geq \kappa > 0) = 1,$$

where, $\mathscr{C} = \{\Delta \in \mathscr{R}^p : ||\left(2W_{S^c} - I_{p-s}\right)\Delta_{S^c}||_1 \leq ||\left(2W_S + I_s\right)\Delta_S||_1\}$

A5 There exists positive constants c_2, c_3 and c_4 such that

$$P(\limsup_{n\to\infty} \Phi_n(n) \leq c_2) = 1, P(\liminf_{n\to\infty} \phi_n(c_3 \log n) \geq c_4) = 1,$$

and $P(\phi_n(n) > 0) = 1, \forall n$.

A6 We consider standardized covariates: $E(X_{ij}) = 0, E(X_{ij})^2 = 1$. Also, there exists a constant $B \in (0,\infty)$ such that $P(|X_{ij}| < B) = 1$.

These assumptions can be relaxed at the expense of more intricate proofs. Our objective is to learn the sparsity structure by estimating the true index set S through the selection of a feature set $\hat{D}_n$ that ensures control over the associated false discovery rate (FDR) under a predefined threshold q. Specifically, we aim for $FDR = E\left(\frac{|\hat{D}_n \cap S^c|}{|\hat{D}_n|}\right) < q$. Additionally, while maintaining FDR control, we strive to maximize the $Power = E\left(\frac{|\hat{D}_n \cap S|}{|S|}\right)$ in order to strike a stable trade-off between type-I and type-II errors, thereby enhancing the overall performance of the proposed methodology. Next, we will provide a detailed description of the multiple steps involved in our proposed methodology.

2.2.2 Screening step

Assuming that the cardinality of set S is much smaller than the dimension of the feature space, p, the majority of features are found in the complement of S_0, denoted as S_0^c. Hence, during the screening step, our primary focus is on identifying an active set, $\hat{S}_n$, with a much smaller cardinality than p, such that the probability of S being a subset of $\hat{S}_n$ approaches 1 as n approaches infinity. This is known as the sure screening property, first introduced in Fan and Lv (2008), which guarantees that all relevant predictors are retained in $\hat{S}_n$. The remaining predictors, denoted as $X_j, j \in \hat{S}_n^c$, are removed from further analysis. This is just a dimension reduction step. We proceed as follows:

- First, we randomly split the data into two groups $\mathbf{D}_1$ and $\mathbf{D}_2$, each with n_1 and n_2 observations. In $\mathbf{D}_1$, we fit the lasso model with tuning parameter $\lambda \in \Lambda$, where Λ is some index set, and get
$$\hat{\beta}(\lambda) = \text{argmin}_{\beta \in \mathbb{R}^{p_n}} \frac{1}{2n_1} \parallel Y_{\mathbf{D}_1} - X_{\mathbf{D}_1}\beta \parallel_2^2 + \lambda \parallel \beta \parallel_1 \tag{2.1}$$
We further denote the set of selected variables as a function of the tuning parameter λ: $\tilde{S}_n(\lambda) = \{j : |\hat{\beta}_j(\lambda)| \neq 0\}$.

- In $\mathbf{D}_2$, we minimize the squared error loss to get the optimum value of the tuning parameter λ^*, i.e., $\lambda^* = \text{argmin}_{\lambda \in \Lambda} \hat{L}(\lambda)$, where $\hat{L}(\lambda) = \frac{1}{n_2}\sum_{i \in \mathbf{D}_2}(Y_i - X_i^{'}\hat{\beta}(\lambda))^2$.

- Therefore our active set $\hat{S}_n$ is the set of variables corresponding to the cross-validated λ, i.e., $\hat{S}_n = \tilde{S}_n(\lambda^*)$, where $\lambda^* = \text{argmin}_{\lambda \in \Lambda} \hat{L}(\lambda)$

By assumption (A1)-(A6) in 2.2.1, it can be shown as in Wasserman and Roeder (2009) that this $\hat{S}_n$ enjoys sure screening property; i.e. $P(\hat{S}_n \supset S) \to 1$ as $n \to \infty$. Hence, although the dimension is reduced in the active set, we can further clean the active features eliminating the null features still retained in the active set $\hat{S}_n$. This is our goal in the next step.

2.2.3 Cleaning step

We start from the output y and the active features $X_{\hat{S}_n}$. In order to characterize the relevant features, we first define an idea of importance score for the active features. It measures the strength of the association of the predictor with the response. For a given weight vector $\{w_j, j \in \hat{S}_n\}$ and a long sequence of asymptotically increasing tuning parameter $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_2 r$ we fit weighted and unweighted lasso r times. For any tuning parameter $\lambda_i, i = 1, 2, \ldots, r$, the weighted lasso takes the following form:

$$\tilde{\beta}^w(\lambda_i) = \underset{\beta \in \mathbf{R}^{|\hat{\mathbf{S}}_\mathbf{n}|}}{\text{argmin}} \frac{1}{2n} \parallel Y - X_{\hat{S}_n}\beta_{\hat{S}_n} \parallel_2^2 + \lambda_i \parallel W\beta_{\hat{S}_n} \parallel_1 \tag{2.2}$$

On the other hand, the unweighted lasso is the standard lasso expressed as:

$$\tilde{\beta}^{uw}(\lambda_i) = \underset{\beta \in \mathbf{R}^{|\hat{\mathbf{S}}_\mathbf{n}|}}{\operatorname{argmin}} \frac{1}{2n} \parallel Y - X_{\hat{S}_n}\beta_{\hat{S}_n} \parallel_2^2 + \lambda_i \parallel \beta_{\hat{S}_n} \parallel_1 \tag{2.3}$$

where $W = diag(w_1, w_2, \ldots, w_{|\hat{S}_n|})$.

Then we calculate the importance score for the j-th predictor I_j^w by measuring how much it has survived through the lambda sequence before vanishing off; i.e.

$$\mathscr{I}_j^w = \frac{1}{r}\sum_{i=1}^{r} 1(\tilde{\beta}_j^w(\lambda_i) \geq \tau), \mathscr{I}_j^{uw} = \frac{1}{r}\sum_{i=1}^{r} 1(\tilde{\beta}_j^{uw}(\lambda_i) \geq \tau) \tag{2.4}$$

for some pre-specified small threshold τ. This is analogous to the thresholded lasso approach by Wang et al. (2017). The superscript 'w' and 'uw' indicates the usage of the weighted or unweighted version of lasso respectively. Next, we carefully select a sequence of tuning parameters, which is a crucial step in our proposed approach, to fully exploit the disparity between the weighted and unweighted lasso across a wide range of penalty levels. In order to achieve this, we consider the ordered sequence of tuning parameters $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_r \leq \lambda_{r+1} \leq \cdots \leq \lambda_{2r}$. To understand the behavior of lasso, we set the smallest r λ-values (i.e. $\lambda_1, \lambda_2, \ldots, \lambda_r$) in order of $\sqrt{\frac{log(p)}{n}}$. Now in order to understand the behavior of the weighted lasso, we set the largest r λ-values in a much higher level $o(\sqrt{n^\gamma})$ for $\gamma > 0$. Specifically, in practice, to select our sequence of tuning parameters λ to further calculate the importance score, we proceed as follows:

1. First, a sequence of $r-1$ values for $\lambda_1, \ldots, \lambda_{r-1}$ is generated in a logarithmic scale; such as $\lambda_i = 2\sqrt{\frac{log(p)}{n}} * \phi^{\frac{r-i}{r}}, \forall i = \{1, 2, \ldots, r-1\}$.

2. Then, we set the λ_r at $\lambda_r = \max_{j \in \{1,2,\ldots,p\}} |X_j' y|/n$.

3. Next, for the last and largest r values of the sequence, we start with setting $\lambda_{2r} = \max_{j \in \{1,2,\ldots,p\}} |X_j' y|/n + \tilde{c}n^\gamma$, where $\tilde{c}$ is a small constant, typically set as $\tilde{c} = 0.0001$ and $\gamma < \frac{1}{2}$.

4. Finally, we generate the remaining $r-1$ largest λ-values (i.e. $\lambda_{r+1}, \ldots, \lambda_{2r}$) using a similar logarithmic scale such as: $\lambda_{r+i} = \lambda_{2r} * \phi^{\frac{r-i}{r}}, \forall i = \{1, 2, \ldots, r-1\}$.

The value of ϕ is typically set at 0.001, and r is set to 100. Despite its apparent complexity, this construction of the sequence of penalty parameters directly emerges from the asymptotic analysis of the weighted lasso along its regularization path, as we have extensively discussed in Section 2.3. Using this λ-sequence, we bound the unweighted importance scores below 1; i.e. $\mathscr{I}_j^{uw} < 1, \forall j \in \{1, 2, \ldots, p\}$. The underlying idea is that the important predictors should survive longer even for the higher value of the tuning parameter λ. This importance score has great potential in identifying the association strength of the covariates with the response by utilizing adaptive penalization. Suppose, the importance score for the j-th variable obtained from the unweighted lasso (i.e. taking $w_j = 1, \forall j$) is denoted by $\mathscr{I}_j^{uw}$. Then for a typical simulated dataset, the following Figure 2.2.3 demonstrates the behavior of weighted and unweighted importance scores. We can particularly note that the weighted importance scores for true non-null variables are significantly greater than the whole bootstrap distribution of the unweighted importance scores. On the other hand, the weighted importance score for the true null variable is comparable to their unweighted counterparts. This disparity between the unweighted and weighted version of the importance score makes it apt for replacing the p-values in the FDR estimation in this cleaning step. we proceed as follows.

1. We want to select the variables with high-importance scores. So, for any cutoff Δ, we define, $N_+(\Delta)$ =number of total discovered variables = $\sum_{j \in \hat{S}_n} 1\left(\mathscr{I}_j^{w} \geq \Delta\right)$ and $e_0(\Delta)$ = number of falsely discovered variables = $\sum_{j \in \hat{S}_n} 1\left(\mathscr{I}_j^{w} \geq \Delta, j \not\in S\right)$

2. While the random variable $N_+(\Delta)$ is observable, we need to estimate somehow the unobserved $e_0(\Delta)$ by $\hat{e}_0(\Delta)$ and thus we estimate the False Positive Rate (FPR) as

$$F\hat{P}R(\Delta) = \frac{\hat{e}_0(\Delta)}{N_+(\Delta)} = \frac{\sum_{j \in \hat{S}_n} 1\left(\mathscr{I}_j^{uw} \geq \Delta\right)}{\sum_{j \in \hat{S}_n} 1\left(\mathscr{I}_j^{w} \geq \Delta\right)}$$

3. We calculate the optimum cutoff Δ^*, for which the $F\hat{P}R$ is controlled at some pre-specified threshold q:

$$\Delta^* = \min\left\{\Delta > 0 : 0 < \frac{\hat{e}_0(\Delta)}{N_+(\Delta)} \leq q\right\} \tag{2.5}$$

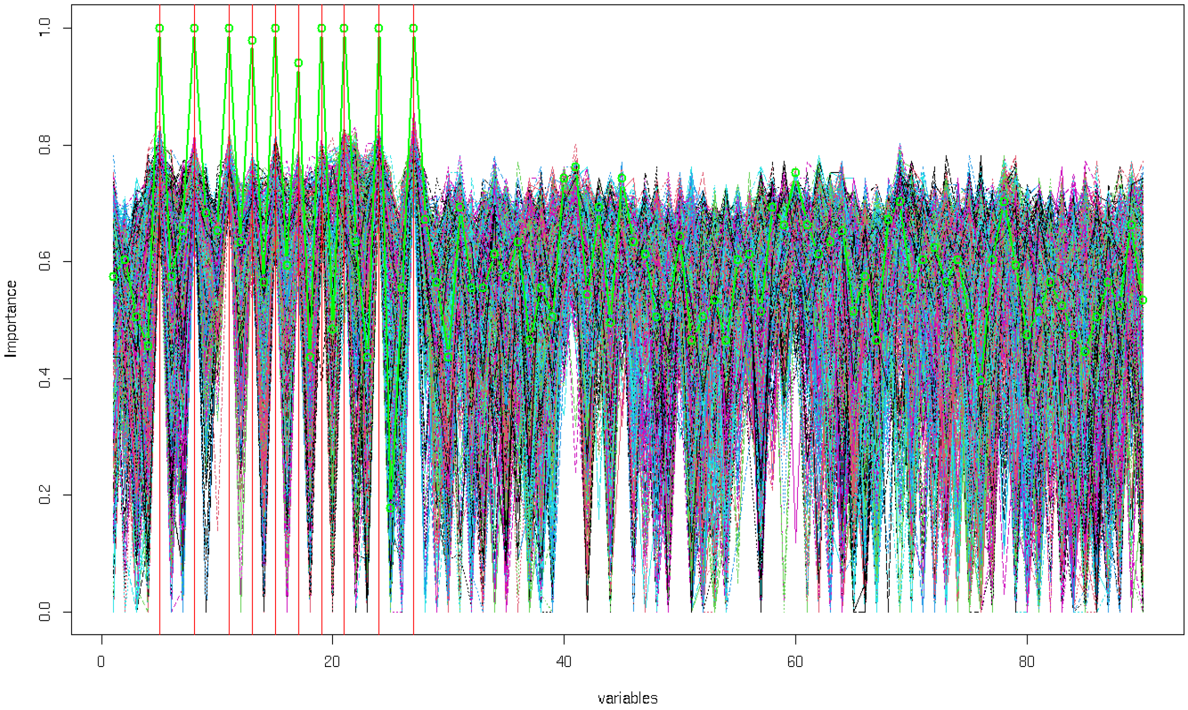

Figure 2.1 Important scores - weighted vs unweighted - A demonstration using simplistic simulation setting: We simulate $n = 400$ iid copies of $(y \in \mathscr{R}, X \in \mathscr{R}^{100})$, where the outcome y is generated from a linear model: $y = 0.5(X_{20}+X_{40}+X_{60}+X_{80}+X_{100})+\epsilon, \epsilon \sim N(0,1)$ and the features $X \sim N_{100}(0,\Sigma), (\Sigma)_{ij} = \rho^{|i-j|}$. The ten Vertical Red lines are indicating the true nonnull predictors. The Solid Green is indicating the weighted importance scores and the dense lines are indicating the unweighted importance scores for 1000 bootstrap versions of the data.

This minimum is taken over $\Delta > 0$ taking values in the set $\left\{\mathscr{I}_1^w, \mathscr{I}_2^w, \ldots, \mathscr{I}_p^w\right\}$

4. Finally the selected set of variables:

$$\hat{D}_n = \left\{j \in \hat{S}_n : \mathscr{I}_j^w \geq \Delta^*\right\} \tag{2.6}$$

Intuitively, the weighted and unweighted importance scores behave similarly for the null variables and the proposed method utilizes this characteristic to identify the null features. In the next section, we further show theoretically that with mild assumption, our proposed above-mentioned will control the FDR below the user-specified threshold q.

2.2.4 Obtaining appropriate weights

The proposed method relies on obtaining suitable weights, where smaller values are assigned to the relevant predictors in set S to minimize their penalty, while larger values are assigned to null features in set S^c to increase their penalization. To address this, we introduce a perturbation bootstrap approach in this section. Perturbation bootstrap is a resampling technique used to estimate the sampling variability and uncertainty associated with a statistical model or estimator by perturbing the observed data. It involves generating new datasets by introducing small perturbations or variations to the original data, allowing for the assessment of the stability and robustness of the statistical analysis. A more detailed review of the perturbation bootstrap can be found elsewhere, such as Minnier et al. (2011); Das and Lahiri (2019). In our setting, we define the perturbation bootstrap in the following way.

$$\hat{\beta}^b(\lambda) = \text{argmin}_{\beta\in\mathbb{R}^{|\hat{S}_n|}} \frac{1}{2n}\sum_{i=1}^{n} u_i(Y_i - X_{\hat{S}_n,i}\beta_{\hat{S}_n})^2 + \lambda \parallel \beta_{\hat{S}_n} \parallel_1 \tag{2.7}$$

where $\{u_1, u_2, \ldots, u_n\}$ is a set of positive values iid samples generated from a bounded distribution like $Uniform(1,2)$. These random samples stochastically perturb the objective function and repeating this a large number of times helps in recognizing patterns in the solution and assessing the underlying uncertainty.

To generate the weights $w_j, j \in \hat{S}_n$, , we set the tuning parameter maintaining $\lambda_n^w = o(\sqrt{\frac{log(p)}{n}})$. We randomly generate the iid samples $U_b^* = \{u_1^b, u_2^b, \ldots, u_n^b\}$ from a non-degenerate bounded positive-valued distribution and repeat the process B times to generate B sets of $\{U_b^*\}_{b=1}^B$. Let $\hat{\beta}^b(\lambda_n^w)$ be the perturbed lasso estimator 2.2.4 perturbed by U_b^* with tuning parameter λ_n^w. In Section 2.3, we demonstrate this perturbed lasso enjoys asymptoticaly vanishing L_2 error $||\hat{\beta}^w(\lambda) - \beta^*||_2$. Finally,we calculate the weight w_j for the variable $X_j, j \in \hat{S}_n$ as

$$w_j = \frac{1}{B}\sum_{b=1}^{B} 1\left(|\hat{\beta}_j^b(\lambda_n^w)| < \tau\right) + \frac{\tilde{c}}{n^{\frac{1}{2}-\gamma}} \tag{2.8}$$

for some pre-specified small threshold $\tilde{c}, \tau > 0$ and $\gamma < \frac{1}{2}$ as mentioned in Section 2.2.3. By selecting a suitably small value for τ, the consistency of the perturbed lasso method guarantees that the weights $w_j, j \in \hat{S}_n \cap S$ will approach zero suggesting no penalization for the true features. On the other hand, the null weights $w_j, j \in \hat{S}_n \cap S^c$ would tend to 1 indicating higher penalization. However, to maintain their rate of convergence we add the small decreasing sequence $\frac{\tilde{c}}{n^{\frac{1}{2}-\gamma}}$. We discuss and prove this in detail in Section 2.3

2.3 Theoretical properties

In this section, we establish the asymptotic properties of the proposed method, including the guarantee of FDR control. For ease of demonstration, we divide the theoretical study into the following parts: Section 2.3.1 focuses on the asymptotic properties of the weights, while Section 2.3.2 provides a detailed analysis of the weighted lasso. We examine the differences between the weighted and unweighted lasso methods for an asymptotically increasing sequence of tuning parameters, which leads to the attainment of FDR control. Additionally, our analysis demonstrates that the method achieves asymptotic power approaching unity. Moreover, in accordance with the assumptions stated in Section 2.2.1, our screening step directly utilizes the framework proposed in Wasserman and Roeder (2009), which has been proven to satisfy the sure screening property, denoted as $P(\hat{S}_n \supset S) \to 1$ as $n \to \infty$. Therefore, conditioned on this screening step, our focus shifts to the analysis of the FDR and power of our proposed method on the reduced dimensional data $\left(Y, X_{\hat{S}_n}\right)$. For the sake of notational simplicity, throughout this section, the term "X" refers specifically to the active features $X_{\hat{S}_n}$ obtained from the screening step unless stated otherwise. Consequently, $\tilde{p} = |\hat{S}_n|$ and $\beta = \beta_{\hat{S}_n}$

2.3.1 Asymptotic properties of the weights

We start from the perturbed lasso estimator defined in 2.2.4.

$$\begin{aligned} \hat{\beta}^b(\lambda) &\in \text{argmin}_{\beta \in \mathbb{R}^{\tilde{p}}} \frac{1}{2n} \sum_{i=1}^{n} u_i (Y_i - X_i \beta)^2 + \lambda \parallel \beta \parallel_1 \\ &\in \text{argmin}_{\beta \in \mathbb{R}^{\tilde{p}}} \frac{1}{2n} ||\tilde{U}_b Y - \tilde{U}_b X \beta||_2^2 + \lambda \parallel \beta \parallel_1 \end{aligned}$$

where $\tilde{U}_b = diag\{U_b^*\} = diag\{u_1^b, u_2^b, \ldots, u_n^b\}$.

Lemma 2.3.1 (L_2 error bound of perturbed Lasso)**.** *Under the assumptions mentioned in Section 2.2.1 and with bounded perturbation variables $U_b^* \in [c_L, c_U]^n$, the L_2 estimation error of perturbed lasso converges to zero; i.e.*

$$||\hat{\beta}^b - \beta||_2 \leq o\left(\sqrt{\frac{log(\tilde{p})}{n}}\right) \tag{2.9}$$

Proof. As $\hat{\beta}^b$ minimizes the objective function in 2.2.4, we note that

$$\begin{aligned}
&\frac{1}{2n} \parallel \tilde{U}Y - \tilde{U}X\hat{\beta}^b \parallel_2^2 + \lambda \parallel \hat{\beta}^b \parallel_1 \leq \frac{1}{2n} \parallel \tilde{U}Y - \tilde{U}X\beta^* \parallel_2^2 + \lambda \parallel \beta^* \parallel_1 \\
&\implies 0 \leq \frac{1}{2}(\hat{\beta}^b - \beta^*)'(\frac{X'\tilde{U}'\tilde{U}X}{n})(\hat{\beta}^b - \beta^*) \leq \frac{1}{n}\epsilon'\tilde{U}'\tilde{U}X(\hat{\beta}^b - \beta^*) + \lambda||\beta^*||_1 - \lambda||\hat{\beta}^b||_1 \\
&\leq \lambda \left[\frac{1}{2}||\hat{\beta}^b - \beta^*||_1 + ||\beta^*||_1 - ||\hat{\beta}^b||_1\right]
\end{aligned} \tag{2.10}$$

by taking $\lambda \geq 2||\frac{1}{n}\epsilon'\tilde{U}'\tilde{U}X||_\infty$. Now, we note that from equation 2.3.1,

$$\begin{aligned}
0 &\leq ||\hat{\beta}^b - \beta^*||_1 + 2||\beta^*||_1 - 2||\hat{\beta}||_1 \\
&= ||\hat{\beta}^b_S - \beta^*_S||_1 + ||\hat{\beta}^b_{S^c}||_1 + 2||\beta^*_S||_1 - 2||\hat{\beta}^b_S||_1 - 2||\hat{\beta}^b_{S^c}||_1 \\
&\leq \sum_{j\in S} |\hat{\beta}^b_j - \beta^*_j| + 2\sum_{j\in S} |\hat{\beta}^b_j - \beta^*_j| + \sum_{j\in S^c} |\hat{\beta}^b_j| - 2\sum_{j\in S^c} |\hat{\beta}^b_j| \\
&= ||3(\hat{\beta}^b_S - \beta^*_S)||_1 - ||\hat{\beta}^b_{S^c}||_1
\end{aligned} \tag{2.11}$$

Hence, $\hat{\beta}^b - \beta^* \in \mathscr{C}$, where the cone $\mathscr{C}$ is defined in equation A4. Additionally, we note that, as $u^b_i \geq c_L > 0, \forall i \in \{1, 2, \ldots, n\}$, for any $\Delta \in \mathscr{C}$,

$$\Delta'\frac{X'\tilde{U}'\tilde{U}X}{n}\Delta = \frac{1}{n}||\tilde{U}X\Delta||_2^2 \geq \frac{1}{n}\min_{i\in\{1,2,\ldots,n\}}(u^b_i)^2||X\Delta||_2^2 \geq c_L\kappa \tag{2.12}$$

Hence the RE assumption in A4 holds for perturbed Lasso with the new lower bound $c_L\kappa$. Hence, continuing from equation 2.3.1,

$$\begin{aligned}
\frac{c_L\kappa}{2}||\hat{\beta}^b - \beta^*||_2^2 &\leq \frac{\lambda}{2}\left[||3(\hat{\beta}^b_S - \beta^*_S)||_1 - ||\hat{\beta}^b_{S^c}||_1\right] \\
&\leq \frac{\lambda}{2}\left[||3(\hat{\beta}^b_S - \beta^*_S)||_1\right] \\
&\leq \frac{\lambda}{2}\left[3\sqrt{s}||\hat{\beta}^b_S - \beta^*_S)||_2\right] \\
&\leq \frac{\lambda}{2}\left[3\sqrt{s}||\hat{\beta}^b - \beta^*)||_2\right]
\end{aligned} \tag{2.13}$$

Hence, $||\hat{\beta}^b - \beta^*||_2 \leq \frac{3\lambda\sqrt{s}}{c_L\kappa}$, when $\lambda \geq 2||\frac{1}{n}\epsilon'\tilde{U}'\tilde{U}X||_\infty$. Now we note that as by assumption A6,

$$P(|X_{ij}| < B) = 1, \implies \frac{1}{n}||\tilde{U}X_j||_2^2 \leq c_U^2 B^2 \tag{2.14}$$

Now, let $Z_j = \frac{1}{n}\epsilon'\tilde{U}X_j \sim N\left(0, \frac{\sigma^2}{n}||\tilde{U}X_j||_2^2\right), j = 1, 2, \ldots, \tilde{p}$, conditional on $\tilde{U}$ and X. Then $||\frac{1}{n}\epsilon'\tilde{U}X||_\infty = \max_{1\leq j\leq\tilde{p}}|Z_j|$. This implies,

$$\begin{aligned} P\left(\frac{2}{n}||\epsilon'\tilde{U}X||_\infty \geq t\right) \leq \sum_{j=1}^{\tilde{p}} P\left(|Z_j| > \frac{t}{2}\right) &\leq \sum_{j=1}^{\tilde{p}} 2e^{-\frac{t^2n^2}{8||\tilde{U}X_j||_2^2\sigma^2}} \\ &\leq 2\tilde{p}e^{-\frac{t^2n^2}{8\sigma^2c_U^2B^2n}} \\ &= 2\tilde{p}e^{-\frac{t^2n}{8\sigma^2c_U^2B^2}} \\ &= 2e^{-\frac{n\delta^2}{8}} \to 0 \end{aligned} \tag{2.15}$$

by setting $t = c_U B\sigma\left(\sqrt{\frac{8\log(\tilde{p})}{n}} + \delta\right)$

Extending this result, we can show that with probability tending to 1,

$$||\hat{\beta}^b - \beta^*||_2 \leq \frac{3\lambda\sqrt{s}}{c_L\kappa} \text{ if, } \lambda \geq o\left(\sqrt{\frac{log(\tilde{p})}{n}}\right) \tag{2.16}$$

□

Lemma 2.3.1 establishes that with $\lambda = o\left(\sqrt{\frac{log(\tilde{p})}{n}}\right)$, the L_2 error bound $||\hat{\beta}^b - \beta^*||_2 \leq o_P(1)$.Next, we utilize this result to show that our proposed weights mentioned in Section 2.2.4 achieve the desired properties.

First, we observe that by the error bound in Lemma 2.3.1, for any $\mathbf{j} \in \mathbf{S}$, we can write, $|\beta_j^*| = |\beta_j^* - \hat{\beta}_j^b + \hat{\beta}_j^b| \leq |\hat{\beta}_j^b| + |\hat{\beta}_j^b - \beta_j^*|$ and that implies, $|\hat{\beta}_j^b| \geq |\beta_j^*| - |\hat{\beta}_j^b - \beta_j^*| \geq \psi + o_P(1)$. Consequently, for any $\mathbf{j} \in \mathbf{S}$, $|\hat{\beta}_j^b| \leq o_P(1)$. Hence, with sufficiently small $\tau < \psi$ and $\lambda_n^w = o(\sqrt{\frac{log(p)}{n}})$, $1\left(|\hat{\beta}_j^b(\lambda_n^w)| < \tau\right) \xrightarrow{P} 0$. Similarly, for $\mathbf{j} \in \mathbf{S^c}$, $1\left(|\hat{\beta}_j^b(\lambda_n^w)| < \tau\right) \xrightarrow{P} 1$, as $n \to \infty$.

For $0 < \gamma < \frac{1}{2}$, we recall,

$$w_j = \frac{1}{B}\sum_{b=1}^{B} 1\left(|\hat{\beta}_j^b(\lambda_n^w)| < \tau\right) + \frac{\tilde{c}}{n^{\frac{1}{2}-\gamma}} \tag{2.17}$$

Lets denote, $\tilde{w}_j = \frac{1}{B}\sum_{b=1}^{B} 1\left(|\hat{\beta}_j^b(\lambda_n^w)| < \tau\right)$. by Markov inequality, we see that, for any $\alpha > 0$ and

$j \in S$,

$$P(\tilde{w}_j > 2e^{-\frac{n(\delta^2+\alpha)}{8}}) < \frac{E(\tilde{w}_j)}{2e^{-\frac{n(\delta^2+\alpha)}{8}}} = \frac{P\left(|\hat{\beta}_j^b(\lambda_n^w)| < \tau\right)}{2e^{-\frac{n(\delta^2+\alpha)}{8}}}$$
$$\leq \frac{2e^{-\frac{n\delta^2}{8}}}{2e^{-\frac{n(\delta^2+\alpha)}{8}}}$$

for $\delta > 0$. Hence, $w_j = o\left(e^{-\frac{n(\delta^2+\alpha)}{8}} + \frac{1}{n^{\frac{1}{2}-\gamma}}\right)$ for $j \in S$.

For $j \not\in S$, for any $\alpha > 0$, similarly as above $w_j \to 1$ as $n \to \infty$

$$\begin{aligned} P\left(\tilde{w}_j < 1 - 2e^{-\frac{n(\delta^2+\alpha)}{8}}\right) = P\left(1 - \tilde{w}_j > 2e^{-\frac{n(\delta^2+\alpha)}{8}}\right) &\leq \frac{E(1-\tilde{w}_j)}{2e^{-\frac{n(\delta^2+\alpha)}{8}}} \\ &\leq \frac{2e^{-\frac{n\delta^2}{8}}}{2e^{-\frac{n(\delta^2+\alpha)}{8}}} \text{ as } n \to \infty \end{aligned} \tag{2.18}$$

Hence, $|1 - w_j| < o\left(e^{-\frac{n(\delta^2+\alpha)}{8}} + \frac{1}{n^{\frac{1}{2}-\gamma}}\right)$ for $j \not\in S$. Consequently, the proposed weights offer valuable proxy information regarding β^*. They tend to approach zero for relevant features in S, indicating their amplified significance, while converging towards 1 for null features in S^c. This capability to discern between null and non-null features renders them suitable for utilization in adaptive penalization strategies.

2.3.2 Asymptotic FDR control and power analysis

We start by expressing the FPR in the following way:

$$\begin{aligned} FPR = \frac{\sum_{j=1}^p 1\left(\mathscr{I}_j^w \geq \Delta^*, j \not\in S\right)}{\sum_{j=1}^p 1\left(\mathscr{I}_j^w \geq \Delta^*\right)} = \frac{\sum_{j\not\in S} 1\left(\mathscr{I}_j^w \geq \Delta^*\right)}{\sum_{j=1}^p 1\left(\mathscr{I}_j^w \geq \Delta^*\right)} &= \frac{\sum_{j=1}^p 1\left(\mathscr{I}_j^{uw} \geq \Delta^*\right)}{\sum_{j=1}^p 1\left(\mathscr{I}_j^w \geq \Delta^*\right)} \cdot \frac{\sum_{j\not\in S} 1\left(\mathscr{I}_j^w \geq \Delta^*\right)}{\sum_{j=1}^p 1\left(\mathscr{I}_j^{uw} \geq \Delta^*\right)} \\ &\leq q \cdot R(\Delta^*) \end{aligned} \tag{2.19}$$

where $R(\Delta) = \frac{\sum_{j\not\in S} 1\left(\mathscr{I}_j^w \geq \Delta\right)}{\sum_{j=1}^p 1\left(\mathscr{I}_j^{uw} \geq \Delta\right)}$. The last inequality holds by the construction of Δ^* in equation 3.

Now in order to show the FDR control, we only need to show that,

$$\lim_{n\to\infty} E\left(R(\Delta^*)\right) \leq 1$$

Now, decomposing the $E\left(R(\Delta^*)\right)$ we get

$$E\left(R(\Delta^*)\right) = E\left(\frac{\sum_{j\not\in S} 1\left(\mathscr{I}_j^w \geq \Delta^*\right)}{\sum_{j\in S} 1\left(\mathscr{I}_j^{uw} \geq \Delta^*\right) + \sum_{j\not\in S} 1\left(\mathscr{I}_j^{uw} \geq \Delta^*\right)}\right)$$

Following Theorem 2.3.2 first establishes the L_2 error bound for the weighted lasso in equation 2.2.3. The proof is relegated to Appendix B.

Theorem 2.3.2 (L_2 error bound of weighted Lasso)**.** *Assume that the weights corresponding to the true features are among the first m elements of the ordered list of weights defined as* $w_{order} = \{w_{(1)}, w_{(2)}, \ldots, w_{(\tilde{p})}\}$ *and let's denote the set of m ordered weights as* $w_T = \{w_{(j)}, 1 \le j \le m\}$. *Under the basic assumptions mentioned in Section 2.2.1 and with positive constants c and c′, the weighted lasso maintains the following L_2 error bound*

$$||\hat{\beta}^w(\lambda) - \beta^*||_2 \le \frac{\lambda}{\kappa}||w_T||_2 \text{ , } where\ \lambda \ge c\sigma\left(\frac{\sqrt{m}}{\sqrt{n}||w_T||_2} + \sqrt{\frac{log(\tilde{p})}{n}}\frac{1}{w_{(m+1)}}\right) \qquad (2.20)$$

when $\frac{||w_T||_2}{w_{(m+1)}}\sqrt{\frac{log(\tilde{p})}{n}} + \sqrt{\frac{m}{n}} = o(1)$. *This further implies* $||\hat{\beta}^w - \beta^*||_\infty \le o_P(1)$

Theorem 2.3.2 demonstrates how the proper utilization of weights in adaptive penalization improves the accuracy. We can retrieve the traditional L_2 bound for unweighted lasso by setting $w_j = 1, \forall 1 \le j \le \tilde{p}$. Next, in the following lemma, we demonstrate some key characteristics for our proposed weighting scheme in Section 2.2.4.

Lemma 2.3.3 (Key characteristics for our proposed weighting scheme in Section 2.2.4)**.** *Suppose the weights are generated using our proposed weighting scheme in Section 2.2.4. Then, with high probability, (1) The* $w_S = \{w_j, j \in S\}$ *belongs to to the first s elements in the ordered list* w_{order}*; i.e.,*$P(|m - s| > \epsilon) \to 0$*, and (2)* $P(\tilde{w}_{min} < 1 - \frac{1}{n^{\frac{1}{2}+\gamma}}) \to 0$

Proof. We prove these statements one by one.

Claim 1: The w_S belongs to to the first s elements in the ordered list w_{order}; i.e.,$P(|m - s| > \epsilon) \to 0$

Proof of Claim 1: This is true as by lemma 2.3.1, for any $\alpha > 0$

$$P(\sup_{j\in S} w_j > e^{-\frac{n(\delta^2+\alpha)}{8}} + \frac{1}{n^{\frac{1}{2}-\gamma}}) \le \sum_{j\in S} P(w_j > e^{-\frac{n(\delta^2+\alpha)}{8}} + \frac{1}{n^{\frac{1}{2}-\gamma}}) \le 2se^{-\frac{n\alpha}{8}} \text{ and}$$

$$P(\inf_{j\in S^c} w_j < (1 - e^{-\frac{n(\delta^2+\alpha)}{8}} + \frac{1}{n^{\frac{1}{2}-\gamma}})) \le \sum_{j\in S^c} P(w_j < 1 - e^{-\frac{n(\delta^2+\alpha)}{8}} + \frac{1}{n^{\frac{1}{2}-\gamma}}) \le 2\tilde{p}e^{-\frac{n\alpha}{8}}$$

After the screening, the dimension of the active set $|\hat{S}_n| = \tilde{p} = o(n^3)$ Wasserman and Roeder (2009) and by our assumption in Section 2.2.1, $s = O(1)$; hence, both the probabilities tend to zero with a high convergence rate. The weight w_S corresponds to the first s elements in the ordered list w_{order}, denoted as $w_{(1)}, w_{(2)}, \dots, w_{(\tilde{p})}$. Consequently, for any $\epsilon > 0$, $P(|m-s| > \epsilon) \leq P(\sup_{j\in S} w_j > e^{-\frac{n(\delta^2+\alpha)}{8}} + \frac{1}{n^{\frac{1}{2}-\gamma}}) + P(\inf_{j\in S^c} w_j < (1-e^{-\frac{n(\delta^2+\alpha)}{8}} + \frac{1}{n^{\frac{1}{2}-\gamma}})) \leq 4\tilde{p}e^{-\frac{n\alpha}{8}} \to 0$.

Claim 2: Define $\tilde{w}_{min} = \min_{j=m+1,\dots,\tilde{p}} w_{(j)}$. Then, $P(\tilde{w}_{min} < 1-\frac{1}{n^{\frac{1}{2}+\gamma}}) \to 0$ and $P(||w_{(1:m)}||_2 > \frac{1}{n^{\frac{1}{2}-\gamma}}) \to 0$.

Proof of Claim 2: The proof directly follows from the proof of claim 1. First, we note that

$$\begin{aligned}
P(\tilde{w}_{(m+1)} < 1-e^{-\frac{n(\delta^2+\alpha)}{8}} + \frac{1}{n^{\frac{1}{2}-\gamma}}) &= P(w_{(s+1)} < 1-e^{-\frac{n(\delta^2+\alpha)}{8}} + \frac{1}{n^{\frac{1}{2}-\gamma}}) + P(m > s) \\
&\leq P(\inf_{j\in S^c} w_j < (1-e^{-\frac{n(\delta^2+\alpha)}{8}} + \frac{1}{n^{\frac{1}{2}-\gamma}})) + P(m > s) \\
&\leq 6\tilde{p}e^{-\frac{n\alpha}{8}}
\end{aligned}$$

□

Therefore, these weights fulfill all the assumptions of theorem 2.3.2, making the bound directly applicable to the proposed weighting scheme. Particularly, we note that,

$$\frac{||w_T||_2}{w_{(m+1)}}\sqrt{\frac{log(\tilde{p})}{n}} \leq \frac{e^{-\frac{n(\delta^2+\alpha)}{8}} + \frac{1}{n^{\frac{1}{2}-\gamma}}}{1-e^{-\frac{n(\delta^2+\alpha)}{8}} + \frac{1}{n^{\frac{1}{2}-\gamma}}}\sqrt{\frac{log(\tilde{p})}{n}} = o(1)$$

Next, in the following lemma 2.3.4, we discuss the order of the sequence of tuning parameters $\lambda_1 < \lambda_2, \dots, \lambda_{2r}$ which we specifically designed in Section 2.2.3 for the importance score. Additionally, we show the order of the penalty level λ_{max} after which the unweighted lasso returns an empty model with all estimated coefficients $\tilde{\beta}^{uw}(\lambda_{max}) = 0$

Lemma 2.3.4 (Order of the designed tuning parameters)**.** *Among the sequence of tuning parameters considered, which has a length of* $2r$*, the first* r *smallest tuning parameters follow an order of* $o\left(\sqrt{\frac{lop(\tilde{p})}{n}}\right)$*, while the last* r *largest tuning parameters are* $o(n^{\gamma})$ *for* $\gamma < \frac{1}{2}$*. Additionally,* $\lambda_{max} < o(n^{\gamma})$ *where* $\lambda_{max} = \sup\{\lambda : \max_{j\in 1,2,\dots,\tilde{p}} |\tilde{\beta}^{uw}(\lambda)| = 0\}$

Proof. By construction, the first r smallest tuning parameters follow an order of $o\left(\sqrt{\frac{lop(\tilde{p})}{n}}\right)$. On the other hand, the sequence of r largest tuning parameter is greater than the order of $\frac{1}{n^{\frac{1}{2}-\gamma}}$.

Hence, $\frac{\sqrt{m}}{\sqrt{n}||w_T||_2} \leq \frac{\sqrt{m}}{\sqrt{n}\frac{1}{n^{\frac{1}{2}-\gamma}}} = o(n^{\gamma})$ and our chosen λ-sequence can be application for the L_2 bound.

Next, in order to study λ_{max}, we first note that it is the level of penalty where the first variable enters the lasso model along the regularization path. Hence,

$$\begin{aligned}\lambda_{max} = \max_{j\in\{1,2,\dots,\tilde{p}\}} |X_j' y|/n &\leq \frac{1}{n} \max_{j\in\{1,2,\dots,p\}} ||X_j||_2||y||_2 \\ &\leq \frac{B\sqrt{n}}{n}||y||_2 = \frac{B}{\sqrt{n}}||y||_2\end{aligned} \tag{2.21}$$

Now, $||y||_2^2 = \sum_{i=1}^{n} y_i^2$, where $y_i \sim N(X_i'\beta^*, \sigma^2)$ independently. Hence, $||y||_2^2 = \sum_{i=1}^{n} \sigma^2 u_i^2$, where $u_i^2 = \frac{y_i^2}{\sigma^2} \sim \chi_1^2\left((X_i'\beta^*)^2\right)$; which implies

$$\begin{aligned}E||y||_2^2 = \sum_{i=1}^{n} \sigma^2 \left(1 + (X_i'\beta^*)^2\right) &\leq \sum_{i=1}^{n} \sigma^2 \left(1 + (B\sum_{j\in S} \beta_j^*)^2\right) \\ \implies \frac{E||y||_2^2}{n^{1+\gamma}} \to 0 \implies \frac{E||y||_2}{n^{\frac{1+\gamma}{2}}} \to 0\end{aligned} \tag{2.22}$$

for any $\frac{1}{2} > \gamma > 0$. Hence, by Markov inequality, $P(||y||_2 > \frac{1+\gamma}{2}) \leq \frac{E||y||_2}{n^{\frac{1+\gamma}{2}}} \to 0$. This implies $\lambda_{max} \leq \frac{B}{\sqrt{n}}||y||_2 \leq n^{\frac{\gamma}{2}} \leq \lambda_{r+1}$. □

Lemma 2.3.4 highlights several crucial characteristics: firstly, it establishes the asymptotic ordering of our carefully designed sequence of penalty parameters, i.e., $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_r \leq \lambda_{r+1} \leq \cdots \leq \lambda_{2r}$; secondly, it guarantees that the unweighted importance scores $\mathscr{I}_j^{uw}$ will always remain strictly below 1 since our chosen sequence of tuning parameters mostly exceeds λ_{max}. It also, demonstrates that, due to the construction, $\lambda_{r+1} \leq \cdots \leq \lambda_{2r}$ are greater than the order of $o(\frac{\sqrt{s}}{\sqrt{n}||w_S||_2})$. The following lemma 2.3.2 demonstrates how the error bound in theorem 2.3.2 further characterizes the behavior of the null importance scores defined in 2.4. This shows that for the null features, the weighted and unweighted importance scores become similar asymptotically.

Lemma 2.3.5 (Asymptotic behaviour of the importance scores)**.** *For any cutoff* $\Delta > 0$*, as* $n \to \infty$

$$P\left(\sum_{j\in S^c} |1\left(\mathscr{I}_j^{w} \geq \Delta\right) - 1\left(\mathscr{I}_j^{uw} \geq \Delta\right)| > \epsilon\right) \to 0 \tag{2.23}$$

Proof. For any $\mathbf{j} \in \mathbf{S^c}$, $|\hat{\beta}_j^w| \le o_P(1)$ for whole sequence of tuning parameters $\lambda_1, \dots, \lambda_{2r}$. Hence, we have, for any $\Delta > 0$,

$$P\left(\mathscr{I}_j^w > \Delta\right) = P\left(\sum_{i=1}^{2r} 1\left(|\hat{\beta}_j^w(\lambda_i)| > \tau\right) > \Delta\right) \le \frac{\sum_{i=1}^{2r} P(|\hat{\beta}_j^w| > \tau)}{\Delta} \le \frac{2re^{-c_5 n}}{\Delta}$$

for $c_5 > 0$. Similarly,

$$\begin{aligned} P\left(\mathscr{I}_j^{uw} > \Delta\right) &= P\left(\sum_{i=1}^{2r} 1\left(|\hat{\beta}_j^{uw}(\lambda_i)| > \tau\right) > \Delta\right) \\ &\le P\left(\sum_{i=1}^{r} 1\left(|\hat{\beta}_j^{uw}(\lambda_i)| > \tau\right) > \frac{\Delta}{2}\right) + P\left(\sum_{i=r+1}^{2r} 1\left(|\hat{\beta}_j^{uw}(\lambda_i)| > \tau\right) > \frac{\Delta}{2}\right) \\ &\le \frac{re^{-c_5 n}}{\Delta} \end{aligned}$$

So, with a fixed $\Delta \in (0,1)$, for $\mathbf{j} \in \mathbf{S^c}$, we have $1\left(\mathscr{I}_j^w \ge \Delta\right) \xrightarrow{P} 0$ and $1\left(\mathscr{I}_j^{uw} \ge \Delta\right) \xrightarrow{P} 0$. Hence, for a null feature, the weighted and unweighted importance scores behave similarly. Additionally, due to the monotonicity wrt Δ, these convergences are uniform for $\Delta \in (0,1)$.

So, for any $j \in S^c$, for $\Delta \in (0,1)$ and $\epsilon > 0$,

$$\begin{aligned} P\left(|1\left(\mathscr{I}_j^w \ge \Delta\right) - 1\left(\mathscr{I}_j^{uw} \ge \Delta\right)| > \epsilon\right) &\le P\left(1\left(\mathscr{I}_j^w \ge \Delta\right) + 1\left(\mathscr{I}_j^{uw} \ge \Delta\right) > \epsilon\right) \\ &\le P\left(1\left(\mathscr{I}_j^w \ge \Delta\right) > \frac{\epsilon}{2}\right) + P\left(1\left(\mathscr{I}_j^{uw} \ge \Delta\right) > \frac{\epsilon}{2}\right) \le \frac{6re^{-c_5 n}}{\epsilon\Delta} \end{aligned}$$

Furthermore, for a sequence $u_n = o(\frac{1}{n})$,

$$P\left(\sum_{j \in S^c} |1\left(\mathscr{I}_j^w \ge \Delta\right) - 1\left(\mathscr{I}_j^{uw} \ge \Delta\right)| > u_n\right) = \tilde{p} \max_{j \in S^c} P\left(|1\left(\mathscr{I}_j^w \ge \Delta\right) - 1\left(\mathscr{I}_j^{uw} \ge \Delta\right)| > u_n\right) \to 0$$

as the basic convergences are in order of exponential to n and $\tilde{p} = o(n^3)$. □

Theorem 2.3.6 (Asymptotic FDR control guarantee of the proposed method)**.** *With* Δ^* *as the data-dependent optimum cutoff defined in eq. 3,*

$$\lim_{n \to \infty} FDR = E\left(\frac{\sum_{j=1}^{p} 1\left(\mathscr{I}_j^w \ge \Delta^*, j \notin S\right)}{\sum_{j=1}^{p} 1\left(\mathscr{I}_j^w \ge \Delta^*\right)}\right) = 0$$

Also, the asymptotic power with the optimum cutoff Δ^* *approaches 1; i.e.,*

$$\lim_{n \to \infty} Power = E\left(\frac{\sum_{j=1}^{p} 1\left(\mathscr{I}_j^w \ge \Delta^*, j \in S\right)}{|S|}\right) = 1$$

The proof is relegated to Appendix A.

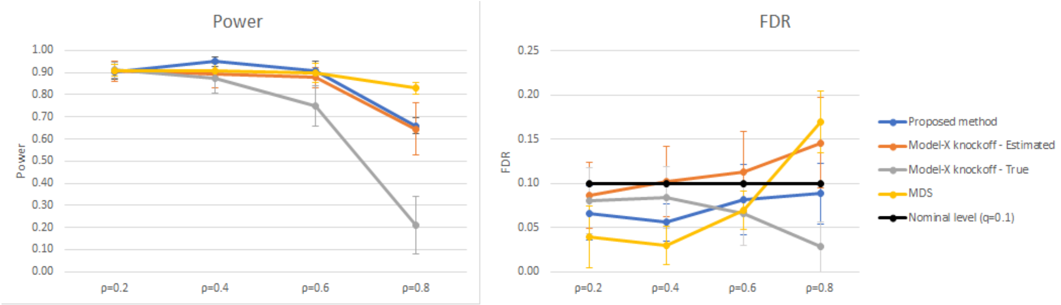

Figure 2.2 Correlation-wise comparison: the effect size and sparsity levels are fixed.

2.4 Simulation Studies

To evaluate the performance of the proposed variable selection method, we conducted a comprehensive simulation study. We simulated data under different scenarios, including varying correlation structures, effect sizes, sparsity, and noise levels.

Specifically, we consider the following high-dimensional linear regression setup: $Y^{n\times 1} \sim N_p(X^{n\times p}\beta^{p\times 1}, \sigma^2 I_n)$ with p=1000 and $n = 600$.The true index set S is randomly generated maintaining $|S| = s$ with $\beta_{S^c} = 0$ and the values of β_S are randomly drawn from N(β_0,0.1). For the design matrix, each $X_i, i = 1, 2, ..., n$ are randomly drawn from $N_p(0, \Sigma)$ where $\Sigma_{ij} = cov(X_i, X_j) = \rho^{|i-j|}$.

1. **Correlation wise comparison**: we fix the effect size $|\beta_0| = 1$ and correlation between X_i and X_j varied as $\rho = \{0.2, 0.4, 0.6, 0.8\}$ keeping $s = 20$.

2. **Effect-size wise comparison**: We fix ρ at 0.5, $s = 20$ and varied the effect size with $|\beta_0| = \{0.6, 0.8, 1, 1.2\}$.

3. **Sparsity-size wise comparison**: We fix ρ at 0.5 and β at 1, then gradually increase $s = \{10, 20, 30, 40, 50\}$

We compared the performance of our proposed methods with several existing methods, including the Model-X knockoff Candès et al. (2018) and the multiple data splitting (MDS) approach Dai et al. (2022) for FDR control. As we discussed in Section 2.1, the Model-X knockoff Candès et al. (2018) is a variable selection method that controls the FDR by constructing a set of "knockoff" variables that mimic the correlation structure of the original variables. On the other hand,

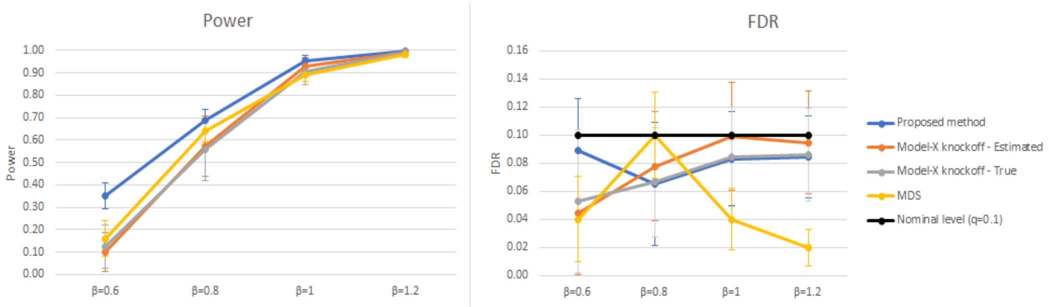

Figure 2.3 Effect size-wise comparison: the autocorrelation and sparsity levels are fixed.

the MDS is a recently developed method based on the random splitting based technique that assesses the stability of selected features by repeating the variable selection process on multiple random subsamples of the data. These two methods represent a wide class of feature selection models in the current literature and hence we choose these two as other competing methods for the empirical study.

We evaluated the performance of these methods using the FDR as the primary metric, controlling for a pre-specified FDR level at $q = 0.10$. We also computed the true positive rate (TPR) to assess the power of the methods. We repeated each simulation scenario 100 times to ensure statistical reliability and calculated the mean and standard deviation of the FDR, and Power across replications. For the implementation of the knockoff procedure, we adopt the two-stage approach. First, we randomly split the data into two halves. In the first part, we apply the screening process as in Wasserman and Roeder (2009) and screen out most of the null variables with sure screening property, while in the second part, we apply the knockoff procedure to clean the noise variables with FDR control. For the knockoff variables, we generate second-order Gaussian knockoffs using the estimated model parameters with full semidefinite programming (SDP) construction. In spite of its computational complexity, the SDP knockoffs are statistically superior by having higher power than its other alternatives for creating knockoff variables. For the test statistic for the knockoff approach, we considered the signed maximum statistic: $W_j = max(Z_j, \tilde{Z}_j) \cdot sign(Z_j - \tilde{Z}_j)$, where Z_j and $\tilde{Z}_j$ are the maximum values of λ at which the jth variable and its knockoff, respectively, enter the generalized linear model. To implement the MDS method, we use the R-code provided in their paper Dai et al. (2022).

Figure 2.4, 2.4, and 2.4 illustrates the power and FDR comparisons in the correlation-wise, effect size-wise, and sparsity-wise experiments respectively. We note that the performance of the knockoffs is dependent on the estimation of the feature distribution in order to generate the knockoffs. Hence to separate this out, we consider the "Model-X knockoff - true" where the true feature's distribution is used to generate the knockoffs. This is possible here as thein the simulation setup, we know the true data-generating distributions. For a more realistic application, we show "Model-X knockoff - Estimated" where the distribution of the feature is estimated from the data assuming Gaussian distribution. Our simulation results indicate several interesting observations. First, the proposed method quickly gains power for moderate correlation or effect sizes, while successfully maintaining the FDR below the specified threshold 10%. Second, in Figure 2.4, we can see the estimated Model-X knockoff achieves higher power; but it also loses its FDR control. However, the Model-X knockoff-True is not affected by the higher correlation and maintains the FDR control at the cost of reduced power. This experiment further substantiates the claim of Barber et al. (2020) that for high multicollinearity, the error in estimating the feature's distribution increases which further indices the inflated FDR. Third, consistently, for all the cases, the MDS method is highly conservative, although in the simulation setting its power is comparable to the proposed method.

In conclusion, our simulation study demonstrates that the proposed method performs well in terms of controlling the FDR and achieving high power for variable selection in high-dimensional settings. These findings suggest that the proposed method can be a valuable tool for high-dimensional variable selection in various applications.

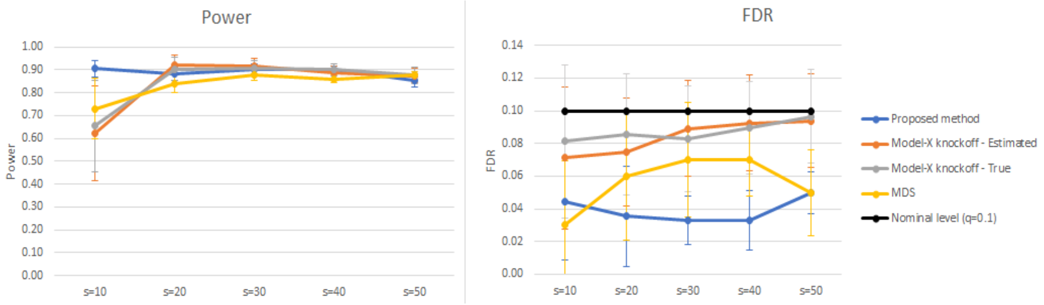

Figure 2.4 Sparsity-wise comparison: the autocorrelation and effect size are fixed.

Table 2.1 Drug-sensitive genes identified by the feature selection methods

Drug	Genes selected			Confirming references
	the proposed method	Model-X knockoff	MDS	
Topotecan	***SLFN11 (10)***, SF3A2 (10), RPL18 (7), AC018755.1 (5), THG1L (5),PSAP (5), THRB (5)	RP1.199J3.5 (9), ***SLFN11 (9)***, MGST3 (8), RPL18 (8), CLCN4 (8), OXCT2P1 (7), GAREML (7), BHLHE40 (7), THG1L (7), AP000974.1 (7), AC018755.1 (7), DCHS1 (7), LRRC7 (6), SGK223 (6), MFSD5 (6), RP11.562A8.5 (6), SUSD2 (6), CA5B (6), FAM86EP (5), UFSP2 (5), PLIN2 (5), PC (5), SH3BP1 (5), CDKN2C (5), PSAP (5), SF3A2 (5), KANK3 (5)	***SLFN11 (7)***, RP1.199J3.5 (5)	Barretina et al. (2012) Li et al. (2012)
Irinotecan	***SLFN11 (7)***, WTAPP1 (6), FKBP2 (5), SYT13 (5), IFITM10 (5), AC068580.5 (5)	ADORA2A (8), ***SLFN11 (7)***, TCEANC2 (7), IFITM10 (7), CUEDC1 (6), LMNB1 (5), C7orf26 (5), SYT13 (5), C16orf71 (5), RRAD (5), AC068580.5 (5), RPL7AP66 (5), THG1L (5), ARHGAP19 (5), MBNL2 (5), GOLGA5 (5), SLC48A1 (5), FKBP2 (5), ARL2 (5), WTAPP1 (5), CETN2 (5)	No genes met the selection criteria of MDS	Li et al. (2012)
17-AAG	MB21D1 (10), KCNK7 (10), ***NQO1 (9)***, MMP24 (9), SERPINF1 (9), DNAJC17 (9), CSNK1E (9), NAPG (8), FTSJ2 (8), RP11.442H21.2 (7), CTDSP1 (7), SEPW1 (6), SUSD4 (6), RP4.816N1.6 (6), HOXA11 (6), DIMT1 (6), RP11.143J12.3 (6), ERLIN1 (5), SLC12A7 (5), RP13.15M17.1 (5), TBC1D4 (5), LYNX1 (5), LINC01006 (5), CPED1 (5), THRB (5), WBP2 (5), DHRS4.AS1 (5), NEK6 (5)	KCNK7 (10), MB21D1 (10), DNAJC17 (10), CSNK1E (10), FTSJ2 (9), ERLIN1 (9), SERPINF1 (9), MMP24 (9), CTDSP1 (8), RP11.442H21.2 (8), DHRS4.AS1 (8), GNPNAT1 (8), ***NQO1 (8)***, NAPG (8), RP11.218L14.4 (8), DIMT1 (8), HOXA11 (8), CPED1 (8), LINC01006 (8), OSBPL9 (7), THRB (7), SLC12A7 (7), RP4.816N1.6 (7), WBP2 (7), RP11.143J12.3 (7), SEPW1 (7), LYNX1 (7), WWP2 (7), NEK6 (7), RP13.15M17.1 (6), SUSD4 (6), ATP6V0E1 (6), SEMA4G (6), TBC1D4 (6), ZNF571.AS1 (6), PFKFB1 (6), PLA2G4A (6), BICC1 (6), DGAT2 (6), RNF121 (6), CLSTN2 (5), OTUD4 (5), NARS2 (5), ZNF420 (5), RP11.661A12.12 (5), RP11.432J22.2 (5), RP11.644F5.11 (5), C12orf73 (5), AMN (5), UACA (5), GPRC5B (5), ZNF506 (5)	No genes met the selection criteria of MDS	Hadley and Hendricks (2014) , Barretina et al. (2012)
PaclitaXel	ELOVL1 (10), ABCB1 (10), ***BCL2L1 (9)***, PRODH (9), STX10 (9), SHANK2 (9), SDHAP3 (9), SUCO (8), PCDHGA2 (7), RUNDC3B (7), TYMP (6), NMB (6), C8orf46 (6), UQCRFS1P1 (6), BUB1B (5), INHBA (5), ANKRD36BP2 (5), NOS1AP (5), PRKX (5), SDF2 (5), SH3BP1 (5), LRRC16B (5), RIMKLA (5)	ELOVL1 (9), SUCO (9), ABCB1 (9), SHANK2 (8), ARL6IP1 (8), STX10 (8), PRODH (8), SDHAP3 (8), RUNDC3B (7), C8orf46 (7), YTHDF1 (7), TYMP (7), LRRC16B (7), NOS1AP (6), PCDHGA2 (6), LSMEM1 (6), BUB1B (6), SPATA5L1 (6), RP11.862L9.3 (6), PRKX (6), UQCRFS1P1 (6), BAK1 (6), RIMKLA (5), PDZK1IP1 (5), ANKRD36BP2 (5), ENC1 (5), INHBA (5), RP11.644F5.11 (5), NMB (5), CSNK2A1 (5), ***BCL2L1 (5)***, EXOC5P1 (5), SDF2 (5), CETN2 (5), SH3BP1 (5)	ABCB1 (5), LRRC16B (5)	Dorman et al. (2016) Lee et al. (2016)
AEW541	***IQGAP2*** (8), SOAT1 (8), KCNAB1 (8), DERL3 (7), LINC00324 (7), SLC44A1 (7), IRS2 (7), AC008132.12 (7), CASP10 (7), TMEM101 (6), TSPYL5 (5)	TSPYL5 (8), SLC44A1 (8), LINC00324 (8), DERL3 (8), SOAT1 (7), IRS2 (7), TMEM101 (7), KCNAB1 (7), ***IQGAP2 (6)***, CYP1B1.AS1 (6), CASP10 (6), VILL (6), AC004840.9 (6), ATP11B (6), AC008132.12 (6), ANO10 (5), MAF (5), UNC119 (5)	TSPYL5 (5)	Liang et al. (2018)

2.5 Real data analysis

In this section, we present real data analysis to illustrate the utility of our proposed statistical method for high-dimensional data analysis. For this application, we choose the CCLE dataset that was generated by the Cancer Cell Line Encyclopedia project (CCLE, link available here. The dataset contains gene expression profiles of over 100 cancer cell lines across different cancer types, making it a valuable resource for studying the molecular basis of cancer and identifying potential drug targets. With the ultimate goal of improving precision medicine, the CCLE dataset has been extensively used in previous studies to investigate various aspects of cancer biology, such as identifying genetic and epigenetic markers associated with drug sensitivity or resistance, characterizing molecular subtypes of different cancer types, and exploring the genetic basis of cancer evolution and progression.

More specifically, the CCLE dataset contains dose-response curves for 24 different drugs across over $n = 400$ cell lines, with the expression data of $p = 18,926$ genes for each cell line considered as features. To measure drug sensitivity, we used the activity area Barretina et al. (2012). Specifically, in this study, we aim to identify the set of genes associated with the sensitivity of five specific anticancer drugs, namely Topotecan, 17-AAG, Irinotecan, Paclitaxel, and AEW541, which have been used to treat various cancer types including ovarian and lung cancer. Previous studies have already investigated these drugs and related gene expression data, more details can be found at Barretina et al. (2012). We implemented the proposed method along with the two other competing feature selection methods Model-X knockoffs Candès et al. (2018) and MDS Dai et al. (2022) on the CCLE dataset, at the nominal FDr control level $q = 0.2$. Each method selected some genes to indicate that they are highly associated with these five above-mentioned drugs. However, in a real setting, assessing the performance of a feature selection method is difficult as the underlying data-generating mechanism is completely unknown. So, We validate the results of these methods in two ways. First, we check if the selected genes can be confirmed using the existing domain knowledge. Second, for more empirical validation, we perform a 10-fold cross-validation and check the out-of-sample test accuracy only using the handful of selected genes by each of the methods. The results are summarized below.

Table 2.1 reports the genes selected in at least five validation out of the 10-fold cross-validation for each of the three methods considered. The selected genes for these drugs exhibit a high degree of consistency with our current knowledge and consistently appeared in multiple cross-validation runs. For instance, in the case of Topotecan and Irinotecan, the proposed method identifies *SLFN11* as the top drug-sensitive gene in the majority of the validation iterations. This finding is consistent with prior research, as Barretina et al. (2012); Zoppoli et al. (2012) previously reported that *SLFN11* is highly predictive for both drugs. Similarly, for 17-AAG, the proposed method identifies *NQO1* as the topmost important gene, which is known to be highly sensitive to 17-AAG Hadley and Hendricks (2014). In Table 2.1, the genes that are confirmed by previous research are highlighted in red. One would observe that, for all the drugs, the proposed method selects these important genes more consistently for repeated validations compared to the other competing methods. Although the FDR control level is the same for all these three methods, knockoff consistently selected a higher number of genes. This is justifiable as the knockoff's FDR inflates Barber et al. (2020) if the feature distribution is not estimated properly. Also, similar to the simulation study, MDS is highly conservative and fails to discover any gene under the 20% FDR control.

For a more empirical validation, Table 2.2 reports the average number of genes selected in the 10 cross-validation runs and the average cross-validation test Mean Square Error (MSE). Additional to the Model-X knockoff and MDS, here we also show the results for the cross-validated lasso, which under mild regularity conditions enjoys the sure screening property. Table 2.2 shows the proposed method consistently maintains the prediction accuracy compared to the other methods while selecting much fewer genes. This indicates that the proposed method even after sufficient dimension reduction, the proposed method successfully retains the important genes with higher predictive ability.

Table 2.2 Drug-sensitive gene selection and related prediction performance

Drug	selected number of genes				Test MSE			
	Lasso	Proposed method	Model-X knockoff	MDS	Lasso	Proposed method	Model-X knockoff	MDS
Topotecan	96.3	36.2	50.4	3.8	1.08	1.12	1.61	1.56
17-AAG	103.6	40.1	71.7	1.02	0.87	0.94	1.07	1.14
Irinotecan	52.5	11.9	23.9	0.7	0.63	1.04	1.96	5.01
Paclitaxel	127.3	40.3	57.8	4.6	1.36	1.94	2.21	2.91
AEW541	48.6	16.2	30.4	1.3	0.33	0.35	0.37	0.36

2.6 Conclusion

In this paper, we have proposed a novel p-value-free FDR controlling method for ultrahigh dimensional variable selection. We effectively utilized the adaptive penalization framework and used it to distinguish between the null and nonnull features. Due to the dimension reduction in the screening step, our method is computationally very efficient compared to the Model-X knockoff or the MDS method. Our empirical results show that the proposed method has an excellent performance in terms of FDR control while maintaining high power compared to other existing methods. Unlike the knockoff-based methods, the proposed methodology does not need much prior knowledge of the joint distribution of features, making it conceptually simple and easy to implement. We have provided theoretical guarantees for FDR control under mild assumptions on the design matrix and the response variable.

There are several promising directions for future development that are worth considering. The proposed method is applicable to both linear and generalized linear models. As the idea of adaptive penalization can be easily formalized to more complex nonparametric models, it would be interesting to investigate the potential use and theoretical properties of the proposed methodology in handling neural networks and other non-linear models, particularly with regard to more intricate data types such as natural language and computer vision. Additionally, as the idea of L_1 penalization is used in longitudinal data Barber et al. (2017), another idea of potential future direction is to extend the methodology to incorporate longitudinal or hierarchical structures datasets which is quite common in finance or imaging studies.

BIBLIOGRAPHY

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., Reimherr, M., and Schill, T. (2017). The function-on-scalar LASSO with applications to longitudinal GWAS. *Electronic Journal of Statistics*, 11(1):1351 – 1389.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., et al. (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607.

Bates, S., Candè s, E., Janson, L., and Wang, W. (2020). 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 (Methodological)*, 57(1):289–300.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370.

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(3):pp. 551–577.

Dai, C., Lin, B., Xing, X., and Liu, J. S. (2022). False discovery rate control via data splitting. *Journal of the American Statistical Association*, pages 1–18.

Das, D. and Lahiri, S. N. (2019). Distributional consistency of the lasso by perturbation bootstrap. *Biometrika*, 106(4):957–964.

Dorman, S. N., Baranova, K., Knoll, J. H. M., Urquhart, B. L., Mariani, G., Carcangiu, M. L., and Rogan, P. K. (2016). Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning. *Mol. Oncol.*, 10(1):85–100.

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.

Gimenez, J. R., Ghorbani, A., and Zou, J. (2019). Knockoffs for the mass: new feature importance statistics with false discovery guarantees.

Hadley, K. E. and Hendricks, D. T. (2014). Use of nqo1 status as a selective biomarker for oesophageal squamous cell carcinomas with greater sensitivity to 17-aag. *BMC cancer*, 14(1):1–8.

Lee, H. J., Hanibuchi, M., Kim, S.-J., Yu, H., Kim, M. S., He, J., Langley, R. R., Lehembre, F., Regenass, U., and Fidler, I. J. (2016). Treatment of experimental human breast cancer and lung cancer brain metastases in mice by macitentan, a dual antagonist of endothelin receptors, combined with paclitaxel. *Neuro-oncology*, 18(4):486–496.

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

Liang, F., Li, Q., and Zhou, L. (2018). Bayesian neural networks for selection of drug sensitive genes. *Journal of the American Statistical Association*, 113(523):955–972.

Minnier, J., Tian, L., and Cai, T. (2011). A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association*, 106(496):1371–1382.

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

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Wang, S., Weng, H., and Maleki, A. (2017). Which bridge estimator is optimal for variable selection? *arXiv preprint arXiv:1705.08617*.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of statistics*, 37(5A):2178.

Xie, H. and Huang, J. (2009). Scad-penalized regression in high-dimensional partially linear models.

Zoppoli, G., Regairaz, M., Leo, E., Reinhold, W. C., Varma, S., Ballestrero, A., Doroshow, J. H., and Pommier, Y. (2012). Putative dna/rna helicase schlafen-11 (slfn11) sensitizes cancer cells to dna-damaging agents. *Proceedings of the National Academy of Sciences*, 109(37):15030–15035.

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. (2003). Regression shrinkage and selection via the elastic net, with applications to microarrays. *JR Stat Soc Ser B*, 67:301–20.

APPENDIX A

PROOF OF THEOREM 2.3.6

We recall that in order to show the FDR control, we only need to verify $\lim_{n\to\infty} E(R(\Delta^*)) \leq 1$, where Δ^* is the data-dependent optimum cutoff in eq. 3 and

$$R(\Delta^*) = \frac{\sum_{j\notin S} 1\left(\mathscr{I}_j^w \geq \Delta^*\right)}{\sum_{j\in S} 1\left(\mathscr{I}_j^{uw} \geq \Delta^*\right) + \sum_{j\notin S} 1\left(\mathscr{I}_j^{uw} \geq \Delta^*\right)}$$

By constructing Δ^*, the denominator of the $R(\Delta^*)$ is strictly positive. So, in order to show $\lim_{n\to\infty} E(R(\Delta^*)) \leq 1$, we show the following:

1. $P(\Delta^* \geq \frac{1}{2}) \to 1$ as $n \to \infty$.

2. $\sum_{j\in S} 1\left(\mathscr{I}_j^{uw} \geq \Delta^*\right) > 0$, and

3. $P\left(|\sum_{j\notin S} 1\left(\mathscr{I}_j^w \geq \Delta^*\right) - \sum_{j\notin S} 1\left(\mathscr{I}_j^{uw} \geq \Delta^*\right)| > u_n\right) \to 0$, as $n \to \infty$ with $u_n = o(\frac{1}{n})$.

We will show this one by one.

Part 1:, Now, by the error bound in Lemma 2.3.2, for any $\mathbf{j} \in \mathbf{S}$,

1. $P\left(|\hat{\beta}_j^{uw}| > \tau\right) > 1 - o(e^{-nc_6}), \forall \lambda_1, \ldots, \lambda_r$, but

2. $P\left(|\hat{\beta}_j^{uw}| < \tau\right) > 1 - o(e^{-nc_6}), \forall \lambda_{r+1}, \ldots, \lambda_{2r}$.

3. $P\left(|\hat{\beta}_j^{w}| > \tau\right) > 1 - o(e^{-nc_6}), \forall \lambda_1, \ldots, \lambda_{2r}$

Additionally, for any $\mathbf{j} \in \mathbf{S^c}$, $P\left(|\hat{\beta}_j^{w}| < \tau\right) > 1 - o(e^{-nc_6}), \forall \lambda_1, \ldots, \lambda_{2r}$. Similar inequality holds for the unweighted lass0 as well. Hence, with sufficiently small $\tau < \psi$, $\mathscr{I}_j^w = \frac{1}{2r}\sum_{i=1}^{2r} 1\left(|\hat{\beta}_j^w(\lambda_i)| > \tau\right) \xrightarrow{P} 1$. Similarly, for $\mathbf{j} \in \mathbf{S}$, $\mathscr{I}_j^{uw} \xrightarrow{P} \frac{1}{2}$, as $n \to \infty$.

Now, for any cutoff Δ, we observe,

$$F\hat{P}R(\Delta) = \frac{\sum_{j=1}^{p} 1\left(\mathscr{I}_j^{uw} \geq \Delta\right)}{\sum_{j=1}^{p} 1\left(\mathscr{I}_j^{w} \geq \Delta\right)} = \frac{\sum_{j\in S} 1\left(\mathscr{I}_j^{uw} \geq \Delta\right) + \sum_{j\notin S} 1\left(\mathscr{I}_j^{uw} \geq \Delta\right)}{\sum_{j\in S} 1\left(\mathscr{I}_j^{w} \geq \Delta\right) + \sum_{j\notin S} 1\left(\mathscr{I}_j^{w} \geq \Delta\right)} = \frac{Z_1(\Delta) + Z_2(\Delta)}{Z_3(\Delta) + Z_4(\Delta)} \qquad \text{(A.1)}$$

Now, for any $\epsilon > 0$ and $\Delta \leq \frac{1}{2}$, as $\mathscr{I}_j^{uw} \xrightarrow{P} 1$ as $n \to \infty$ for $j \in S$

$$P(Z_1(\Delta) < s - \epsilon) = P\left(\exists \text{ at least one } j \in S \text{ for which } \mathscr{I}_j^{uw} < \Delta\right) \leq \sum_{j \in S} P\left(\mathscr{I}_j^{uw} < \Delta\right) \to 0$$

Hence, $Z_1(\Delta) \xrightarrow{P} s$ for fixed $\Delta \leq \frac{1}{2}$.

Further we observe, as $\mathscr{I}_j^{uw} \xrightarrow{P} 0$ as $n \to \infty$ for $j \notin S$ with exponential rate,

$$P(Z_2(\Delta) > \epsilon) = P\left(\exists \text{ at least one } j \notin S \text{ for which } \mathscr{I}_j^{uw} \geq \Delta\right) \leq \sum_{j \notin S} P\left(\mathscr{I}_j^{uw} \geq \Delta\right) \to 0$$

Hence, $Z_2(\Delta) \xrightarrow{P} 0$ for fixed $\Delta \leq \frac{1}{2}$.

Similarly, $Z_3(\Delta) \xrightarrow{P} s$ and $Z_4(\Delta) \xrightarrow{P} 0$ for fixed $\Delta \leq \frac{1}{2}$. Hence,

$$\frac{Z_1(\Delta) + Z_2(\Delta)}{Z_3(\Delta) + Z_4(\Delta)} \xrightarrow{P} 1 \Rightarrow F\hat{P}R(\Delta) \xrightarrow{P} 1 \text{ for any } \Delta \leq \frac{1}{2}$$

Next, the uniform convergence can be established by observing that each of the process $Z_i(\Delta)$ is monotonic with respect to Δ for $i = 1, 2, 3, 4$ and hence $\inf_{\Delta \leq \frac{1}{2}} Z_1(\Delta) \xrightarrow{P} s, \sup_{\Delta \leq \frac{1}{2}} Z_2(\Delta) \xrightarrow{P} 0, \inf_{\Delta \leq \frac{1}{2}} Z_3(\Delta) \xrightarrow{P} s$, and $\sup_{\Delta \leq \frac{1}{2}} Z_4(\Delta) \xrightarrow{P} 0$. These conjectures further implies $P(\Delta^* \geq \frac{1}{2}) \to 1$ as $n \to \infty$.

Part 2:, We note that $\sup_{\Delta \in (\frac{1}{2}, 1)} P\left(Z_1(\Delta) \geq Z_2(\Delta)\right) \to 1$, as $n \to \infty$. Hence, by construction of Δ^*, $\sum_{j=1}^{p} 1\left(\mathscr{I}_j^{uw} \geq \Delta^*\right) > 0 \Rightarrow P\left(\sum_{j \in S} 1\left(\mathscr{I}_j^{uw} \geq \Delta^*\right) > 0\right) \to 1$.

Part 3:, Following Lemma 2.3.2, we note that,

$$\begin{aligned} &\sup_{\Delta \in (\frac{1}{2},1)} P\left(|\sum_{j \notin S} 1\left(\mathscr{I}_j^{w} \geq \Delta\right) - \sum_{j \notin S} 1\left(\mathscr{I}_j^{uw} \geq \Delta\right)| > \epsilon\right) \\ &\leq \sup_{\Delta \in (\frac{1}{2},1)} P\left(\sum_{j \in S^c} |1\left(\mathscr{I}_j^{w} \geq \Delta\right) - 1\left(\mathscr{I}_j^{uw} \geq \Delta\right)| > \epsilon\right) \to 0 \end{aligned}$$

as $n \to \infty$ due to the exponential bound in Lemma 2.3.2.

These conjectures further imply that $\lim_{n \to \infty} E(R(\Delta^*)) \leq 1$, guaranteeing the asymptotic FDR control.

Next, to study the asymptotic power. we start by observing that $\hat{e}_0(\Delta^*) = \sum_{j \in \hat{S}_n} 1\left(\mathscr{I}_j^{uw} \geq \Delta^*\right) > 0$. We define, $\tilde{\Delta} = \sup_{\Delta > 0}\{\hat{e}_0(\Delta) > 0\}$ and $\tilde{\tilde{\Delta}} = \inf_{\Delta > 0}\{\mathscr{I}_j^{w}, j \in S\}$. Now, by construction, $P(\tilde{\Delta} < \frac{1}{2}) >$

$1-o(e^{-nc_6})$. Additionally, for any $\epsilon > 0$, $P(\tilde{\tilde{\Delta}} < 1-\epsilon) \leq \sum\limits_{j\in S} P(\mathscr{I}_j^w < 1-\epsilon) \leq \sum\limits_{j\in S}\sum\limits_{i=1}^{2r} P\left(|\hat{\beta}_j^w(\lambda_i)| < \tau\right) < o(e^{-nc_6})$.

This implies, $P(\tilde{\tilde{\Delta}} > \tilde{\Delta}) > P\left(\tilde{\tilde{\Delta}} \geq 1-\epsilon, \tilde{\Delta} < \frac{1}{2}\right) > P\left(\tilde{\tilde{\Delta}} \geq 1-\epsilon\right) + P\left(\tilde{\Delta} < \frac{1}{2}\right) - 1 \geq 1 - o(e^{-nc_6}) + 1 - o(e^{-nc_6}) - 1 = 1 - 2o(e^{-nc_6})$. Hence, with high probability the final discovered set $\hat{D}_n = \left\{j \in \hat{S}_n : \mathscr{I}_j^w \geq \Delta^*\right\} \subset \left\{j \in \hat{S}_n : \mathscr{I}_j^w \geq \tilde{\tilde{\Delta}}\right\}$. Hence this implies that the asymptotic power converges to one.

APPENDIX B

PROOF OF THEOREM 2.3.2

Here we prove the major theorem 2.3.2 on the L_2 error bound on the weighted Lasso. This proof consists of three parts. First in Lemma B.0.1, we show that for specifically designed events $\mathscr{A}$ and $\mathscr{B}$, we achieve the desired error bound on $\mathscr{A} \cap \mathscr{B}$. Then, in lemma B.0.2 and B.0.3, we show that the probability of the events $\mathscr{A}$ and $\mathscr{B}$ converges to one. We define again the following notations: $m = \max_{j \in S}\{\text{the rank of } w_j \text{ (from smallest to largest)}\}$, the maximum rank of weights for the true features and $T = \{1 \le j \le p : \text{rank}(w_j) \le m\}$. Consequently, we define, w_T = the subspace of $w = w_1, w_2, \ldots, w_p$ indexed by T; and $w_{min} = \min_{j \in S^c} w_j$. In Section 2.3, we show our proposed weighting scheme satisfies all these assumptions, hence enjoying the tight error bound for weighted lasso even for an increasing sequence of tuning parameter λ_n.

Lemma B.0.1. *Define the probability events:*

$$\mathscr{A} = \left\{ \lambda \ge 2max \left\{ \frac{||\frac{1}{n}\epsilon' X_T||_2}{||w_T||_2}, max_{j \in T^c} |\frac{1}{n}\epsilon' X_j w_j^{-1}| \right\} \right\}, \tag{B.1}$$

$$\mathscr{B} = \left\{ \min_{\Delta \in \mathscr{C}} \frac{\Delta' \frac{1}{n} X' X \Delta}{||\Delta||_2^2} \ge \kappa_n \right\}, \tag{B.2}$$

where $\mathscr{C} = \left\{ \Delta \in \mathscr{R}^p : \sum_{j \in T^c} w_j |\Delta_j| \le 3 ||w_T||_2^2 ||\Delta_T||_2^2 \right\}$*. Then on the event,* $\mathscr{A} \cap \mathscr{B}$*, it holds that*

$$||\hat{\beta}^w - \beta^*||_2 \le \frac{3}{2} \frac{\lambda}{\kappa_n} ||w_T||_2 \tag{B.3}$$

Proof. We first note that,

$$\frac{1}{2n} \parallel Y - X\hat{\beta}^w \parallel_2^2 + \lambda \sum_{j=1}^{p} w_j |\hat{\beta}_j^w| \le \frac{1}{2n} \parallel Y - X\beta^* \parallel_2^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j^*|$$

$$\implies 0 \le \frac{1}{2n} \parallel X\hat{\beta}^w - X\beta^* \parallel_2^2 \le \frac{1}{n}\epsilon' X(\hat{\beta}^w - \beta^*) + \lambda \sum_{j=1}^{p} w_j|\beta_j^*| - \lambda \sum_{j=1}^{p} w_j|\hat{\beta}^w| \tag{B.4}$$

$$\le \frac{\lambda}{2}\left[2\sum_{j\in T} w_j|\beta_j^*| - 2\sum_{j\in T} w_j|\hat{\beta}_j^w| - 2\sum_{j\in T^c} w_j|\hat{\beta}_j^w|\right] \tag{B.5}$$

$$+ \frac{1}{n}\epsilon' X_T(\hat{\beta}^w - \beta^*)_T + \frac{1}{n}\epsilon' X_{T^c}(\hat{\beta}^w - \beta^*)_{T^c} \tag{B.6}$$

$$\le \frac{\lambda}{2}\left[||w_T||_2||(\hat{\beta}^w - \beta^*)_T||_2 + 2\sum_{j\in T} w_j|\hat{\beta}_j^w - \beta_j^*| - \sum_{j\in T^c} w_j|\hat{\beta}_j^w - \beta_j^*|\right] \tag{B.7}$$

As by the construction of λ,

1. $\frac{1}{n}\epsilon' X_T(\hat{\beta}^w - \beta^*)_T \le ||\frac{1}{n}\epsilon' X_T||_2||(\hat{\beta}^w - \beta^*)_T||_2 \le \frac{\lambda}{2}||w_T||_2||(\hat{\beta}^w - \beta^*)_T||_2$, and
2. $\frac{1}{n}\epsilon' X_{T^c}(\hat{\beta}^w - \beta^*)_{T^c} \le \max_{j\in T_c}|\frac{1}{n}\epsilon' X_j w_j^{-1}| \sum_{j\in T_c} w_j|\hat{\beta}_j^w - \beta_j^*|$.

Hence,the error vector $\Delta = \hat{\beta}^w - \beta^*$ belongs to the cone $\mathscr{C} = \{\Delta \in \mathscr{R}^p : \sum_{j\in T^c} w_j|\Delta_j| \le 2\sum_{j\in T} w_j|\Delta_j| + ||w_T||_2||\Delta_T||_2 \le 3||w_T||_2||\Delta_T||_2\}$. Now, with the RE property in 2.2.1, we continue from eq. B.7,

$$\Delta^T \frac{1}{n} X' X \Delta \le \frac{\lambda}{2}\left[||w_T||_2||\Delta_T||_2 + 2\sum_{j\in T} w_j|\Delta_j|\right] \tag{B.8}$$

$$\implies \kappa_n||\Delta||_2^2 \le \frac{3}{2}\lambda||w_T||_2||\Delta_T||_2 \tag{B.9}$$

$$\implies ||\Delta_T||_2 \le \frac{3}{2}\frac{\lambda}{\kappa_n}||w_T||_2 \tag{B.10}$$

□

Now to show that $\mathscr{A} \cap \mathscr{B}$ holds with high probability, we assume each row of X is iid from a mean zero sub-Gaussian distribution (with bounded sub-Gaussian norm) with $E(X_i X_i') = \Sigma$ such that $0 \le \rho_0 \le \lambda_{min}(\Sigma) \le \lambda_{max}(\Sigma) \le \rho_1 < \infty$. The assumptions we consider in Section 2.2.1 satisfy this sub-gaussianity condition.

Lemma B.0.2. *Recall* $\mathscr{A} = \left\{\lambda \ge 2max\left\{\frac{||\frac{1}{n}\epsilon' X_T||_2}{||w_T||_2}, max_{j\in T^c}|\frac{1}{n}\epsilon' X_j w_j^{-1}|\right\}\right\}$. *Further, choose,* $\lambda \ge 2\tau\sigma\left[\sqrt{\frac{m}{n}}\frac{1}{||w_T||_2}\right]$ *with* $\tau = 2\sqrt{\rho_1} + \frac{2}{\sqrt{\rho_0}}(2 + \frac{m}{n})$. *Then,* $P(\mathscr{A}) \ge 1 - p^{-c} - 2e^{-\tilde{c}n} - 2e^{-\tilde{\tilde{c}}m}$.

Proof. First, note that, $P(\mathscr{A}^c) \le \underbrace{P\left(||\frac{1}{n}\epsilon' X_T||_2 \ge \lambda\frac{||w_T||_2}{2}\right)}_{I} + \underbrace{P\left(\max_{j\in T^c}|\frac{1}{n}\epsilon' X_j w_j^{-1}| \ge \frac{\lambda}{2}\right)}_{II}$.

Now, with $\tilde{\epsilon} \sim N(0, I_n)$,

$$
\begin{aligned}
I &\leq P\left(||\frac{1}{n}\tilde{\epsilon}' X_T||_2 \geq \tau\frac{m}{n}\right) \\
&\leq p\left(|||\frac{1}{n}\tilde{\epsilon}' X_T||_2 - ||\frac{1}{n}X_T||_F| > (\tau - c_1)\sqrt{\frac{m}{n}}, ||\frac{1}{n}X_T||_2| \leq c_2\frac{1}{\sqrt{n}}\right) \\
&+ P\left(||\frac{1}{n}X_T||_F > c_1\sqrt{\frac{m}{n}}\right) + P\left(\sqrt{\frac{m}{n}} > c_2\frac{1}{\sqrt{n}}\right)
\end{aligned}
$$

By theorem 6.3.2 in Vershynin (2018), we get,

$$
\begin{aligned}
I &\leq 2exp\left[-c - 3c_2^{-2}(\tau - c_1)^2 m\right] + P\left(||\frac{1}{n}X_T||_F > c_1\sqrt{\frac{m}{n}}\right) + P\left(||\frac{1}{n}X_T||_2 > c_2\frac{1}{\sqrt{n}}\right) \\
&\leq 2exp\left[-c - 3c_2^{-2}(\tau - c_1)^2 m\right] + 2exp(-\tilde{c}n) + 2exp(-\tilde{c}n)
\end{aligned}
$$

by setting $c_1 = c_2 = \frac{1}{\sqrt{\rho_0}}(2 + c\sqrt{\frac{m}{n}})$ and using theorem 5.19 in Vershynin (2010). Following similar arguments in the standard Lasso analysis,

$$
\begin{aligned}
II &\leq P\left(\max_{j \in T^c} |\frac{1}{n}\epsilon' X_j w_j^{-1}| \geq \tau\sigma\sqrt{\frac{log(p)}{n}}\right) \\
&\leq 2exp(-clog(p)) + P(\max_{j \in T^c} ||\frac{1}{n}X_j||_2^2 > \tau^2) \\
&\leq 2exp(-clog(p)) + 2exp(-cn(\frac{\tau}{\sqrt{\rho_1}} - 1)^2)
\end{aligned}
$$

□

Lemma B.0.3. *Recall* $\mathscr{B} = \left\{\min_{\Delta \in \mathscr{C}} \frac{\Delta' \frac{1}{n} X' X \Delta}{||\Delta||_2^2} \geq \kappa_n\right\}$, *where*

$$
\mathscr{C} = \left\{\Delta \in \mathscr{R}^p : \sum_{j \in T^c} w_j |\Delta_j| \leq 3||w_T||_2^2 ||\Delta_T||_2^2\right\}
$$

Then, $P(\mathscr{B}) \geq 1 - exp(-\frac{||w_T||_2}{w_{min}}\sqrt{log(p)} - \sqrt{m})$

Proof. Setiing $\tilde{X} = X\Sigma^{-\frac{1}{2}}$, and $\tilde{\Delta} = \Sigma^{\frac{1}{2}}\Delta$, we start by observing,

$$\begin{aligned}
P(\mathscr{B}^c) &= P\left(\min_{\Delta\in\mathscr{C},||\Delta||_2=1} \frac{1}{\sqrt{n}}|X\Delta||_2 \le \sqrt{\kappa_n}\right) \\
&= P\left(\min_{\tilde{\Delta}\in\Sigma^{\frac{1}{2}}\mathscr{C},||\Sigma^{-\frac{1}{2}}\tilde{\Delta}||_2=1} |\tilde{X}\tilde{\Delta}||_2 \le \sqrt{\kappa_n}\right) \\
&\le P\left(-\max_{\tilde{\Delta}\in\Sigma^{\frac{1}{2}}\mathscr{C},||\Sigma^{-\frac{1}{2}}\tilde{\Delta}||_2=1} |||\tilde{X}\tilde{\Delta}||_2 - \sqrt{n}||\tilde{\Delta}||_2| + \min_{\Delta\in\mathscr{C},||\Delta||_2=1}\sqrt{n}||\tilde{\Delta}||_2 \le \sqrt{n\kappa_n}\right) \\
&\le P\left(\max_{\tilde{\Delta}\in\Sigma^{\frac{1}{2}}\mathscr{C},||\Sigma^{-\frac{1}{2}}\tilde{\Delta}||_2=1} |||\tilde{X}\tilde{\Delta}||_2 - \sqrt{n}||\tilde{\Delta}||_2| \ge \sqrt{n}(\sqrt{\rho_0} - \sqrt{\kappa_n})\right) \\
&\le 2exp(-n^2)
\end{aligned}$$

by setting $\sqrt{\kappa_n} = \sqrt{\rho_0} - \frac{c}{\sqrt{n}}\left[\sup_{\tilde{\Delta}\in\Sigma^{\frac{1}{2}}\mathscr{C},||\Sigma^{-\frac{1}{2}}\tilde{\Delta}||_2=1} ||\tilde{\Delta}||_2 + E\sup_{\tilde{\Delta}\in\Sigma^{\frac{1}{2}}\mathscr{C},||\Sigma^{-\frac{1}{2}}\tilde{\Delta}||_2=1} |<g,\tilde{\Delta}>|\right]$, where $g \sim N(0, I_p)$.

Now,

1. $\sup_{\tilde{\Delta}\in\Sigma^{\frac{1}{2}}\mathscr{C},||\Sigma^{-\frac{1}{2}}\tilde{\Delta}||_2=1} ||\tilde{\Delta}||_2 \le \sqrt{\rho_1}$.

2. $E\sup_{\tilde{\Delta}\in\Sigma^{\frac{1}{2}}\mathscr{C},||\Sigma^{-\frac{1}{2}}\tilde{\Delta}||_2=1} |<g,\tilde{\Delta}>| = E\sup_{\Delta\in\mathscr{C},||\Delta||_2=1} |<\Sigma^{\frac{1}{2}}g,\Delta>|$

$$\begin{aligned}
&\le E\sup_{\Delta\in\mathscr{C},||\Delta||_2=1} |\sum_{j\in T^c}\frac{\tilde{g}}{w_j}w_j\Delta_j| + E\sup_{\Delta\in\mathscr{C},||\Delta||_2=1} |<\tilde{g}_T,\Delta_T>| \\
&\le E\max_{j\in T^c}|\tilde{g}_j|\frac{1}{w_{min}}3||w_T||_2 + E||\tilde{g}_T||_2 \\
&\le \frac{||w_T||_2}{w_{min}}\sqrt{log(p)} + \sqrt{m}
\end{aligned}$$

This implies as long as $\frac{||w_T||_2}{w_{min}}\sqrt{\frac{log(p)}{n}} + \sqrt{\frac{m}{n}} = o_P(1)$, κ_n can be choosen to be $\frac{1}{2}\rho_0$.

□

Hence, combining lemmas B.0.1-B.0.3, choosing

$$\lambda \ge c\sigma\left[\sqrt{\frac{m}{n}}\frac{1}{||w_T||_2} + \frac{1}{w_{min}}\sqrt{\frac{log(p)}{n}}\right]$$

as long as

$$\frac{||w_T||_2}{w_{min}}\sqrt{\frac{log(p)}{n}}+\sqrt{\frac{m}{n}}=o_P(1),$$

with high probability,

$$||\hat{\beta}^w-\beta^*||_2\leq\frac{3}{2}\frac{\lambda}{\kappa_n}||w_T||_2$$

CHAPTER 3

SCIDNET: ERROR CONTROLLED FEATURE SELECTION FOR ULTRA HIGH DIMENSIONAL AND HIGHLY CORRELATED FEATURE SPACE USING DEEP LEARNING

3.1 Introduction

In modern applications (e.g., genetics and imaging studies), the investigator is often interested in uncovering the true pattern between a quantitative response and a large number of features. The key working assumption, oftentimes, is that there is an underlying sparsity pattern buried in the high dimensional data setting. Selecting the essential features aids in further scientific investigations by offering improved interpretability and explainability, reduced computational cost for prediction and estimation, and less memory usage due to lower dimensional manifolds of the feature space being estimated. Under the linear model (LM) framework, this problem has been extensively studied over the past few decades producing popular algorithms such as Lasso, Elastic net, SCAD, and MCP. A detailed review of this literature can be found in Fan and Lv (2010) and thus is omitted here. However, regardless of their ubiquitous applications, the LM has limited usage, especially when the underlying mechanism is highly nonlinear, with potential interaction effects. Relaxing the linearity assumption, the Artificial Neural Network (ANN) models are well known for efficiently approximating complicated functions. From an information-theoretic viewpoint, Elbrächter et al. (2021) established that deep neural networks (DNN) provide an optimal approximation of a nonlinear function, covering a wide range of functional classes used in signal processing. This property has promoted the use of Deep Learning (DL) models for feature selection, an approach that has generated much research interest over the past few years. A major caveat, however, is that the DL models are often used as a black box in many applications. Following the intriguing arguments in Rudin (2019), caution must be exercised regarding the application of DL models for decision-making in real-world problems. Employing only the relevant predictors to construct a predictive model is the right step toward explainable machine learning. However, as suggested in Ghorbani et al. (2019), oftentimes, the feature importance in DL-based algorithms varied drastically under small perturbations in the input or in the presence of added noise.

As a solution to this problem, we focus on the reproducible nonlinear variable selection using DL models with some error control. We adopt the False Discovery Rate (FDR) first proposed by Benjamini and Hochberg (1995), known for being suitable for large-scale multiple testing problems. To formally define the FDR, we consider the random variable, FDP, representing the False Discovery Proportion: $FDP = \frac{e_0}{N_+ \wedge 1}$, where e_0= number of falsely selected variables, N_+= number of total discoveries. Then, FDR is defined as $FDR = E(FDP)$. Estimating this expectation poses a unique challenge for the model-free variable selection problem, which many authors have tried to solve from various perspectives. For example, a p-value approach has been proposed as a feature importance criterion in multiple testing literature; see Tansey et al. (2018); Xia et al. (2017); Li and Barber (2019); Lei and Fithian (2018) for a more detailed overview. However, for DL models, generating interpretable p-values is still an unrevealed research problem. To circumvent this limitation, the knockoff framework has been proposed by Candès et al. (2018). Essentially, this is a model-free variable selection algorithm with provable FDR control, assuming one has prior knowledge of the predictors' distribution. Lu et al. (2018) further proposed the *DeepPINK* algorithm by integrating the knockoff framework with the DL architecture for improved explainability of the DL models. However, in real-world applications, the predictor's distribution needs to be estimated to generate the knockoff variables, which adds another layer of uncertainty to the analysis. Recently Barber et al. (2020) showed that the knockoff framework might yield inflation in false discoveries, consistent with the error incurred in estimating the predictor's distribution. This problem is exacerbated by highly correlated features. An empirical illustration is provided in Appendix 4.2.2, showing how model-X knockoff (Candès et al., 2018) typically fails to control FDR under a simplistic setting with high multicollinearity. In some cases, it may be possible to have prior knowledge of the correlation pattern among the features. For example, in genetics studies, there is a common notion of linkage disequilibrium, which helps to specify the dependency pattern among the alleles at polymorphisms (Sesia et al. (2018)). However, this information is typically unavailable in many other domain sciences.

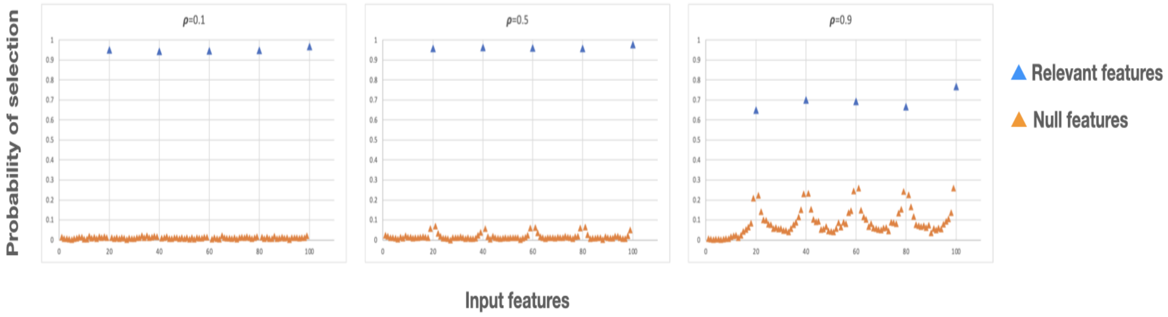

Figure 3.1 How multicollinearity affects feature selection - A demonstration using simplistic simulation setting: We simulate $n = 400$ iid copies of $(y \in \mathscr{R}, X \in \mathscr{R}^{100})$, where the outcome y is generated from a linear model: $y = 0.5(X_{20} + X_{40} + X_{60} + X_{80} + X_{100}) + \epsilon, \epsilon \sim N(0,1)$ and the features $X \sim N_{100}(0, \Sigma), (\Sigma)_{ij} = \rho^{|i-j|}$. We implement the Lasso algorithm for 500 Monte Carlo replication of the data, and the y-axis shows the proportion of time each feature is selected out of 500 replications. For a higher autocorrelation ρ, the selection probability of the true features was significantly reduced whereas the null features associated with true features got selected more frequently.

Hence any model-specific knockoff generation (Candès et al., 2018; Sesia et al., 2018) would be inefficient in those contexts. Recently, DL-based flexible knockoff generating algorithms have been proposed (Liu and Zheng, 2019; Jordon et al., 2019; Romano et al., 2020); however they are trained in a typical big-n-small-p setting, and it is unclear how they will perform when the sample size n is significantly smaller than the dimension of the covariates p, and the predictors are highly correlated. We next discuss in detail the multicollinearity issue.

In many modern high-dimensional datasets arising in genetics and imaging studies, the other challenge is extreme multicollinearity - the predictors are typically correlated among themselves in a complex manner, often with pairwise sample correlations exceeding 0.99. In a simplistic setting of a linear model, Figure 3.1 shows how increased autocorrelation typically reduces the selection probability of the true non-null features. Because extremely correlated features become almost indistinguishable, it would be unrealistic to claim that a particular feature from a cluster is associated with the outcome in a regression setup. Hence, accounting for the uncertainty, it would be pragmatic to aim for group-level variable selection and claim that at least one variable from a densely correlated group is important for the outcome. This approach is not entirely new; as beautifully argued in Brzyski et al. (2017) that in genetics, the discovery of a specific genomic region is treated equivalently as a particular variant-wise discovery in that location. In this context, the term '*true discovery*' implies that the selected cluster can serve as a good proxy for at least one element in the true index set of significant features. However, a complication of this approach is that the notion of FDR becomes non-trivial. For this reason, following Siegmund et al. (2011), we adopt the cluster version of the FDR as the *expected value of the "proportion of clusters that are falsely declared among all declared clusters"*. We denote this as *cFDR* henceforth. Looking at the extreme multicollinearity problem from a slightly different angle, several algorithms have been proposed in the hierarchical testing literature including *CAVIAR* (Hormozdiari et al., 2014), *SUSIE* (Wang et al., 2020), *KnockoffZoom* (Sesia et al., 2019). While the knockoff-based procedures have the limitation of generating knockoffs from an unknown distribution with a very small sample size, other methods lack applicability in non-linear-nonparametric setups as they typically depend on p-values.

Our contribution To address the complications mentioned above in variable selection and unexplored gap while applying DL, we propose **SciDNet- Screening & Cleaning Incorporated Deep Neural Network** - a novel method for the reproducible high-dimensional nonlinear-nonparametric feature selection with highly correlated predictors. The screening step is a dimension reduction step. We screen out most of the null features and create a set of multi-resolution clusters that collectively contain all the proxy variables needed to cover the truly significant features with high probability. In the cleaning step, using a properly tuned DL model under an appropriate resampling scheme, an estimator of the FDR is proposed. Finally, we select some clusters of highly correlated predictors by controlling the estimated FDR. To this end, 'FDR observed for SciDNet' would implicitly mean the value of the cFDR discussed above. Our major contributions can be summarized below:

- The proposed method SciDNet is based on a combination of techniques from statistical machine learning on sparse modeling. We introduce a resampling-based FDR estimation scheme, which allows us to identify the most relevant features while discarding irrelevant ones in an FDR-controlled setting. Additionally, the proposed algorithm is specifically tailored for highly correlated features, which is proved to be problematic for traditional feature selection methods.

- The proposed approach relies on minimal modeling assumptions and is entirely free from p-value, unlike existing state-of-the-art methods, providing a better understanding of the sparse relationship between the outcome and the high-dimensional predictors. Our theoretical study consolidates the empirical results by showing SciDNet's provable FDR control guarantee in an asymptotic setting.
- To the best of our knowledge, in a high-dimensional setting, no other method in the literature accommodates the multicollinearity issue via data adaptive cluster formation, followed by a nonlinear-nonparametric error-controlled feature selection integrated with DL. The results from our extensive simulations and real data analyses demonstrate the proposed method's validity in general as a proof of concept by achieving higher power, controlled FDR, and higher prediction accuracy. Additionally, we conducted an ablation study which provides a systematic analysis of the contribution of each individual step to the overall performance of the feature selection method.

Overall, our contributions offer a powerful tool for researchers and practitioners who face the challenge of selecting relevant variables from highly correlated ultrahigh dimensional datasets applicable in a variety of fields, from biology to finance to social sciences. For the rest of the article, in Section 3.2, we describe the proposed screening and cleaning method, followed by an extensive simulation study in Section 4.2 and an analysis of two real-world gene expression datasets in Section 3.4. Finally, Section 4.4 concludes with a summary and future directions.

3.2 The Algorithm

3.2.1 Notation and assumptions

Under the supervised learning framework, let Y denote a continuous response variable, and $X = (X_1, \ldots, X_p)$ denote p continuous covariates. Let $F_y(\cdot)$ denote the CDF of the response variable Y, and let $F_k(\cdot)$ denote the CDF of the predictor X_k. Assuming a sample size n, we consider the ultrahigh-dimensional setting where $p = O(exp(n^{\tau})), \tau > 0$. We assume no specific functional relationship between the outcome Y and the predictors X but we impose a high-level assumption on the distribution of X. In the spirit of Liu et al. (2009), we assume that the predictors follow the nonparanormal distribution; i.e., there exist unknown differentiable functions $f(X) = \{f_j(X_j), j \in \{1,2,\ldots,p\}\}$, such that $f(X) \sim N(\mu^{p\times 1}, \Sigma^{p\times p})$. This nonparanormal distribution covers a wide range of parametric families of distributions and its main beauty lies in the fact that $f(X)$ preserves the conditional dependency structure of the original variables X. Maintaining the sparsity condition, we may assume that there exists a subset $S_0 \subset \{1,2,\ldots,p\}, |S_0| = O(1)$, such that, conditional on features in S_0, the response Y is independent of features in S_0^c. In other words, $S_0 = \{k : f(y|X)$ depends on $X_k\}$, where $f(y|X)$ is the conditional density of y given X. Our goal is to learn the sparsity structure by estimating S_0.

3.2.2 Screening Step

Under the assumption that the cardinality of S_0 is much smaller than the feature space dimension p, most of the features belong to S_0^c. Hence in the screening step, we focus primarily on finding an active set $\hat{S}_n$ with $|\hat{S}_n| << p$ such that $P(S_0 \subset \hat{S}_n) \to 1$ as $n \to \infty$. This property is called the sure screening property (Fan and Lv, 2008), which ensures that all the significant predictors are still retained in $\hat{S}_n$ and the other predictors $\{X_j, j \in \hat{S}_n^c\}$ are henceforth eliminated from the remaining analysis. As these active variables are highly correlated among themselves, in the second step we further cluster them by exploiting the conditional dependency structure.

3.2.2.1 Finding the active set of variables

To find the active set, we first consider the nonparanormal transformation on (Y, X) and then perform the Henze–Zirkler's (HZ) test on the transformed variable. While the first transforms all the variables to a joint Gaussian variable maintaining their conditional covariance structure; the second test confirms, by pairwise testing, if there is significant dependence in the transformed response and predictors. This workflow has been proposed by Liu et al. (2009),Henze and Zirkler (1990), Xue and Liang (2017). The strategy proceeds as follows:

1. **nonparanormal transformation**: We first consider the following transformation: $T_y(Y) = \Phi^{-1}(F_y(Y)), T_k(X_k) = \Phi^{-1}(F_k(X_k)), k = 1, 2, \dots, p$, where $\Phi(\cdot)$ denotes the CDF of the standard Gaussian distribution. However, in practice, the cdf of Y and X_k are unknown, we can estimate it by the truncated empirical cdf as suggested by Liu et al. (2009). Henceforth, let $(\tilde{T}_y(Y), \tilde{T}_k(X_k))$ denote the corresponding transformations.

2. **HZ test**: By the basic properties of CDF, it is easy to see that $(T_y(Y), T_k(X_k))$ will jointly follow a bivariate Gaussian distribution $N_2(0, I_2)$ if and only if Y is independent of X_k. This can be tested using HZ test, Henze and Zirkler (1990), where the test statistic for the predictor X_k can be expressed as $w_k = \int_{\mathscr{R}^2} |\psi_k(t) - exp(-\frac{1}{2}t't)|^2 \phi_\beta(t) dt, k = 1, 2, \dots, p$; where $\psi_k(t)$ is the characteristic function of $(T_y(Y), T_k(X_k))$ and $exp(-\frac{1}{2}t't)$ represents the characteristic function of $N_2(0, I_2)$. It typically measures the disparity between the joint distribution of $(T_y(Y), T_k(X_k))$ and $N_2(0, I_2)$ and is expected to be typically high for the non-null predictors $X_j, j \in S_0$ indicating significant evidence against the independence of the transformed variable $(T_y(Y), T_k(X_k))$.

   Next, as in practice, we proceed with $(\tilde{T}_y(Y), \tilde{T}_k(X_k))$, we calculate the the HZ test statistic as

   $$\tilde{w}_k^* = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n e^{-\frac{\beta^2}{2} d_{ij}} - \frac{2}{n(1+\beta^2)} \sum_{i=1}^n e^{-\frac{\beta^2}{2(1+\beta^2)} d_i} + \frac{1}{1+2\beta^2} \tag{3.1}$$

   where $d_{ij} = (\tilde{T}_k(x_{ki}) - \tilde{T}_k(x_{kj}))^2 + (\tilde{T}_Y(y_i) - \tilde{T}_Y(y_j))^2$ and $d_i = \tilde{T}_k^2(x_{ki}) + \tilde{T}_Y^2(y_i)$. Consistent with the existing literature, we choose the value of the smoothing parameter β as $\frac{(1.25n)^{1/6}}{\sqrt{2}}$, which corresponds to the optimal bandwidth for a nonparametric kernel density estimator with Gaussian kernel (Henze and Zirkler (1990)). The observed test statistics $\tilde{w}_k^*$ converge to w_k as shown in Xue and Liang (2017).

3. Next, we select the active set of predictors $\hat{S}_n$ according to the larger values of $\tilde{w}_k^*$, i.e., $\hat{S}_n = \{1 \le k \le p : \tilde{w}_k^* > cn^{-\kappa}\}$ where where c and κ are predetermined threshold values.

This active set $\hat{S}_n$ contains all the predictors significantly correlated with the response marginally. Under very mild regularity conditions on the signal strength of the nonnull predictors where

$\min_{k\in S_0} w_k \geq 2cn^{-\kappa}$ with c as a constant and $0 \leq \kappa \leq \frac{1}{4}$, the screening process enjoys the advantage of *sure screening property*, i.e., $P(S_0 \subset \hat{S}_n) \to 1$, as $n \to \infty$. More details on the theoretical guarantee can be found in Xue and Liang (2017). A common practice is to set the active set size $|\hat{S}_n|$ at $\nu_n = [n/log(n)]$. However, as we further cluster the active variables in the next step, our proposed method is fairly robust in terms of the $|\hat{S}_n|$ as long as we retain most of the significant variables. We propose to select a bigger active set with a size proportional to ν_n.

3.2.2.2 Clustering the active predictors using the precision matrix

Algorithm 3.1 Finding clusters and the representatives

```
Input :($X \in \mathscr{R}^{n\times p}, Y \in \mathscr{R}^n$), The Active set $\hat{S}_n$ , $|\hat{S}_n| = p_1 < p$
Estimate the precision matrix: $\hat{\Sigma}^{-1} = (\hat{\sigma}^{ij})_{i,j\in\{1,2,...,p\}}$ using Nodewise Lasso
Define the clusters $C_i = \{j \in \hat{S}_n : \hat{\sigma}^{ij} \neq 0\}, i \in \hat{S}_n$
for $1 \leq i \leq p_1$ do
    for $1 \leq j \leq p_1, j \neq i$ do
        Define $\Omega^{ij} = \left\{corr(X_{C_i}, X_{C_j})\right\} = \{\rho(X_l, X_{l'}), l \in C_i, l' \in C_j\} \in \mathscr{R}^{|C_i||C_j|}$
        if $\max\{\Omega^{ij}\} \geq r$ then
            $C_i = C_i \cup C_j$
            $C_j = \phi$
        end
    end
end
Retain only the non-null clusters: $C = \{C_i : C_i \neq \phi, i \in \hat{S}_n\}$
Find the cluster representatives $\tilde{S}_n = \{R_j, 1 \leq j \leq |C| : R_j = \underset{l\in C_j}{argmax}\{\tilde{w}^*_l\}\}$
Output :Clusters $C_1, C_2, ..., C_{|C|}$ and corresponding cluster representatives $\tilde{S}_n = \{R_j, 1 \leq j \leq |C|\}$
```

As the Henze–Zirkler test focuses on the (pairwise) marginal correlation among the predictors and response, it typically includes the null predictors with strong associations with a significant predictor; thus they are highly correlated among themselves. Hence to reduce the high correlation in the active set, our strategy is to exploit their conditional dependency structure and divide the active variables $\{X_j : j \in \hat{S}_n\}$ into $p_c(<< p)$ non-overlapping clusters: $C_1, C_2, \ldots, C_{p_c}$. By sure screening property, with asymptotically high probability, $S_0 \subset \bigcup_{j=1}^{p_c} C_j$. The use of a sparse precision matrix to understand the dependence structure in a high-dimensional feature space has been well acknowledged in statistics literature (e.g., Lauritzen (1996), Shojaie and Michailidis

(2010)) due to its scalability. In some contexts, it brings more insight compared to the analysis of a simple covariance matrix. For example, in the human brain, two separate regions can be highly correlated with no direct relation and only due to their strong interaction with a common third region. So, understanding the conditional dependence structure and using it in clustering the brain regions is more informative in the context of understanding the functional connectivity in the human brain (Das et al., 2017). Otherwise, simple correlation-based clustering will result in huge cluster sizes with less interpretable groups of brain regions.

To this end, in order to estimate the precision matrix we implement the nodewise Lasso algorithm (Van de Geer et al., 2014) on the transformed variables $(\tilde{T}_y(Y), \tilde{T}_k(X_k)), k \in \hat{S}_n$. Nodewise Lasso regression is generally entertained to estimate a sparse precision matrix in the context of the Gaussian graphical model by performing simultaneous Lasso regression on each predictor. The tuning parameters in each nodewise Lasso are typically selected using cross-validation. More details on this algorithm and its theoretical guarantees can be found in (Meinshausen and Bühlmann, 2006). Let $\hat{\Sigma}^{-1}$ be the estimated precision matrix by the nodewise Lasso algorithm and $\rho(Z, Z')$ denotes any correlation metric for two random variables Z and Z', e.g. Pearson's correlation. Algorithm 3.1 summarizes the clustering step.

Here not only we are clustering the active predictors, but also selecting an appropriate representative from each cluster. First, for each active predictor $X_i \in \hat{S}_n$, we collect all the other active predictors conditionally dependent on X_i, and make cluster C_i. Although clustering using conditional dependence produces smaller clusters, there might be some overlaps owing to the complex association in the original predictor space. Hence, to reduce the excessive intercluster correlation, we merge all those clusters having a maximum correlation greater than some prespecified threshold r (we typically set r=0.9). Next, each cluster is updated by adding all the other features conditionally dependent on the existing cluster members. Finally, to find the appropriate cluster representatives, we focus on the HZ-test statistic $\tilde{w}_k^*$ in 3.1 which measures the extent of resemblance between the distribution of each nonparanormally transformed variable in a cluster and the null distribution $N(0, I_2)$. So, for cluster C_i, we select the variable $R_i = argmax_{j \in C_i}\{\tilde{w}_j^*\}$

indicating its strongest association with the response variable compared to the other predictors in the cluster.

3.2.3 Cleaning with Deep Neural Network (DNN)

We start the cleaning step by modeling the response Y and the cluster representatives $X_{\tilde{S}_n}$ obtained through 3.2.2.2. In order to perform the error-controlled variable selection, each representative will be assigned an importance score followed by a resampling algorithm to finally control the FDR.

While it is possible to adopt any other generic sparsity-inducing DNN procedure, here we focus on the LassoNet algorithm recently proposed by Lemhadri et al. (2021) for its elegant mathematical frameworks which naturally sets the stage for nonlinear feature selection. To approximate the unknown functional connection, it considers the class of all fully connected feed-forward residual neural networks; namely, $\mathscr{F} = \{f \equiv f_{\theta,W} : x \mapsto \theta^T x + h_W(x)\}$. Here, W denotes the network parameters, K denotes the size of the first hidden layer, $W^{(0)} \in \mathscr{R}^{p\times K}$ denotes the first hidden layer parameters, $\theta \in \mathscr{R}^p$ denotes the residual layer's weights. In order to minimize the reconstruction error: the LassoNet objective function can be formulated as:

$$\min_{\theta,W} L(\theta, W) + \lambda||\theta||_1 \text{ subject to } ||W_j^{(0)}||_\infty \le M|\theta_j|, j = 1, 2, \ldots, p \tag{3.2}$$

With $L(\theta, W) = \frac{1}{n}\sum_{i=1}^n l(f_{\theta,W}(x_i), y_i)$ as the empirical loss on the training data and x_i as the vector of cluster representatives observed for the i^{th} individual. While the main feature sparsity is induced by the L_1 norm on residual layer parameter θ, the second constraint controls the total amount of nonlinearity of the predictors. As mentioned in Lemhadri et al. (2021), LassoNet can be argued as an extension of the celebrated Lasso algorithm to nonlinear variable selection.

In L_1 penalization framework, the importance of a specific feature is naturally embedded into the highest penalization level up to which it can survive in the model. So, to measure the importance of each representative, the LassoNet algorithm is executed over a long range of tuning parameter $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_r$ on $(Y, X_{\tilde{S}_n})$. In practice, a small value is fixed for λ_1 where all the variables are present in the model. Then we gradually increase the value of the tuning parameter and stop at λ_r, where no variables are present in the model. Next, the importance score for the j-th cluster is defined as $\hat{\lambda}_j$= maximum value of λ up to which the j-th representative exists in the model, and then the following rank statistic is computed: $\mathscr{I}_j = \sum_{j' \neq j} 1\left(\hat{\lambda}_j \leq \hat{\lambda}_{j'}\right)$ for $j = 1, 2, \ldots, C$. A lower $\mathscr{I}_j$ means that the j-th cluster representative stays in the model up to a higher value λ implying its high potential as a significant cluster. In contrast, a higher $\mathscr{I}_j$ indicates the corresponding cluster leaves the model even for a smaller value of λ as a consequence of being simply a collection of null features. Hence, we should only focus on the clusters with lower ranks. Additionally, in order to control the FDR, understanding the behavior of the predictors under the null distribution is important. In traditional FDR controlling algorithms, this is typically done by generating the p-values. Here, as a p-value-free algorithm, we propose the following resampling-based approach:

1. Generate B bootstrap versions of the data $\left\{Y^b, X^b_{\tilde{S}_n}\right\}_{b=1}^{B}$ considering only the cluster representatives $\tilde{S}_n$. For each bootstrap version, run the LassoNet algorithm parallelly, and calculate the importance of each representative by measuring $\hat{\lambda}^b_j$= maximum value of λ up to which the j-th predictor exists in the model for b-th bootstrap version, and then the ranks $\mathscr{I}^b_j = \sum_{j' \neq j} 1\left(\hat{\lambda}^b_j \leq \hat{\lambda}^b_{j'}\right)$.

   Therefore, the averaged rank is: $\bar{\mathscr{I}}_j = \frac{1}{B} \sum_{b=1}^{B} \mathscr{I}^b_j$

2. For an arbitrary threshold δ, we would select the cluster representatives with averaged rank $\bar{\mathscr{I}}_j$ lower than δ; so we define, $N_+(\delta) = \sum_{j \in \tilde{S}_n} 1(\bar{\mathscr{I}}_j \leq \delta)$ representing the number of selected clusters with respect to the cutoff δ.

3. Next, to estimate the expected number of falsely discovered clusters, define $\mathscr{R}^b = \{j : \mathscr{I}^b_j \leq \delta\}$, the number of cluster representatives with higher importance score so that the corresponding rank is lower than the cutoff δ in the b-th bootstrap version. Additionally, define a neighbourhood $\mathscr{N}(\bar{\mathscr{I}}_j, \kappa) = \{l \in \{1, 2, \ldots, C\} : \bar{\mathscr{I}}_j - l \leq \kappa\}$, for some specific small number κ.

4. Further, we estimate the number of falsely discovered clusters and hence an estimator of the FDR can be constructed as $F\hat{D}R(\delta) = \frac{\hat{e}_0(\delta)}{N_+(\delta)}$ where,

$$\hat{e}_0(\delta) = \tfrac{2}{B}\textstyle\sum_{b=1}^{B}\left\{\sum_{j\in\mathscr{R}^b} 1(\mathscr{I}_j^b \not\in \mathscr{N}(\bar{\mathscr{I}}_j, \kappa))\right\} \tag{3.3}$$

5. The $F\hat{D}R$ is sequentially estimated with $\delta = \bar{\mathscr{I}}_{(1)}, \bar{\mathscr{I}}_{(2)}, \ldots, \bar{\mathscr{I}}_{(C)}$ and the optimum threshold is $\Delta^* = \max\{\delta > 0 : F\hat{D}R(\delta) < q\}$ for some pre-specific FDR control level q. The final selected set of clusters with controlled FDR is given by $\hat{D}_n = \{C_j, j = 1, 2, \ldots, C : \bar{\mathscr{I}}_j \leq \Delta^*\}$

The proposed method certainly has a close resemblance with an FDR-controlling approach. The notion of false discovery is incorporated into the algorithm via the resampling: if a null predictor gets a relatively higher importance score, that is most possibly due to that specific bootstrap version which creates the spurious relation, however, that would not be consistent for all the other bootstraps in general. On the other hand, all the bootstrap versions should consistently produce higher importance scores for the significant predictors. As a consequence, the variability in the ranks of the importance scores will be much higher for the null predictors compared to their nonnull counterparts. This notion was introduced in the statistics literature in the last twenty years as bagging methods (Breiman, 1996; Bühlmann and Yu, 2002) for reducing the variance of a black-box prediction. The proposed method utilizes this phase transition in the feature selection framework to effectively identify the false discoveries; an empirical illustration of which is provided below. The theoretical investigation is relegated to Appendix A, where in the spirit of Ng and Newton (2022), we use the idea of random-weighted Group Lasso penalization to mimic the resampling setup.

How to choose the hyperparameter κ: Choosing an appropriate value of κ has a significant effect on the performance of SciDNet. A higher value of κ might lead to weaker control over the inclusion of false discoveries, whereas choosing a small κ will create tighter error control resulting in reduced power. However, we propose an effective way to tune the κ with the assistance of phase transition in the ranks of the importance score $\bar{\mathscr{I}}_j$ of the cluster representatives. For an

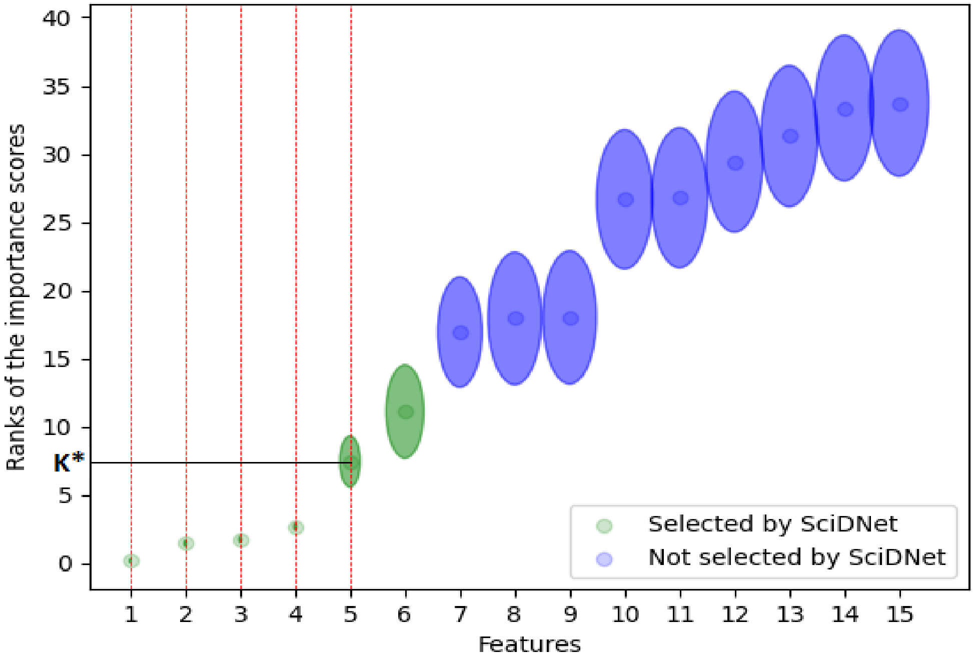

Figure 3.2 Illustration of phase transition using synthetic data: Features selected by SciDNet at $q = 0.2$ clearly have lower bootstrap variability compared to the other irrelevant features.

illustration, in Figure 3.2 we consider a single index model (see section 4.2 for more details of the data-generating mechanism), and the features with top 15 importance scores are shown along the x-axis. The first 5 representative features are the only relevant predictors (indicated by the vertical dotted red line). Along the y-axis, the center of the ellipse for each feature represents the rank of the importance scores $\bar{\mathscr{I}}_j$ averaged over 50 bootstrap replications and the area of each ellipse represents the bootstrap variability around the averaged score. One would observe a clear phase transition in the bootstrap distribution of the ranks. For the significant features, the ranks are lower with extremely precise estimates. On the other hand, for the rest of the null features, the averaged ranks possess much higher values coupled with huge variability producing bigger ellipses. Hence, for a compact neighborhood $\mathcal{N}(\bar{\mathscr{I}}_j, \kappa)$ to capture only the small variability in the bootstrap ranks of the significant features, we simply fix κ= **K*** (in figure 3.2), the phase transition point for the averaged rank. This phase transition property is further illustrated on real data in Appendix B.0.2.

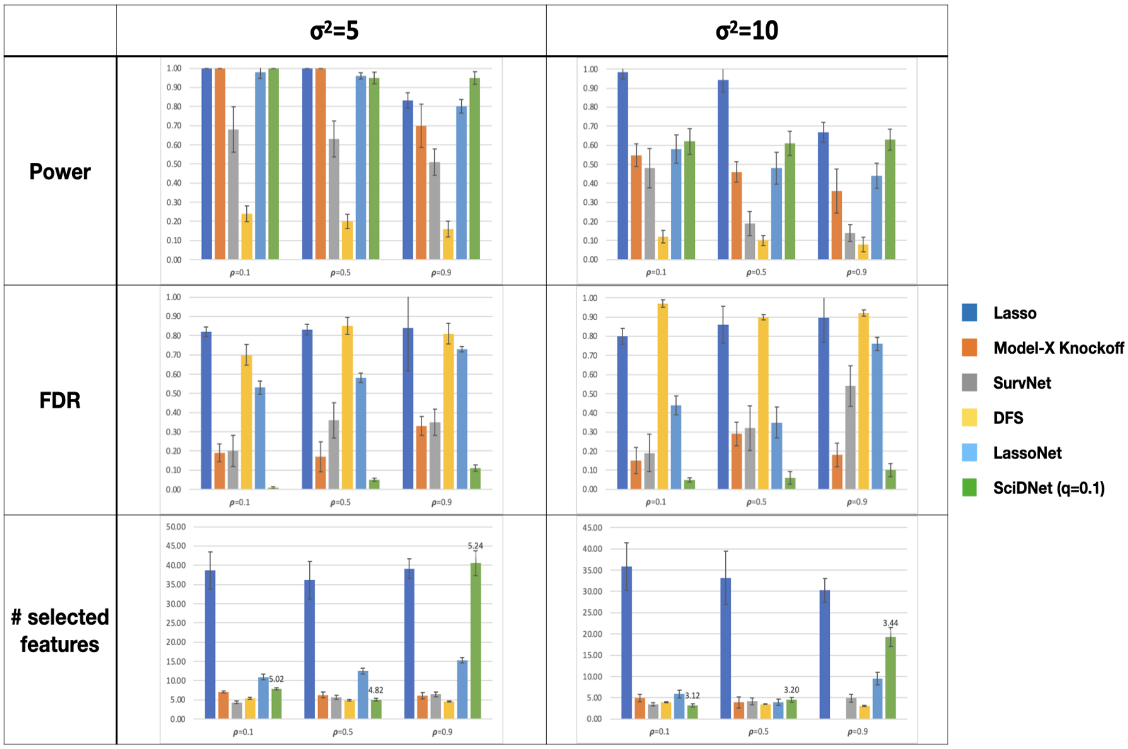

Figure 3.3 Illustration of the effect of multicollinearity on the feature selection methods.

3.3 Numerical Illustrations

In this section, we investigate the finite-sample performance of SciDNet using a wide spectrum of simulation scenarios. We compare SciDNet to several baseline methods which are widely used in practice. We conduct this simulation study on synthetic data generated from a high-dimensional regression problem.

Data Generation: We first consider the single index model for the data-generating mechanism, which is a straightforward yet flexible example of nonlinear models. Here the response is related to a linear combination of the features through an unknown nonlinear, monotonic link function, i.e., $y = g(x'\beta) + \epsilon$. We choose the following link functions: $g(x) = \frac{x^3}{10} + 3\frac{x}{10}$.

We set $n = 400$ and $p = 5000$. The coefficients $\beta \in \mathscr{R}^p$ is sparse with the true nonzero locations $S_0 = \{50, 150, 250, 350, 450\}$, $s = |S_0| = 5$, where $\beta_{S_0^c} = 0, \beta_{S_0} \sim N_5(u\beta_0, 0.1I^{5\times 5})$ with $u = \{\pm 1\}^5$. The value of β_0 is set at $\beta_0 = 2, 4$ to incorporate varying signal strength. The random error $\epsilon \sim N(0, \sigma^2)$, with three increasing noise level as $\sigma^2 = 1, 5, 10$. The high dimensional predictors are generated from $X \sim N_p(0, \Sigma)$ where the covariance matrix Σ is chosen as a Toeplitz matrix with $\Sigma_{ij} = \rho^{|i-j|}$.

To check the effect of multicollinearity, we consider three different settings: $\rho = 0.1, 0.5, 0.95$.

Evaluation Metrics: Let $\hat{D}_n$ denote the selected set of features by some algorithm, then we use the following three metrics to evaluate the performance of these feature selection algorithms:

1. Power= $\frac{|\hat{D}_n \cap S_0|}{|S_0|}$, the proportion of relevant features that are correctly identified

2. Empirical FDR = $\frac{|\hat{D}_n \cap S_0^c|}{|\hat{D}_n|}$, the proportion of falsely identified features among all the identified features

3. $n\_var = |\hat{D}_n|$, the number of total features selected and $n\_clust = |\tilde{S}_n|$, the number of clusters selected by a group feature selection method like SciDnet.

The whole experiment is repeated over 100 Monte Carlo replications and We summarize the results from the empirical evaluations in the following three subsections: (1) The effect of multicollinearity on major existing feature selection methods, (2) the Power vs. FDR balance of SciDNet, and (3) an ablation study showing the effectiveness of all the steps for SciDNet.

3.3.1 The effect of multicollinearity on major existing feature selection methods

we present a simulation study to evaluate the performance of our proposed algorithm in comparison with the following five baseline feature selection methods

1. **Lasso** (Tibshirani, 1996): the L_1 penalized linear regression to prevent overfitting and improve model interpretability.

2. **Model-X knockoff** (Candès et al., 2018): A theoretically guaranteed statistical method for FDR control used in high-dimensional variable selection. It constructs "knockoff" variables that mimic the correlations between the original variables and their relationship with the response variable, allowing for control of the false discovery FDR while identifying important variables.

3. **SurvNet** (Song and Li, 2021): A DNN-based FDR control method for feature selection applicable to high-dimensional large datasets.

4. **Deep Feature Selection (DFS)** (Chen et al., 2021): A novel DNN method for feature selection in a high-dimensional setting with complex nonlinear relationships utilizing L_0 penalization.

5. **LassoNet** (Lemhadri et al., 2021): The nonlinear extension of Lasso. It combines the advantages of both L_1 penalization and neural network structures to identify the important features.

Although several other existing methods exist in the literature for ultra-high dimensional feature selection, we choose these five baseline methods because of their wide applicability and reliable theoretical guarantees. Also, these five methods can be thought of as representative of different classes of algorithms. For example, Model-X knockoff represents the big class of knockoff-based algorithms; whereas SurvNet, DFS, and LassoNet show the effectiveness of the DL-based feature selection methods. Among these five baseline methods, Model-X knockoff and SurvNet are designed to control the FDR and we use $q = 0.2$ as the FDR-control threshold. Also, Lasso, LassoNet, and DFS require proper tuning for their L_1 or L_0 penalty parameters, which can be done via a grid search. For this purpose, we optimize a BIC-type criterion (Chen and Chen, 2008) to tune these hyperparameters, as suggested by the authors of Chen et al. (2021). For the practical implementation of these baseline methods, we used the code/hyperparameters provided by the authors in the respective papers.

Figure 3.3 demonstrates the Power, FDR, and the n_var of all these methods under different correlation strengths and varying noise levels, fixing $\beta_0 = 2$. Also, as SciDNet selects the features as clusters, we show the n_clust in numbers along with n_var, in the third row of Figure 3.3. One would observe that the baseline methods perform poorly as the autocorrelation increases; they result in reduced power and inflated false discoveries. The FDR-controlling methods such as Model-X knockoff and SurvNet fail to control their FDR under the pre-specified threshold $q = 0.2$ for higher autocorrelation. This is somehow expected, as these methods are not tailored for handling such huge multicollinearity. This demonstration empirically motivates as well why we need a feature selection method designed for highly correlated feature space. The DL-based method like LassoNet, DFS, or Survnet additionally suffers from insufficient data under the current big-p-small-n setting. Our additional experiments, presented in the supplementary material, demonstrate how these DL-based models gain better power-FDR balance given sufficient training data and moderate correlation among the features. Compared to these baseline methods, SciDNet successfully maintains its power while controlling the FDR below the pre-specified threshold $q = 0.2$ irrespective of the correlation strength. Also, under moderate multicollinearity, when there is no need for clustering, most of the selected clusters by SciDNet are just singleton sets. On the other hand, under the setting of excessive multicollinearity, the individual features become almost indistinguishable. SciDNet addresses this added uncertainty by selecting larger clusters for higher autocorrelation, as demonstrated in the third row of Figure 3.3.

3.3.2 The Power vs. FDR balance of SciDNet

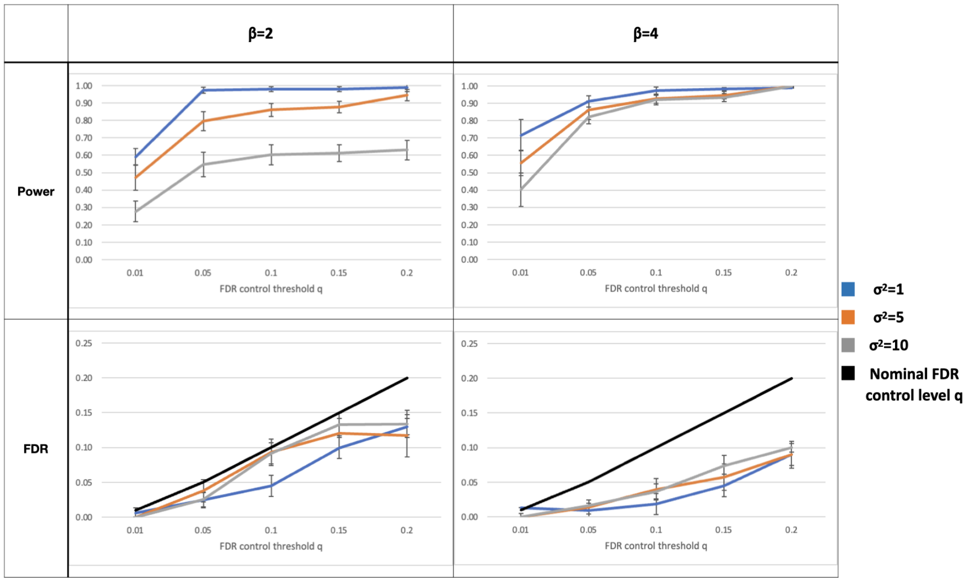

Figure 3.4 Power vs FDR balance for SciDNet.

One major validation for an FDR controlling method is to check how it retrieves its power with respect to gradually increasing the FDR-control threshold q. Figure 3.4 shows this power-FDR trade-off for SciDNet. Also, Section 3.3.1 demonstrates that the baseline methods are not well suited for highly correlated feature space. As SciDNet is specifically designed for this setting,

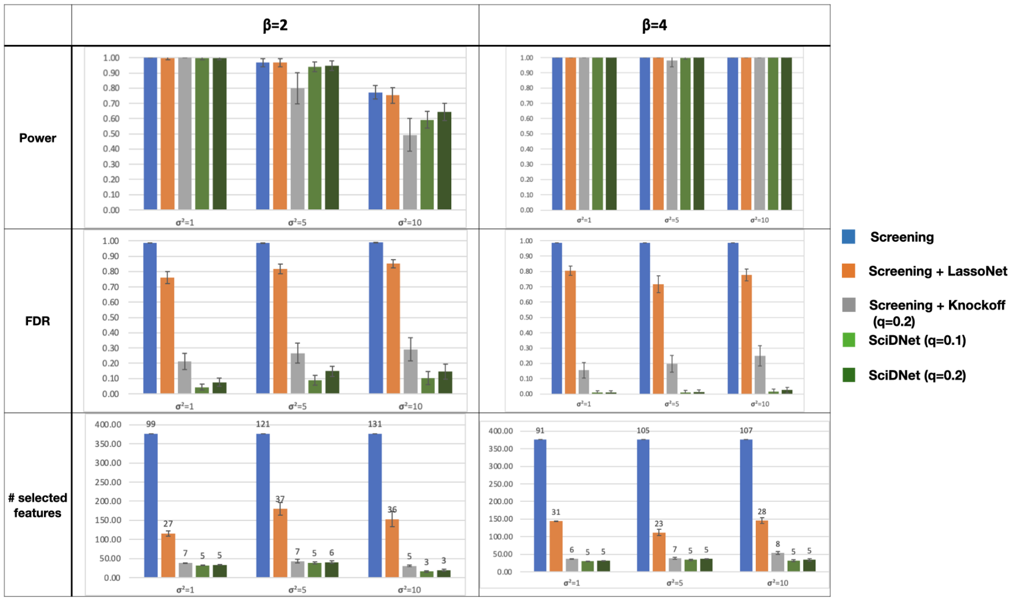

Figure 3.5 Ablation study for SciDNet.

for ease of illustration, we only present the performance of SciDNet setting $\rho = 0.95$. Figure 3.4 illustrates how our method enjoys quick recovery in power when we gradually increase the FDR-controlling threshold q from 0.01 to 0.20 and also maintain the number of false discoveries below the required level. This illustration empirically validates SciDNet as an FDR-controlled feature selection method.

3.3.3 An ablation study

The proposed method SciDNet is a multi-step process. Its screening step reduces the dimension while retaining the main important features and clustering the highly correlated features. Following this, the cleaning step further uses a sparsity-inducing DL model and finally selects some cluster of features by controlling the FDR via resampling. We aim to analyze the impact of these two steps on the overall performance of the model by adding them one by one and observing the change in the evaluation metrics. We also compare our proposed cleaning step (i.e. resampled LassoNet with FDR estimation) with other possible alternatives like using knockoffs for the cleaning.

Hence, for the ablation study, we compare the following four methods under varying signal and noise strength:

1. **Screening only**,
2. **Screening+LassoNet**: Here we consider LassoNet as the cleaning step
3. **Screenning+Knockoff**: Here we consider Model-X knockoff as the cleaning step, with the FDR-controlling threshold $q = 0.2$
4. **Screening+resampled LassoNet**, i.e. **SciDNet**. We use two FDR-controlling thresholds $q = 0.1$, and 0.2.

The power, empirical FDR, n_var are illustrated in Figure 3.5. In addition, the selected number of clusters, i.e., n_clust, are written in numbers in the third row of Figure 3.5. it empirically consolidates several interesting characteristics: (a) Due to the sure-screening property 3.2.2.1, the screening step selects a slightly bigger set of features resulting in high power and high FDR which necessitates further cleaning; (b) All the alternative cleaning steps aim to further eliminate the null features and reduce the FDR while maintaining the power of the screening step; (c) For higher signal and low noise case, SciDNet is comparable to other alternatives. However, for difficult scenarios like low signal and high noise case, which is common in modern genomic and imaging datasets, SciDNet maintains its performance. One would notice that Model-X knockoff loses its FDR control at the nominal level $q = 0.2$ and results in inflated FDR in the presence of high noise. On the other hand, by effectively using the added information from the resampling, SciDNet achieves the best performance (in terms of the power-FDR tradeoff) and continues to preserve its FDR below the nominal level q.

Overall, our ablation study showed that the proposed cleaning following the screening step contributes to the overall performance of SciDNet and that removing any of them leads to a significant drop in accuracy. The screening helps in reducing the dimension of the feature space by removing a large chunk of irrelevant features, which further makes the cleaning step computationally very efficient. Additionally, one would notice the different tasks in the screening and cleaning steps can be easily done in parallel. Specifically, for all the simulation studies SciDNet takes ~ 4.1 minutes to complete. In our experience, this is highly competitive with other methods like DFS, especially when an exhaustive grid search has to be done to optimize the hyperparameters for several baseline methods. We also conducted a further simulation study considering several nonlinear models as data-generating processes with Gaussian and non-Gaussian features. The results are presented in Appendix B.0.3 which further substantiates the above-mentioned results that SciDNet maintains a satisfactory power-FDR balance for various complicated nonlinear models with and without interaction terms. The hyperparameter selection and further implementation details of SciDNet are relegated to Appendix B.0.1 and B.0.5, respectively.

3.4 Real Data Analysis

In addition to the simulation studies, we implemented the proposed algorithm SciDNet in the following two publicly available gene-expression data sets - the CCLE dataset and the riboflavin dataset. We substantiate the findings in two ways: We first provide supporting evidence from the domain research. Additionally, as a more data-aligned validation, we demonstrate that several generic prediction models significantly gain in test accuracy when applied only on the few features selected by SciDNet compared to the prediction result considering the whole feature space. For this purpose, consistent with the other genomic studies, we use the prediction correlation *Corr*(Y_{Pred}, Y_{Test}) in addition to the test MSE as a metric to measure the test performance. To overcome the extra burden of the low sample size and ultrahigh dimensionality in these data sets, we consider 50 independent replications where the data is divided into training and testing maintaining an $8:2$ ratio to get the metrics for the test performance. The final estimate is obtained by averaging all the test MSEs calculated on each of these replications. A similar approach is considered for the correlation metric as well.

3.4.1 Selection of Drug Sensitive Genes using CCLE dataset

A recent large-scale pharmacogenomics study, namely, the cancer cell line encyclopedia (CCLE, link available here), investigated multiple anticancer drugs over hundreds of cell lines. Its main objective is to untangle the response mechanism of anticancer drugs which is critical to precision medicine. The data set consists of dose-response curves for 24 different drugs

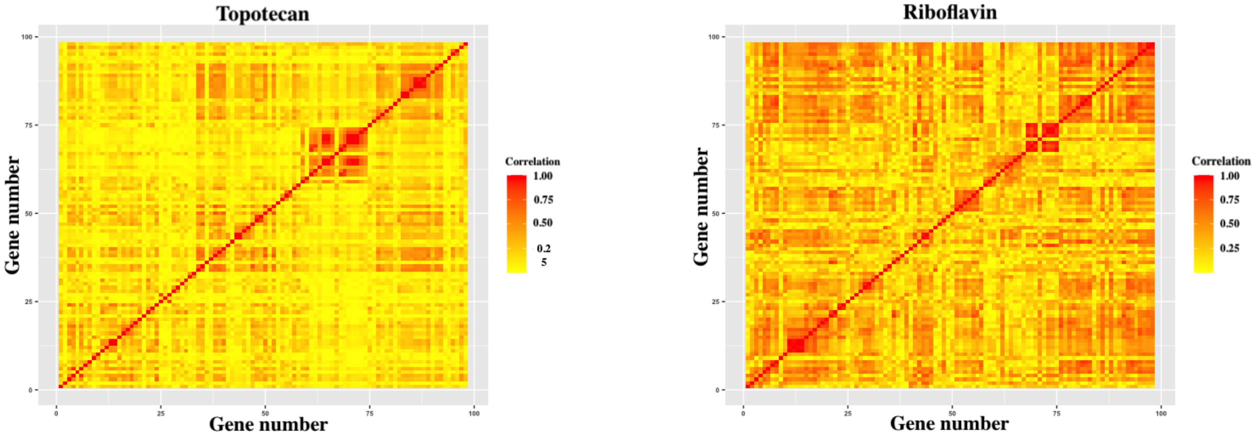

Figure 3.6 A snapshot of correlation strength for first 100 genes considered for the drug Topotecan in CCLE dataset (left) and riboflavin dataset (right).

across over $n = 400$ cell lines. For each cell line, it consists of the expression data of $p = 18,926$ genes, which we consider as features. For the response, we used the activity area (Barretina et al., 2012) to measure the sensitivity of a drug for each cell line. Here we seek to uncover the set of genes associated with the following five specific anticancer drug sensitivity: Topotecan, 17-AAG, Irinotecan, Paclitaxel, and AEW541. These drugs have been used to treat ovarian cancer, lung cancer, and other cancer types. Previous research outputs on these drugs and related gene expression data can be found elsewhere (Barretina et al., 2012).

Table 3.1 Drug-sensitive genes identified by SciDNet and related prediction performance

Drug	# genes (clusters) selected		Test MSE			Corr(Y_{Pred}, Y_{Test})		
	by SciDNet	by LassoNet	LassoNet	SciDNet + MLP	SciDNet + RT	Lassonet	SciDNet + MLP	SciDNet + RT
Topotecan	25 (9)	18469	1.25 (0.21)	1.23 (0.14)	0.81 (0.16)	0.47 (0.11)	0.58 (0.06)	0.69 (0.07)
17-AAG	12 (8)	7152	1.04 (0.16)	1.05 (0.09)	0.83 (0.15)	0.20 (0.16)	0.33 (0.10)	0.49 (0.10)
Irinotecan	18 (7)	17727	0.93 (0.20)	1.09 (0.18)	0.61 (0.13)	0.59 (0.10)	0.63 (0.07)	0.73 (0.08)
Paclitaxel	18 (8)	16437	1.46 (0.33)	1.46 (0.23)	1.11 (0.24)	0.44 (0.14)	0.45 (0.11)	0.59 (0.09)
AEW541	12 (10)	15145	0.33 (0.06)	0.39 (0.09)	0.27 (0.05)	0.30 (0.14)	0.49 (0.10)	0.47 (0.12)

SciDNet produces multi-resolutional clusters of genes for each of the five drugs considered which are interpretable from the domain science perspective. For example, SciDNet discovers *SLFN11* as the top drug-sensitive gene for the drugs Topotecan and Irinotecan. This is consistent with the previous findings as Barretina et al. (2012); Zoppoli et al. (2012) reported the gene *SLFN11* to be highly predictive for both drugs. For another drug 17-AAG, SciDNet discovers the gene *NQO1* as the topmost important gene which is known to be highly sensitive to 17-AAG (Hadley and Hendricks, 2014). The full table containing all the genes selected by SciDNet at q = 15% error-control level and relevant findings from previous genomic studies have been relegated to Appendix B.0.6.

Additionally, as in a real setting, it is difficult to check the performance of a feature selection algorithm, here we present a more data-oriented statistical evaluation for further endorsement of SciDNet's discoveries. From the prediction aspect, one would expect that a prediction model implemented only on a handful of features selected by a successful feature selection algorithm would maintain the similar performance of a model implemented on the whole feature space; in some cases, it might enhance the accuracy. To validate this, we randomly split the whole data into 8:2 for training and testing. First, the SciDNet is implemented in the training part, and then two separate prediction models are used only focusing on the selected features : (1) an MLP - a feed-forward multi-layer perceptron with two hidden layers and (2) bagged regression tree (Breiman, 1996). These two experiments are henceforth called: **"SciDNet+MLP"** and **"SciDNet+RT"**. Next, the test data is used to check the out-of-sample prediction accuracy. Furthermore, similar to the simulation study in section 4.2, we separately implement the LassoNet on the training data for its simultaneous sparsity-induced prediction-optimal characteristics. The summary of the results is presented in table S4 which indicates several interesting points. First, to get the prediction optimal result, LassoNet fails to capture the sparsity and discovers a huge number of genes. This is somehow expected as most of the prediction-optimal sparse methods tend to select a larger set of features to maintain the prediction quality (Wasserman and Roeder, 2009). On the other hand, SciDNet produces only ~ 10 clusters with an average cluster size ~ 2.5. Even with this huge dimension reduction, the added gain in test MSE and $Corr(Y_{Pred}, Y_{Test})$ further proves that ScidNet successfully retains all the significant predictors. One would further notice, SciDNet+RT achieves the best stable performance which is consistent for all the drugs and our simulation study as well. This experiment demonstrates that a black-box predictive model produces more accurate results when applied on the features selected by SciDNet rather than its implementation

Table 3.2 Prediction performance of SciDNet for Riboflavin production data set

Algorithm	Test MSE	Corr(Y_{Pred}, Y_{Test})
LassoNet	0.83 (0.18)	0.36 (0.30)
SciDNet + LassoNet	0.19 (0.15)	0.89 (0.12)
SciDNet + RT	0.42 (0.16)	0.74 (0.15)
SciDNet + MLP	2.64 (0.79)	0.28 (0.37)

on the whole feature space, indicating SciDNet's potential use in both feature selection and prediction.

3.4.2 Selection of associated genes in Riboflavin production data set

We further implement the SciDNet in the context of riboflavin (vitamin B2) production with bacillus subtilis data, a publicly accessible dataset available in the 'hdi' package in R. Here the continuous response is the logarithm of the riboflavin production rate, observed for $n = 71$ samples along with the logarithm of the expression level of $p = 4088$ genes which are treated here as the predictors. Unlike the previous CCLE data, significant multicollinearity is present in most of the Riboflavin data set, as demonstrated in Figure 3.6. Hence, to determine which genes are important for riboflavin production, SciDNet resulted in finding 9 clusters of a total of 160 correlated genes at the $q = 15\%$ FDR control level, making the average cluster size of ~ 17.78, which is much bigger compared to the previous analysis. SciDNet discovered the gene *YCIC_at* as one of the expressive genes related to riboflavin production which was identified by Bühlmann et al. (2014) as a causal gene in this context. The full list of the selected cluster of genes by SciDNet is relegated to Appendix B.0.8.

The results from the empirical validation of SciDNet's feature selection on the riboflavin dataset are presented in table 3.2. However, for the empirical evaluation, SciDNet+MLP is performing poorly as the inflated cluster dimensions make the input layer of the MLP comparatively large where the number of the training data point is ≈ 57. This necessitates the need for a sparse model here, and we adopt the idea of *Relaxed Lasso*, first proposed by Meinshausen (2007). Here we implement the LassoNet again on the selected features by SciDNet, which certainly improves the prediction accuracy. Consistent with the previous experiments, SciDNet+RT effectively maintains its prediction performance. This further consolidates the need for applying an apt feature selection method before fitting a predictive model for an explainable research outcome.

3.5 Discussion

While the explainable AI is the need of the hour, statistical models coupled with cutting-edge ML techniques have to push forward because of their solid theoretical foundation clipped with principled algorithmic advancement. The proposed method SciDNet efficiently exploits several existing statistics and ML literature tools to circumvent some of the complexities that simple adaptation of current DL-based models fail to address appropriately. The basic intuition and exciting empirical results of SciDNet on simulated and real datasets open avenues for further research. For example, one may be interested in developing a theoretical foundation for this ’screening’ and ’cleaning’ strategy for provable FDR control. It would be worth mentioning that although we used the sure independence screening with HZ-test and LassoNet as the main tools, SciDNet puts forward a more generic framework and can be implemented with any other model-free feature screening method and sparsity-inducing DL algorithms like Feng and Simon (2017). In the screening part, a further methodological extension would consider relaxing the assumption of nonparanormally distributed features for a more flexible approach. Additionally, as the dimensionality is reduced after the screening step, it would be interesting to implement model-free knockoff generating algorithms like Romano et al. (2020) in the cleaning step as further algorithmic development. One limitation is that we mainly focus on the regression setup with the continuous outcome because of the requirements of the HZ sure Independence test used in the screening step. For a classification task, any model-free feature screening method like Zhou and Zhu (2018) can be applied in a more general framework.

BIBLIOGRAPHY

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

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., et al. (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607.

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

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Brzyski, D., Peterson, C. B., Sobczyk, P., Candès, E. J., Bogdan, M., and Sabatti, C. (2017). Controlling the Rate of GWAS False Discoveries. *Genetics*, 205(1):61–75.

Bühlmann, P., Kalisch, M., and Meier, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1(1):255–278.

Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *The Annals of Statistics*, 30(4):927 – 961.

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(3):pp. 551–577.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Chen, Y., Gao, Q., Liang, F., and Wang, X. (2021). Nonlinear variable selection via deep neural networks. *Journal of Computational and Graphical Statistics*, 30(2):484–492.

Das, A., Sampson, A. L., Lainscsek, C., Muller, L., Lin, W., Doyle, J. C., Cash, S. S., Halgren, E., and Sejnowski, T. J. (2017). Interpretation of the Precision Matrix and Its Application in Estimating Sparse Brain Connectivity during Sleep Spindles from Human Electrocorticography Recordings. *Neural Computation*, 29(3):603–642.

Das, D. and Lahiri, S. N. (2019). Distributional consistency of the lasso by perturbation bootstrap. *Biometrika*, 106(4):957–964.

Dinh, V. C. and Ho, L. S. (2020). Consistent feature selection for analytic deep neural networks. *Advances in Neural Information Processing Systems*, 33:2420–2431.

Dorman, S. N., Baranova, K., Knoll, J. H. M., Urquhart, B. L., Mariani, G., Carcangiu, M. L., and

Rogan, P. K. (2016). Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning. *Mol. Oncol.*, 10(1):85–100.

Elbrächter, D., Perekrestenko, D., Grohs, P., and Bölcskei, H. (2021). Deep neural network approximation theory. *IEEE Transactions on Information Theory*, 67(5):2581–2623.

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.

Feng, J. and Simon, N. (2017). Sparse-input neural networks for high-dimensional nonparametric regression and classification. *arXiv preprint arXiv:1711.07592*.

Ghorbani, A., Abid, A., and Zou, J. (2019). Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3681–3688.

Hadley, K. E. and Hendricks, D. T. (2014). Use of nqo1 status as a selective biomarker for oesophageal squamous cell carcinomas with greater sensitivity to 17-aag. *BMC cancer*, 14(1):1–8.

Henze, N. and Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in Statistics - Theory and Methods*, 19(10):3595–3617.

Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B., and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. *Genetics*, 198(2):497–508.

Jordon, J., Yoon, J., and van der Schaar, M. (2019). KnockoffGAN: Generating knockoffs for feature selection using generative adversarial networks. In *International Conference on Learning Representations*.

Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press.

Lee, H. J., Hanibuchi, M., Kim, S.-J., Yu, H., Kim, M. S., He, J., Langley, R. R., Lehembre, F., Regenass, U., and Fidler, I. J. (2016). Treatment of experimental human breast cancer and lung cancer brain metastases in mice by macitentan, a dual antagonist of endothelin receptors, combined with paclitaxel. *Neuro-oncology*, 18(4):486–496.

Lei, L. and Fithian, W. (2018). Adapt: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):649–679.

Lemhadri, I., Ruan, F., Abraham, L., and Tibshirani, R. (2021). Lassonet: A neural network with feature sparsity. *Journal of Machine Learning Research*, 22(127):1–29.

Li, A. and Barber, R. F. (2019). Multiple testing with the structure-adaptive benjamini–hochberg algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(1):45–74.

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

Liang, F., Li, Q., and Zhou, L. (2018). Bayesian neural networks for selection of drug sensitive genes. *Journal of the American Statistical Association*, 113(523):955–972.

Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(10).

Liu, Y. and Zheng, C. (2019). Deep latent variable models for generating knockoffs. *Stat*, 8(1):e260.

Lu, Y., Fan, Y., Lv, J., and Stafford Noble, W. (2018). Deeppink: reproducible feature selection in deep neural networks. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Meinshausen, N. (2007). Relaxed lasso. *Comput. Stat. Data Anal.*, 52:374–393.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436 – 1462.

Ng, T. L. and Newton, M. A. (2022). Random weighting in lasso regression. *Electronic Journal of Statistics*, 16(1):3430–3481.

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

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Sesia, M., Katsevich, E., Bates, S., Candès, E., and Sabatti, C. (2019). Multi-resolution localization of causal variants across the genome. *bioRxiv*.

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

Shojaie, A. and Michailidis, G. (2010). Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538.

Siegmund, D. O., Zhang, N. R., and Yakir, B. (2011). False discovery rate for scanning statistics. *Biometrika*, 98(4):979–985.

Song, Z. and Li, J. (2021). Variable selection with false discovery rate control in deep neural networks. *Nature Machine Intelligence*, 3(5):426–433.

Tansey, W., Wang, Y., Blei, D., and Rabadan, R. (2018). Black box FDR. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4867–4876. PMLR.

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

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine-mapping. *bioRxiv*.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of statistics*, 37(5A):2178.

Xia, F., Zhang, M. J., Zou, J. Y., and Tse, D. (2017). Neuralfdr: Learning discovery thresholds from hypothesis features. In *NIPS*.

Xue, J. and Liang, F. (2017). A robust model-free feature screening method for ultrahigh-dimensional data. *Journal of Computational and Graphical Statistics*, 26(4):803–813.

Zhang, A. R. and Zhou, Y. (2020). On the non-asymptotic and sharp lower tail bounds of random variables. *Stat*, 9(1):e314. e314 sta4.314.

Zhou, Y. and Zhu, L. (2018). Model-free feature screening for ultrahigh dimensional datathrough a modified blum-kiefer-rosenblatt correlation. *Statistica Sinica*, 28(3):1351–1370.

Zoppoli, G., Regairaz, M., Leo, E., Reinhold, W. C., Varma, S., Ballestrero, A., Doroshow, J. H., and Pommier, Y. (2012). Putative dna/rna helicase schlafen-11 (slfn11) sensitizes cancer cells to dna-damaging agents. *Proceedings of the National Academy of Sciences*, 109(37):15030–15035.

APPENDIX A

THEORETICAL STUDY

Although the proposed method SciDNet is demonstrated based on the LassoNet algorithm (Lemhadri et al., 2021), any other sparsity-inducing Deep learning framework can be adopted at the cleaning step of SciDNet. Hence, for the theoretical study, we consider a broader framework with a general analytic DNN. Also, as the screening step is theoretically guaranteed by Xue and Liang (2017), we start the theoretical study directly from the cleaning step, assuming the data is $\{Y^i \in \mathbb{R}, X^i \in \mathscr{X}\}_{i=1}^n \overset{iid}{\sim} P_D$, where $\mathscr{X}$ is a bounded open set in $\mathbb{R}^p$, $p < n$ and the input density p_x is positive and continuous on its domain $\mathscr{X}$. In the spirit of Dinh and Ho (2020), we consider the following general analytic neural network model $f_\alpha(x)$, an L-layer neural network with parameters $\alpha = (t_{in}, T_{in}, t_{out}, T_{out}, S)$ and defined by

- **Input layer:** $h_1(x) = t_{in} + T_{in}x$;
- **Hidden layers:** $h_j(x) = \phi_{j-1}(S, h_{j-1}(x), h_{j-2}(x), \ldots, h_1(x)), j = 2, 3, \ldots, L-1$; and
- **Output layers:** $f_\alpha(x) = h_L(x) = t_{out} + T_{out}h_{L-1}(x)$

with d_i = size of the i-th layer, $d_1 = p, d_L = 1$, $T_{in} \in \mathbb{R}^{d_2 \times p}, t_{in} \in \mathbb{R}^{d_2}, T_{out} \in \mathbb{R}^{1 \times D_{L-1}}, t_{out} \in \mathbb{R}$ and $\phi_1, \phi_2, \ldots, \phi_{L-2}$ are analytic functions parameterized by the hidden layers' parameter S. This general framework covers a wide range of models, including feed-forward networks, convolutional networks, and a major subclass of residual networks. For the sake of theoretical study, we further assume that (1) the set of all feasible vectors α of the model is a hypercube $\mathscr{W} = [-A, A]^{n_\alpha}$, (2) both $\mathscr{W}$ and $\mathscr{X}$ are bounded, (3) f_α is analytic in the sense that there exist $C_1, C_2 > 0$ such that $|f_\alpha(x)| \le C_1$, and $|| \bigtriangledown_\alpha f_\alpha(x)||_\infty \le C_2, \forall \alpha \in \mathscr{W}, x \in \mathscr{X}$ and these functions are Lipschtz continuous, and (4) $Y = f_{\alpha^*}(X) + \epsilon$ with $\epsilon \sim N(0, \sigma_\epsilon^2), \alpha^* \in \mathscr{W} = [-A, A]^{n_\alpha}$ and we assume that the "true" model $f_{\alpha^*}(\cdot)$ only depends on x through a subset of significant features $S \in 1, 2, \ldots, p$ while being independent of features in S^c, $|S| = s = O(1)$. To this end, a general group Lasso estimator (Feng and

Simon, 2017; Dinh and Ho, 2020) has been defined as

$$\hat{\alpha}_n =_\alpha \{\frac{1}{n}\sum_{i=1}^{n} l(\alpha, X^i, Y^i) + \lambda_n \sum_{j=1}^{p} ||\alpha^{[:,j]}||\} \tag{A.1}$$

where $l(\alpha, x, y) = (y - f_\alpha(x))^2$ is the square-error loss, $\lambda_n > 0$ is the associated penalty parameter, $||\cdot||$ is the standard Euclidean norm and $\alpha^{[:,j]}$ is the vector of parameters associated with j-th input feature.

Now, we incorporate the notion of resampling/bootstrapping by introducing the random-weighted group lasso defined as follows

$$\hat{\alpha}_n^w =_\alpha \{\frac{1}{n}\sum_{i=1}^{n} W_i l(\alpha, X^i, Y^i) + \lambda_n \sum_{j=1}^{p} ||\alpha^{[:,j]}||\} \tag{A.2}$$

where $W_i \overset{iid}{\sim} F_W$, independent of the data distribution P_D. The random weighting scheme effectively maintains the flavor of perturbation bootstrap (Das and Lahiri, 2019), thus can be used for uncertainty quantification in the context of sparse models. Ng and Newton (2022) studied this in the context of the sparse linear model. Here, we first show that under appropriate random weights, $\hat{\alpha}_n^w$ converges to a set of well-behaved optimal hypotheses in the sense that $\mathscr{K} = \{\alpha \in \mathscr{W} : f_\alpha = f_{\alpha^*}$ and $\alpha^{[:,j]} = 0$, for $j \in S^c\}$.

We note that with the addition of random weights the overall probability measure changed to $P = P_D \times P_W$ where P_W is the probability measure of the triangular array of random weights. We define the following three sigma fields: $\mathscr{F}_n^w = \sigma\{W_1, W_2, \ldots, W_n\}$, $\mathscr{F}_n^x = \sigma\{X_1, X_2, \ldots, X_n\}$, and $\mathscr{F}_n^y = \sigma\{Y_1, Y_2, \ldots, Y_n\}$.

We further define, the risk function $R(\alpha) = E_{P_D}(Y - f_\alpha(X))^2$, the empirical risk function $R_n(\alpha) = \frac{1}{n}\sum_{i=1}^n \left(Y^i - f_\alpha(X^i)\right)^2$ and the weighted empirical risk function adding the random weights as $R_n^w(\alpha) = \frac{1}{n}\sum_{i=1}^n W_i \left(Y^i - f_\alpha(X^i)\right)^2$. The equivalence class can be expressed as $\mathscr{H}^* = \{\alpha \in \mathscr{W} : R(\alpha) = R(\alpha^*)\}$.

For simplicity, in order to obtain a bounded weight, we assume F_W is such that $P_W(C_2 < W < C_3) = 1$, for some $C_3 > 0$.

Lemma A.0.1 (Probabilistic Lipschitzness of the random-weighted empirical risk)**.** *For any $\delta > 0$, there exists $M_\delta > c_0$ such that $R_n^w(\alpha)$ is an M_δ-Lipschitz function with probability at least $1 - \delta$.*

Proof. We note that,

$$\begin{aligned}
|R(\alpha)-R(\beta)| &= |E\left((Y-f_\alpha(X))^2-(Y-f_\beta(X))^2\right)| \\
&\leq E|\left(f_\alpha(X)-f_\beta(X)\right)\left(2Y-f_\alpha(X)-f_\beta(X)\right)| \\
&\leq C_2||\alpha-\beta||E|\left(2Y-f_\alpha(X)-f_\beta(X)\right)| \\
&\leq C_2||\alpha-\beta||\left(2E|Y-f_{\alpha^*}(X)|+E|\left(f_\alpha(X)+f_\beta(X)-2f_{\alpha^*}(X)\right)|\right) \\
&\leq C_2||\alpha-\beta||(2\sigma+4C_1)
\end{aligned}$$

Similarly,

$$\begin{aligned}
|R_n^w(\alpha)-R_n^w(\beta)| &= |\frac{1}{n}\sum_{i=1}^n W_i\Big((Y^i-f_\alpha(X^i))^2-(Y^i-f_\beta(X^i))^2\Big)| \\
&\leq \frac{1}{n}\sum_{i=1}^n W_i|\Big(f_\alpha(X^i)-f_\beta(X^i)\Big)\Big(2Y^i-f_\alpha(X^i)-f_\beta(X^i)\Big)| \\
&\leq C_2||\alpha-\beta||\frac{1}{n}\sum_{i=1}^n W_i\Big(2|Y^i-f_{\alpha^*}(X^i)|+|\Big(f_\alpha(X^i)+f_\beta(X^i)-2f_{\alpha^*}(X^i)\Big)|\Big) \\
&\leq C_2||\alpha-\beta||\Big(4c_1+\frac{2}{n}\sum_{i=1}^n W_i|\epsilon^i|\Big)
\end{aligned}$$

Thus for all $M_\delta > 4C_1C_2$, the proof is completed by noting the following

$$\begin{aligned}
&P\Big(|R_n^w(\alpha)-R_n^w(\beta)| \leq M_\delta||\alpha-\beta||, \forall \alpha,\beta \in \mathscr{W}\Big) \\
&\geq P\Big(\frac{1}{n}\sum_{i=1}^n W_i|\epsilon^i| \leq \frac{M_\delta}{2C_2}-2C_1\Big) = E_{P_W}\left[P\Big(\frac{1}{n}\sum_{i=1}^n W_i|\epsilon^i| \leq \frac{M_\delta}{2C_2}-2C_1|W_1,W_2,\ldots,W_n\Big)\right] \\
&= E_{P_W}\left[1-P\Big(\frac{1}{n}\sum_{i=1}^n W_i|\epsilon^i| \geq \frac{M_\delta}{2C_2}-2C_1|W_1,W_2,\ldots,W_n\Big)\right] \\
&\geq E_{P_W}\left[1-\frac{\frac{1}{n}\sum_{i=1}^n W_iE|\epsilon^1|}{\frac{M_\delta}{2C_2}-2C_1}\right] \geq 1-\frac{C_3E|\epsilon^1|}{\frac{M_\delta}{2C_2}-2C_1}
\end{aligned}$$

□

Lemma A.0.2 (Generalization bound for the random-weighted empirical risk)**.** *For any* $\delta > 0$, $\exists$ $C_4(\delta) > 0$ *such that* $P\Big(|R_n^w(\alpha)-R(\alpha)| \leq C_4\frac{log(n)}{\sqrt{n}}\Big) \geq 1-\delta, \forall \alpha \in \mathscr{W}$

Proof. Note that $nR_n^w(\alpha) = \sum_{i=1}^n W_i\Big(Y^i-f_\alpha(X^i)\Big)^2 = \sum_{i=1}^n W_iZ_i^2$, where $Z_i = Y^i - f_\alpha(X^i)$ which follows $N\Big(f_{\alpha^*}(X^i)-f_\alpha(X^i),\sigma_\epsilon^2\Big)$, conditional on $\mathscr{F}_X$.

Hence, conditional on $\mathscr{F}_X, \mathscr{F}_W$, $\frac{nR_n^w(\alpha)}{\sigma_\epsilon^2} \sim$ a weighted non-central χ^2 distribution. We will use the following Lemma 3 to get a sharp tail bound for weighted non-central χ^2 distribution. Hence,

$$\begin{aligned} P\Big(|R_n^w(\alpha) - R(\alpha)| > \frac{t}{2}\Big) &\leq E_{\mathscr{F}_X,\mathscr{F}_W}\Big[2exp\Big(-\frac{Cn^2t^2}{2n + 2\sum_{i=1}^n W_i\left(f_{\alpha^*}(X^i) - f_\alpha(X^i)\right)^2 - \sum_{i=1}^n W_i}\Big)\Big] \\ &\leq E_{\mathscr{F}_X,\mathscr{F}_W}\Big[2exp\Big(-\frac{Cn^2t^2}{2n + 2\sum_{i=1}^n W_i\left(f_{\alpha^*}(X^i) - f_\alpha(X^i)\right)^2}\Big)\Big] \\ &\leq E_{\mathscr{F}_X,\mathscr{F}_W}\Big[2exp\Big(-\frac{Cn^2t^2}{2n + 2nC_3 4C_1^2}\Big)\Big] \leq 2exp(-\tilde{c}nt^2) \end{aligned}$$

The rest of the proof follows the direct proof of the Lemma 3.3 Dinh and Ho (2020). □

Theorem A.0.3 (Convergence of random-weighted Group Lasso)**.** *For any $\delta > 0$, there exist $C_\delta, C' > 0$ and $N_\delta > 0$ such that for all $n \geq N_\delta$,*

$$d(\hat{\alpha}_n^w, \mathscr{H}^*) \leq C_\delta\Big(\lambda_n^{\frac{\nu}{\nu-1}} + \frac{logn}{\sqrt{n}}\Big)^{\frac{1}{\nu}} \; and \; ||\hat{\alpha}_n^{w[:,S^c]}|| \leq 4C_4\frac{logn}{\lambda_n\sqrt(n)} + C'd(\hat{\alpha}_n, \mathscr{H}^*)$$

where $d(x, Z) = \inf_{z\in Z} ||x - z||$.

The proof directly follows from the above mentioned Lemma 1,2 and the theorem 3.3 from Dinh and Ho (2020).

Theorem A.0.3 demonstrates the convergence of the random-weighted group-Lasso estimates to a set of "well-behaved" optimal hypotheses $\mathscr{K} = \{\alpha \in \mathscr{W} : f_\alpha = f_{\alpha^*}$ and $||\alpha^{[:,S^c]}|| = 0\} \subset \mathscr{H}^*$. However, this depends on the regularization parameter λ_n; finding its optimum value is generally a daunting task in practice. As a solution, the importance score considering the whole regularization path would provide a more robust way to identify false discoveries. Following our proposed method in Section 3.2.3, here we consider the cleaning step in view of the random-weighted group-Lasso penalized DNN. We repeat this step over B independently generated random-weighting scheme and recall the importance scores $\hat{\lambda}_j$ and the corresponding ranks of the importance score $\mathscr{I}_j$ for the b-th scheme as follows:

$\hat{\lambda}_j^b = \max\{\lambda_n \ni ||\alpha_n^{w,[:,j]}|| \neq 0\}$ = maximum value of λ up to which the j-th feature exists in the model for the b-th random scheme, and

$\mathscr{I}_j^b$ = rank of the importance scores = $\sum_{j' \neq j} 1\left(\hat{\lambda}_j \leq \hat{\lambda}_{j'}\right)$ and the averaged rank $\bar{\mathscr{I}}_j = \sum_{b=1}^B \mathscr{I}_j^b$

Our basic strategy is to select the features with high-importance scores consistent in all the bootstrap replication. Hence, for a cutoff δ, define the number of selected features as $N_+(\delta) = \sum_{j=1}^{p} 1\left(\bar{\mathscr{I}}_j \leq \delta\right)$. We further propose our estimate of the FDR as

$$F\hat{D}R(\delta) = \frac{\hat{e}_0(\delta)}{N_+(\delta)},$$

$$\text{where } \hat{e}_0(\delta) = \text{estimated number of false discoveries} = \frac{2}{B}\sum_{b=1}^{B}\left\{\sum_{j\in 1}^{p} 1(\mathscr{I}_j^b \leq \delta, \mathscr{I}_j^b \not\in \mathscr{N}(\bar{\mathscr{I}}_j, \kappa))\right\}$$

We further calculate the data-dependent optimum cutoff as $\Delta^* = \max\{\delta \in \{\bar{\mathscr{I}}_1, \bar{\mathscr{I}}_2, \ldots, \bar{\mathscr{I}}_p\} : F\hat{D}R(\delta) < q\}$ for some pre-specific FDR control level q.

Theorem A.0.4 (Convergence of the estimated FDR)**.** *Using the data-dependent cutoff* Δ^**, the actual FDR is bounded by the user-specified level* q*; that is* $E\left(\frac{\hat{e}_0(\Delta^*)}{N_+(\Delta^*)}\right) < q$ *as* $n \to \infty$*.*

Proof. The feature importance scores $\bar{\mathscr{I}}_j$ evidently provide the information on the survival of the feature X_j over the whole regularization path. Now, under the setup of the random-weighted group-lasso framework, this is uniquely monitored by the KKT condition: $||\alpha^{[:,j]}|| = 0$ if $||\left(\frac{\partial f_\alpha(X)}{\partial \alpha^{[:,j]}}\right)^T_{p_j \times n} diag(W_1, W_2, \ldots, W_n)(Y - f_\alpha(X))_{n\times 1}|| < \lambda\sqrt{p_j}$, where p_j is the number of total parameters associated with the j-th feature. Hence, this L_2 norm in the KKT condition can be treated as the importance score $\hat{\lambda}_j$, mentioned in Section 3.2.3. Now, for the null features $j \in S_0^c$, the derivative term is $o\left(n^{-1/4} log(n)\right)$ by the continuity of $\bigtriangledown_\alpha f_\alpha(x)$ and the convergence of $\hat{\alpha}_n^w$ to $\mathscr{K}$. Also, by Lemma A.0.2, the residual term is $o\left(n^{-1/2} log(n)\right)$. Also, the random weights W's are bounded in $[C_2, C_3]$. Hence, this new $\hat{\lambda}_j, j = 1, 2, \ldots, p$ become exchangeable asymptotically.

As a consequence of the exchangeable variables, the ranks of the importance scores $\hat{\lambda}_j$ follow uniform distribution asymptotically. Hence, for large n, $\mathscr{I}_j^w = \sum_{j' \neq j} 1\left(\hat{\lambda}_j \leq \hat{\lambda}_{j'}\right) \sim U(p - s + 1, p)$. Now, we consider the following random variable names *False Discovery Proportion (FDP)* whose expectation is the FDR.

$$FDP = \frac{\sum_{j=1}^{p} 1\left(\bar{\mathscr{I}}_j \leq \Delta^*, j \in S_0^c\right)}{\sum_{j=1}^{p} 1\left(\bar{\mathscr{I}}_j \leq \Delta^*\right)}$$

$$= \underbrace{\frac{\frac{2}{B}\sum_{b=1}^{B}\left\{\sum_{j\in 1}^{p} 1(\mathscr{I}_j^b \leq \Delta^*, \mathscr{I}_j^b \notin \mathscr{N}(\bar{\mathscr{I}}_j,\kappa))\right\}}{\sum_{j=1}^{p} 1\left(\bar{\mathscr{I}}_j \leq \Delta^*\right)}}_{\leq q} \cdot \underbrace{\frac{\sum_{j=1}^{p} 1\left(\bar{\mathscr{I}}_j \leq \Delta^*, j \in S_0^c\right)}{\frac{2}{B}\sum_{b=1}^{B}\left\{\sum_{j\in 1}^{p} 1(\mathscr{I}_j^b \leq \Delta^*, \mathscr{I}_j^b \notin \mathscr{N}(\bar{\mathscr{I}}_j,\kappa))\right\}}}_{R(\Delta^*)}$$

The first part of FPR is $\leq q$ by the typical choice of Δ^* and hence, in order to show the asymptotic FDR control, all we need to show is $\lim_{n\to\infty} E(R(\Delta^*)) \leq 1$.

$$E(R(\Delta^*)) = E\left(\frac{\sum_{j=1}^{p} 1\left(\bar{\mathscr{I}}_j \leq \Delta^*, j \in S_0^c\right)}{\frac{2}{B}\sum_{b=1}^{B}\left\{\sum_{j\in 1}^{p} 1(\mathscr{I}_j^b \leq \Delta^*, \mathscr{I}_j^b \notin \mathscr{N}(\bar{\mathscr{I}}_j,\kappa))\right\}}\right)$$

$$= E\left(\frac{\sum_{j=1}^{p} 1\left(\bar{\mathscr{I}}_j \leq \Delta^*, \{\mathscr{I}_j^b\}_{b=1}^{B} \sim Uniform\right)}{\frac{2}{B}\sum_{b=1}^{B}\left\{\sum_{j\in 1}^{p} 1(\mathscr{I}_j^b \leq \Delta^*, \mathscr{I}_j^b \notin \mathscr{N}(\bar{\mathscr{I}}_j,\kappa))\right\}}\right)$$

$$= E\left(\frac{\sum_{j=1}^{p} 1\left(\bar{\mathscr{I}}_j \leq \Delta^*, \{\mathscr{I}_j^b\}_{b=1}^{B} \sim Uniform, \mathscr{I}_j^b \in \mathscr{N}(\bar{\mathscr{I}}_j,\kappa), \forall b\right)}{\frac{2}{B}\sum_{b=1}^{B}\left\{\sum_{j\in 1}^{p} 1(\mathscr{I}_j^b \leq \Delta^*, \mathscr{I}_j^b \notin \mathscr{N}(\bar{\mathscr{I}}_j,\kappa))\right\}}\right) +$$

$$E\left(\frac{\sum_{j=1}^{p} 1\left(\bar{\mathscr{I}}_j \leq \Delta^*, \{\mathscr{I}_j^b\}_{b=1}^{B} \sim Uniform, \mathscr{I}_j^b \notin \mathscr{N}(\bar{\mathscr{I}}_j,\kappa), \forall b\right)}{\frac{2}{B}\sum_{b=1}^{B}\left\{\sum_{j\in 1}^{p} 1(\mathscr{I}_j^b \leq \Delta^*, \mathscr{I}_j^b \notin \mathscr{N}(\bar{\mathscr{I}}_j,\kappa))\right\}}\right)$$

$$\leq E\left(\frac{2\sum_{j=1}^{p} 1\left(\bar{\mathscr{I}}_j \leq \Delta^*, \{\mathscr{I}_j^b\}_{b=1}^{B} \sim Uniform, \mathscr{I}_j^b \notin \mathscr{N}(\bar{\mathscr{I}}_j,\kappa), \forall b\right)}{\frac{2}{B}\sum_{b=1}^{B}\left\{\sum_{j\in 1}^{p} 1(\mathscr{I}_j^b \leq \Delta^*, \mathscr{I}_j^b \notin \mathscr{N}(\bar{\mathscr{I}}_j,\kappa))\right\}}\right)$$

$$\leq E\left(\frac{2\sum_{j=1}^{p} 1\left(\bar{\mathscr{I}}_j \leq \Delta^*, \mathscr{I}_j^b \notin \mathscr{N}(\bar{\mathscr{I}}_j,\kappa), \forall b\right)}{\frac{2}{B}\sum_{b=1}^{B}\left\{\sum_{j\in 1}^{p} 1(\mathscr{I}_j^b \leq \Delta^*, \mathscr{I}_j^b \notin \mathscr{N}(\bar{\mathscr{I}}_j,\kappa))\right\}}\right)$$

$$\leq E\left(\frac{\frac{2}{B}\sum_{b=1}^{B}\left\{\sum_{j\in 1}^{p} 1(\mathscr{I}_j^b \leq \Delta^*, \mathscr{I}_j^b \notin \mathscr{N}(\bar{\mathscr{I}}_j,\kappa))\right\}}{\frac{2}{B}\sum_{b=1}^{B}\left\{\sum_{j\in 1}^{p} 1(\mathscr{I}_j^b \leq \Delta^*, \mathscr{I}_j^b \notin \mathscr{N}(\bar{\mathscr{I}}_j,\kappa))\right\}}\right) = 1$$

The third inequality is true by the fact that the neighborhood $\mathscr{N}(\bar{\mathscr{I}}_j,\kappa)$ is much smaller than its complement region. The last inequality is true because of the fact that $P(Z_1 \leq z, Z_2 \leq z, \ldots, Z_p \leq z) \leq \frac{1}{p}\sum_{i=1}^{p} P(Z_i \leq z)$, for any set of random variables $Z_1, Z_2, \ldots, Z_p$. Hence, the actual FDR is controlled at the user-specific bound q by selecting the features with ranks of their importance score greater than the data-dependent threshold Δ^*.

□

Lemma 3: Sharp tail bound for weighted non-central χ^2 distribution

Consider a weighted non-central χ^2 distributed random variable $Y = \sum_{i=1}^{k} u_i Z_i^2, Z_i \sim N(\mu_i, 1)$ independently and $\sum_{i=1}^{k} \mu_i^2 = \lambda$. Then for the centralized random variable $X = Y - \sum_{i=1}^{k} u_i(1+\mu_i^2)$, the following sharp tail bound holds: there exists constants c,c',C>0, such that

$$P(X \geq x) \leq \text{ c } exp(-\frac{ct^2}{2k+2\sum_{i=1}^{k} u_i\mu_i^2 - \sum_{i=1}^{k} u_i}), \forall 0 \leq x \leq c'(2k+2\sum_{i=1}^{k} u_i\mu_i^2 - \sum_{i=1}^{k} u_i)$$

Proof:

The moment-generating function of X:

$$\phi_X(t) = exp\Big[\frac{2t^2}{1-2t}\sum_{i=1}^{k} u_i\mu_i^2 - \frac{k}{2}(log(1-2t)+2t) + t(k - \sum_{i=1}^{k} u_i)\Big]$$

Note, for $0 \leq t \leq 1/2, 2t^2 \leq -log(1-2t) - 2t \leq \frac{2t^2}{1-2t}$.

This implies, for $0 \leq t \leq 2/5$,

$$t^2(2k+2\sum_{i=1}^{k} u_i\mu_i^2 - \sum_{i=1}^{k} u_i) \leq log(\phi_X(t)) \leq 5t^2(2k+2\sum_{i=1}^{k} u_i\mu_i^2 - \sum_{i=1}^{k} u_i)$$

Next, the exact tail bound can be obtained by applying theorem 1 from Zhang and Zhou (2020).

APPENDIX B

ADDITIONAL TECHNICAL DETAILS

In this section, additional technical and implementation details on the proposed algorithm SciDNet are provided.

B.0.1 Hyperparameter Selection

Recently developed Deep Learning (DL) models are generally governed by several hyperparameters and properly tuning them is necessary to get effective results. The proposed SciDNet relies on the following hyperparameters: (1) size of the active set $\hat{S}_n$, (2) the intracluster correlation bound r, (3) LassoNet tuning parameters λ and M and (4) κ used in neighbourhood selection in cleaning step. SciDNet is fairly robust to most of the associated hyperparameters. We discuss a practical way to tune all these hyperparameters here:

1. To choose the size of the active set, we propose to select a bigger active set with size proportional to $v_n = [n/log(n)]$. As we further cluster the active variables in the clustering step, a slightly bigger active set with boost up the confidence of sure screening property, See the section B.0.5 for an example.

2. After clustering, the intra-cluster correlation bound r should be fixed at some higher value (usually at 0.9 or 0.95) otherwise the cluster sizes will be inflated.

3. In the cleaning step, a thorough grid search has been done over λ considering $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_r$; in practice, a small value is fixed for λ_1 where all the variables are present in the model. Then the value of the tuning parameter gradually increased up to λ_r, where there are no variables present in the model. The other hyperparameter for LassoNet is the hierarchy coefficient M for which we follow the path considered in Lemhadri et al. (2021) and set $M = 10$. However, a more flexible approach would be a parallel grid search for M as well.

4. The neighborhood length κ can be chosen using the phase transition in the ranks of the

Table S1 Empirical power and observed FDR of SciDNet with standard error in parentheses for Gaussian features

			Nonlinear Additive				Nonlinear with interaction				Linear			
ρ	snr	q	0.01	0.05	0.1	0.15	0.01	0.05	0.1	0.15	0.01	0.05	0.1	0.15
$\rho = 0.9$	$snr = 9:1$	Power	0.79 (0.19)	0.93 (0.11)	0.96 (0.09)	0.96 (0.09)	0.99 (0.06)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
		FDR	0.00 (0.00)	0.00 (0.02)	0.01 (0.03)	0.02 (0.05)	0.00 (0.00)	0.02 (0.05)	0.03 (0.06)	0.04 (0.08)	0.00 (0.00)	0.01 (0.04)	0.01 (0.05)	0.01 (0.05)
	$snr = 8:2$	Power	0.59 (0.20)	0.82 (0.15)	0.86 (0.14)	0.87 (0.13)	0.84 (0.24)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
		FDR	0.00 (0.03)	0.00 (0.03)	0.03 (0.07)	0.04 (0.08)	0.00 (0.00)	0.01 (0.03)	0.02 (0.06)	0.04 (0.07)	0.01 (0.05)	0.02 (0.06)	0.02 (0.06)	0.02 (0.06)
	$snr = 7:3$	Power	0.42 (0.20)	0.65 (0.12)	0.77 (0.14)	0.81 (0.14)	0.72 (0.26)	0.94 (0.12)	0.95 (0.12)	0.96 (0.11)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
		FDR	0.00 (0.00)	0.00 (0.00)	0.00 (0.03)	0.01 (0.05)	0.01 (0.04)	0.03 (0.06)	0.03 (0.07)	0.05 (0.09)	0.00 (0.02)	0.02 (0.05)	0.02 (0.05)	0.02 (0.06)
$\rho = 0.95$	$snr = 9:1$	Power	0.83 (0.18)	0.96 (0.08)	0.98 (0.06)	0.98 (0.06)	0.99 (0.04)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
		FDR	0.00 (0.03)	0.02 (0.05)	0.03 (0.07)	0.04 (0.07)	0.00 (0.00)	0.04 (0.07)	0.08 (0.09)	0.11 (0.10)	0.00 (0.02)	0.06 (0.09)	0.08 (0.10)	0.11 (0.12)
	$snr = 8:2$	Power	0.60 (0.29)	0.82 (0.17)	0.87 (0.14)	0.89 (0.12)	0.98 (0.08)	0.99 (0.05)	0.99 (0.05)	0.99 (0.05)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
		FDR	0.00 (0.03)	0.02 (0.05)	0.02 (0.06)	0.03 (0.07)	0.01 (0.04)	0.04 (0.07)	0.08 (0.12)	0.12 (0.12)	0.01 (0.03)	0.05 (0.09)	0.07 (0.10)	0.09 (0.11)
	$snr = 7:3$	Power	0.32 (0.22)	0.61 (0.19)	0.79 (0.15)	0.82 (0.15)	0.80 (0.27)	0.96 (0.11)	0.98 (0.05)	0.98 (0.05)	0.93 (0.19)	0.96 (0.11)	0.97 (0.09)	0.97 (0.09)
		FDR	0.00 (0.00)	0.01 (0.04)	0.01 (0.05)	0.03 (0.08)	0.00 (0.02)	0.04 (0.07)	0.07 (0.09)	0.12 (0.10)	0.00 (0.03)	0.04 (0.08)	0.06 (0.08)	0.07 (0.09)

importance scores, as described in Section 3.2.3.

B.0.2 Phase transition observed for the CCLE data

The main reason for the phase transition is that, for a null predictor $X_j, j \in S_0^c$, different bootstrap replicates reshuffle its feature importance each time, whereas, for a nonnull predictor $X_j, j \in S_0$, the feature importance is much stable in different bootstrap replicates. SciDNet effectively captures this characteristic to identify the null features. As a demonstration, here we present in Figure B.1, the bootstrap distribution of rank of the importance scores for the top 25 important cluster representatives via box plots. The green and purple colors respectively indicate if the cluster representatives are selected or rejected by the SciDNet. We can observe the phase transition consistently for all five drugs, and SciDNet selects only those important representatives with reduced variability over the bootstrap replicates.

B.0.3 More Simulation Results

Here we demonstrate finite sample performance of SciDNet under various linear and nonlinear models with varying multicollinearity level under different signal-to-noise-ratio.

B.0.3.1 Using Gaussian Features

For the high dimensional predictors, n i.i.d. copies are first generated from $X \sim N_p(0,\Sigma)$, where $n = 600$, $p = 5000$ and the covariance matric Σ is chosen as a toeplitz matrix with $\Sigma_{ij} = \rho^{|i-j|}$. The value of ρ is varied to explore different correlation strengths. We set the set of truly significant variables $S = \{100,200,300,400,500\}$ with $s = 5$. The response y is generated from

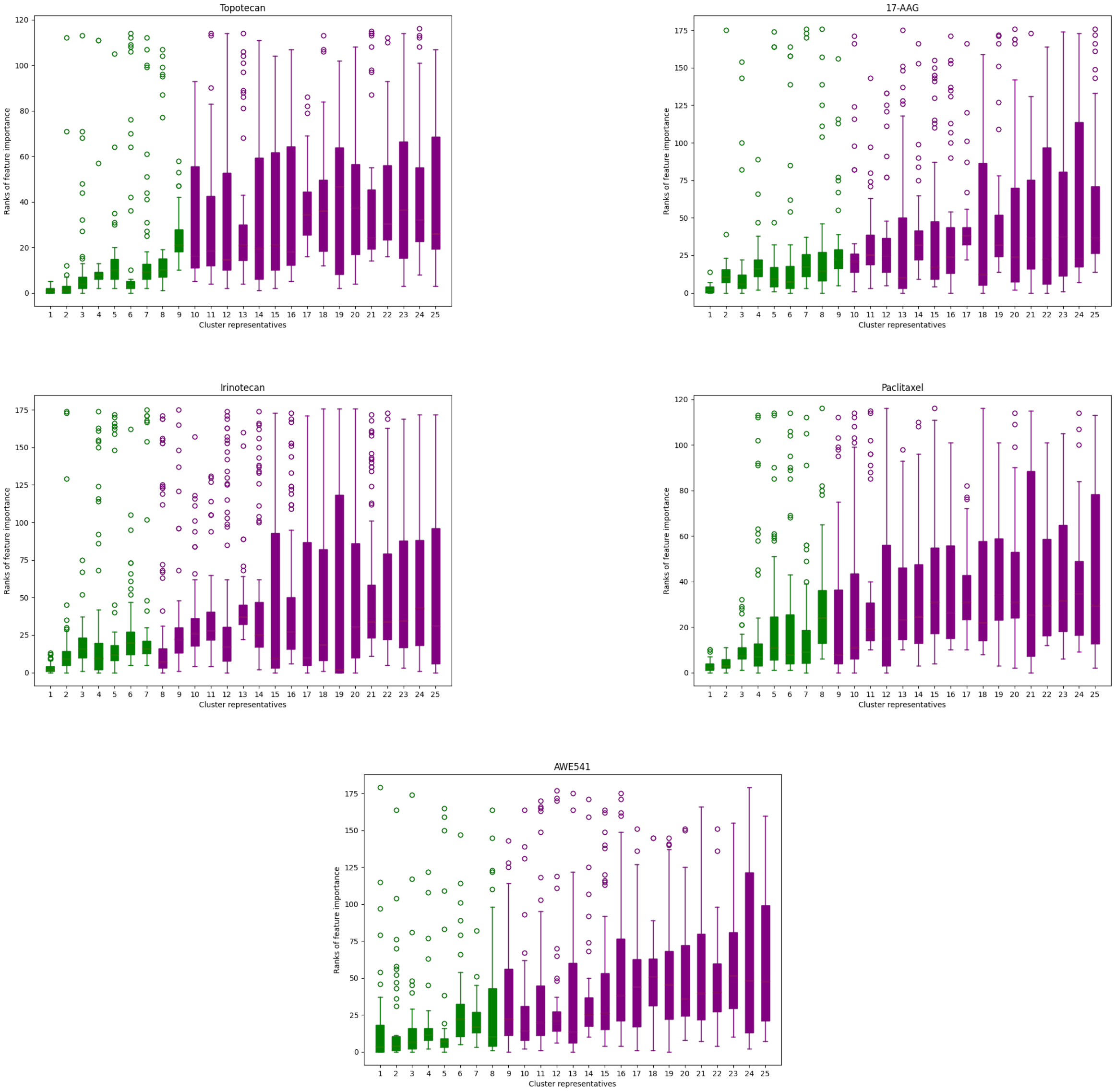

Figure B.1 The phase transition property illustrated for the five anticancer drugs considered: (1) Topotecan, (2) 17-AAG, (3) Irinotecan, (4) Paclitaxel, and (5) AWE541, respectively (from top left).

Table S2 Empirical power and observed FDR of SciDNet with standard error in parentheses for non-gaussian features

			Nonlinear Additive				Nonlinear with interaction				Linear			
ρ	snr	q	0.01	0.05	0.1	0.15	0.01	0.05	0.1	0.15	0.01	0.05	0.1	0.15
$\rho = 0.9$	$snr = 9:1$	Power	0.68 (0.12)	0.96 (0.15)	1.00 (0.09)	1.00 (0.07)	0.92 (0.05)	0.95 (0.02)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
		FDR	0.00 (0.00)	0.00 (0.01)	0.04 (0.01)	0.07 (0.05)	0.00 (0.00)	0.01 (0.06)	0.04 (0.06)	0.06 (0.07)	0.00 (0.00)	0.01 (0.04)	0.01 (0.05)	0.01 (0.05)
	$snr = 8:2$	Power	0.56 (0.24)	0.86 (0.11)	0.94 (0.14)	0.95 (0.13)	0.76 (0.23)	.93 (0.02)	1.00 (0.04)	1.00 (0.03)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
		FDR	0.00 (0.01)	0.00 (0.01)	0.06 (0.05)	0.09 (0.07)	0.00 (0.00)	0.00 (0.02)	0.04 (0.03)	0.07 (0.07)	0.01 (0.05)	0.02 (0.06)	0.02 (0.06)	0.02 (0.06)
	$snr = 7:3$	Power	0.42 (0.21)	0.77 (0.13)	0.91 (0.16)	0.94 (0.12)	0.73 (0.26)	0.94 (0.12)	0.95 (0.15)	0.96 (0.14)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
		FDR	0.00 (0.00)	0.00 (0.01)	0.02 (0.03)	0.06 (0.04)	0.01 (0.05)	0.04 (0.06)	0.05 (0.07)	0.05 (0.09)	0.00 (0.02)	0.02 (0.05)	0.02 (0.05)	0.02 (0.06)
$\rho = 0.95$	$snr = 9:1$	Power	0.81 (0.19)	0.95 (0.07)	0.98 (0.06)	0.98 (0.07)	0.99 (0.04)	0.99 (0.03)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
		FDR	0.00 (0.01)	0.03 (0.06)	0.03 (0.04)	0.05 (0.03)	0.00 (0.00)	0.04 (0.07)	0.08 (0.09)	0.09 (0.13)	0.00 (0.01)	0.03 (0.05)	0.07 (0.11)	0.10 (0.14)
	$snr = 8:2$	Power	0.65 (0.29)	0.84 (0.16)	0.89 (0.17)	0.89 (0.12)	0.94 (0.07)	0.97 (0.04)	0.99 (0.07)	0.99 (0.07)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
		FDR	0.00 (0.03)	0.01 (0.05)	0.04 (0.06)	0.05 (0.06)	0.01 (0.03)	0.04 (0.04)	0.07(0.14)	0.11 (0.11)	0.01 (0.02)	0.05 (0.09)	0.06 (0.14)	0.09 (0.11)
	$snr = 7:3$	Power	0.47 (0.22)	0.64 (0.17)	0.75 (0.19)	0.87 (0.11)	0.82 (0.27)	0.95 (0.10)	0.98 (0.04)	0.98 (0.02)	0.95 (0.10)	0.96 (0.11)	0.97 (0.08)	0.97 (0.06)
		FDR	0.00 (0.00)	0.02 (0.04)	0.02 (0.04)	0.04 (0.09)	0.00 (0.01)	0.04 (0.05)	0.09 (0.03)	0.13 (0.09)	0.00 (0.02)	0.04 (0.05)	0.05 (0.08)	0.08 (0.07)

$y = g(x) + \epsilon$. Here we entertain the following three models:

1. **Linear**: $g(x) = x_S\beta_S$ with β_S generated from $N(2, 0.1)$ independently and $\beta_{S^c} = 0$,

2. **Nonlinear additive**: $g(x) = 2x_{100} + 2x_{200}^3 + e^{x_{300}} + 6\sin x_{400} + 2ReLu(x_{500}^3)$, where ReLu(x)=max(x,0)

3. **Nonlinear with interaction**: $g(x) = 2x_{100} + 2x_{200}^3 + e^{x_{300}} + 6x_{400}x_{500}$

In each case, the random noise ϵ is independently generated from $N(0, \sigma^2)$, where the value of σ^2 is chosen to maintain the signal-to-noise ratio at the desired level. To this end, we define the signal-to-noise ratio as $snr = \frac{var(g(x))}{\sigma^2}$. Here we consider three levels of $snr = 9:1, 8:2$, and $7:3$. Table S1 shows that SciDNet continues to maintain satisfactory power while successfully controlling the FDR below the threshold $q = 0.01, 0.05, 0.1, 0.15$. The average cluster size is observed at 8.3 for $\rho = 0.9$ and 13.4 for $\rho = 0.95$.

B.0.3.2 Using Non-gaussian Features

To check SciDNet's performance under a non-gaussian setup, n iid copies of high-dimensional feature vector X are generated from multivariate $t_p(5)$ distribution considering the same correlation structure as in the previous section B.0.3.1, with n=600, p=5000. The remaining simulation setting is consistent with the previous section 2.1. The performance of SciDNet is presented in table S2 which is quite analogous to the results of gaussian features.

B.0.4 Performance of existing feature selection methods in the presence of high multicollinearity

In this section, we present a numerical illustration of the performance of several recently proposed nonlinear FDR-controlled feature selection algorithms. The predictors are first generated from $X_i \sim N_p(0,\Sigma), i = 1,2,\ldots,n$, for multiple combination of (n, p) and the covariance matric Σ is chosen as a toeplitz matrix with $\Sigma_{ij} = \rho^{|i-j|}, \rho = 0.1, 0.5,$ and 0.9. Under simplistic setting, the response y is generated from $y = x_S\beta_S + \epsilon$, $S = \{5,10,\ldots,50\}, |S| = 10$, with β_S generated from $N(\beta_0, 0.1)$ independently and $\beta_{S^c} = 0$. The random noise $\epsilon \sim N(0,1)$. We focus on the Model-X knockoff (Candès et al., 2018), SurvNet (Song and Li, 2021), and DeepPINK (Lu et al., 2018). For a more rigorous analysis, we consider two different versions of Model-X knockoff - (1) Model-X-Estimated, where the knockoffs are generated using an estimated multivariate Gaussian distribution and (2) Model-X-True, where the knockoffs are generated using the true data generating multivariate gaussian distribution mentioned above. For the knockoff generation, we consider the equicorrelated construction using the R package knockoff: The Knockoff Filter for Controlled Variable Selection. To implement the SurvNet and DeepPINK, we use the codes mentioned in the respective papers Song and Li (2021); Lu et al. (2018). We set $q = 0.15$ as the FDR control threshold.

Table 4.1 reveals several interesting characteristics. Both Model-X-Estimated and Model-X-True maintain the power-FDR balance under a low correlation setup. However with higher multicollinearity, Model-X-Estimated fails to control the FDR below the specified threshold while the Model-X-True controls the FDR efficiently. This disparity indicates Model-X procedure induces inflation in false discoveries if the knockoffs are not generated properly under a 'difficult' situation. As expected, the DL-based algorithms, such as SurvNet and DeepPINK work much better in big-n-small-p and low correlation setups but typically fail in other cases, indicating their reduced effectiveness in ultrahigh dimensional data with small sample sizes.

Table S3 Empirical power and observed FDR of various feature selection algorithms with standard error in parentheses

			$\beta = 2$			$\beta = 4$		
	(n, p)		ρ =0.1	ρ =0.5	ρ =0.9	ρ =0.1	ρ =0.5	ρ =0.9
Model-X-Estimated	$(400, 1000)$	Power	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
		FDR	0.13 (0.17)	0.12 (0.12)	0.27 (0.18)	0.11 (0.19)	0.20 (0.18)	0.27 (0.20)
Model-X-True	$(400, 1000)$	Power	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
		FDR	0.08 (0.13)	0.09 (0.12)	0.14 (0.17)	0.12 (0.14)	0.11 (0.13)	0.08 (0.12)
SurvNet	$(400, 1000)$	Power	0.27 (0.20)	0.32 (0.22)	0.35 (0.24)	0.49 (0.24)	0.52 (0.28)	0.58 (0.29)
		FDR	0.31 (0.36)	0.53 (0.30)	0.59 (0.23)	0.21 (0.21)	0.53 (0.18)	0.60 (0.17)
	$(10000, 60)$	Power	0.99 (0.05)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
		FDR	0.20 (0.15)	0.80 (0.02)	0.78 (0.07)	0.14 (0.11)	0.80 (0.02)	0.56 (0.32)
DeepPINK	$(400, 1000)$	Power	0.01 (0.02)	0.03 (0.04)	0.00 (0.00)	0.03 (0.04)	0.01 (0.03)	0.02 (0.05)
		FDR	0.23 (0.40)	0.35 (0.42)	0.33 (0.47)	0.45 (0.44)	0.24 (0.41)	0.24 (0.40)
	$(10000, 60)$	Power	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
		FDR	0.18 (0.04)	0.29 (0.13)	0.25 (0.11)	0.17 (0.01)	0.24 (0.12)	0.24 (0.12)

B.0.5 Model implementation details and Sensitivity Analysis

In this section, we mention the implementation details of SciDNet that we consider for the simulation study and real data analysis. To select the size of the active set $\hat{S}_n$in the screening step, in consistence with Xue and Liang (2017), we set $|\hat{S}_n| = [\frac{2n}{log(n)}]$ by selecting the predictors with the top $|\hat{S}_n|$ Henze–Zirkler test statistic $\tilde{w}_k^*$, where $[z]$ denotes the integer part of z. In all our simulation scenarios, we set r=0.9, the hyperparameter for intra-cluster correlation bound to further integrate highly correlated conditionally dependent clusters. In the cleaning step, for LassoNet 100 dimensional one-hidden-layer feed-forward neural network has been used; a more detailed model architecture can be found in the appendix in Lemhadri et al. (2021). For creating the compact neighborhood in the cleaning step, each time we choose the value of κ utilizing the phase transition property mentioned in section 2.2 of the main manuscript. The feature selection performance of the SciDNet is demonstrated by calculating the average power and cFDR along with their standard error observed in 50 Monte Carlo replications. Each data set is randomly divided into train, validation, and test with a 70-10-20 split. To assess the prediction performance, the test Mean Square Error (MSE) before and after the variable selection has been shown as part of the simulation study. For the prediction model, a 40-dimensional two-hidden-layer feed-forward neural network with ReLU and linear activation function is considered with Adam as the optimizer. For the regression tree, we used the bagging for further stabilization, as mentioned in Breiman (1996). The number of leaves and nodes is chosen by minimizing the

Table S4 Drug-sensitive genes identified by SciDNet and confirming references

Drug	Selected clusters of genes	Confirming references
Topotecan	*{**SLFN11**},{TUFT1,THRB},{CDT1,SF3A2,SNRPA},{FTH1P10,FTH1},{RPL18},{KLF5}, {RPL11,RPL5P4,RPS8,RPL5,RPL10A,AL162151.3,RPS9,RPL3}, {KIF15,CCNA2,LMNB1,KIF22,AC009133.14},{MATN2,HSPB8},*	Barretina et al. (2012) Li et al. (2012)
17-AAG	*{**NQO1**,CTD.2033A16.1},{BAX},{SLC16A3},{PHPT1,SH3BP1} {SPCS3,DCTD},{CTD.2008A1.2,SORD},{NSMCE4A},{CSK}*	Hadley and Hendricks (2014) Barretina et al. (2012)
Irinotecan	*{**SLFN11**},{KIF15,LMNB1,ARHGAP19},{TCEANC2},{KIF21B},{SQSTM1},{HDAC11} {KHDRBS1,HNRNPA1P35,HMGB2,HNRNPA1,HNRNPA1L2, HNRNPA1P48,HNRNPA1P7,AC021224.1,HNRNPA1P10,RBMX}*	Barretina et al. (2012) Li et al. (2012)
PaclitaXel	*{PARP1,**BCL2L1**},{MMP24},{DIMT1},{RP11.872D17.4,SSRP1,MTA2},{DCUN1D3} {RPL10AP6,RPL10A,EEF2,RPL3},{ARHGAP11B,ARHGAP11A,BUB1B,CASC5},{HCLS1,LCP1}*	Dorman et al. (2016) Lee et al. (2016)
AEW541	*{TCEAL4,**MID2**},{E2F6},{AC096772.6},{SLC44A1},{PGM1} {ATP8B2,RNF122}, {RP11.1017G21.4},{ETNK2},{NHS},{ATG13}*	Liang et al. (2018)

MSE on the validation set.

To access the error bar for the sensitivity analysis, we generate typical data using the polynomial setup (section 3 in the main manuscript, with $\beta = 2, \sigma^2 = 1$) and rerun SciDNet 50 times on the same data and set $q = 0.1$ as FDR-control threshold. The mean and standard deviations from these 50 replications are following: **power** = 0.99 (0.01), **observed FDR** = 0.03 (0.03), **test error by LassoNet** = 1.048 (0.101), **test error by SciDNet+RT** = 0.710 (0.002). To reduce computational complexity, after the screening, the bootstrapped LassoNet can be run in a parallel loop and we conduct all the experiments in a high-performing computing facility with Intel(R) Xeon(R) Platinum 8260 CPU @ 2.40GHz and 4 Tesla V100S. **The codes are available at an anonymous repository (https://anonymous.4open.science/r/SciDNet-3CA8)**.

B.0.6 Important clusters of gene discovered by SciDNet

B.0.7 For CCLE dataset

The following Table S4 presents all the selected clusters of genes by SciDNet for the five anticancer drugs considered. The genes in a single cluster are mentioned in the "{}". Previous research on this gene-expression data has revealed several genes as biologically associated with the corresponding drugs. SciDNet successfully discovers these genes as the top-most important gene associated with the drugs. In Table S4, the selected genes which are confirmed by previous domain research, are highlighted and corresponding references are mentioned in column 3.

Table S5 Selected clusters of genes by SciDNet applied in the riboflavin gene data example

Cluster No.	Genes selected
1	EPR_at, IOLD_at, KAPB_at, PROJ_at, RPLQ_at, UREA_at, YCGB_at, YCGM_at, YCGN_at, YCSN_at, YCGO_at, YCGT_at, YDBM_at, YHXA_at, YKZC_at, YOAB_at, YPJB_at, YUSX_at, YVFH_at
2	COMX_at, CSPC_at, HAG_at, MPR_at, YBDL_at, YDBM_at, YHCB_at, YJFB_at, YHFS_at, YOAB_at, YODF_at, YOAC_at, YONU_at, YOTL_at, YQKI_at, YQZH_at, YTEI_at, YUSV_at
3	HIT_at, KATX_at, LICH_at, NASA_at, OPUCB_at, PHRG_i_at, PHRK_at, ROCB_at, ROCR_at, SACB_at, SPOIIE_at, TMRB_at, YACN_at, YBBJ_at, YBGB_at, YCBF_at, YFKJ_at, YHCS_at, YHXA_at, YJBF_at, YLBA_at, YLOU_at, YPUI_at, YQGY_i_at, YUKE_at, YVYD_at, YXLJ_at, YXZF_at
4	APPA_at, BGLS_at, ccpB_at, MMR_at, SIGY_at, SOJ_at, TREA_at, YBGB_at, YDGF_at, YOPR_at, YQEB_at, YVCI_at, YVDR_at, YWBG_at, YWDE_at, YWFM_at, YXBB_at, YXIL_at, YXIO_at, YXIQ_at, YXJA_at, YXJN_at, YXLC_at, YXLD_at, YXLE_at, YXLF_at, YXLG_at, YXLJ_at, YXZF_at, YYBF_at
5	LYTD_at, SQHC_at, XKDE_at, YFIG_at, YFIH_at, YFII_at, YFNC_at, YHDV_at, YIST_at, YJGA_at, YTCP_at, YTMP_at
6	YCDH_at, YCDI_at, YCEA_at, YCIA_at, YCIB_at, **YCIC_at**, YDAR_at, YHZA_at, YRPE_at, YTGA_at, YTGB_at, YTGC_at, YTGD_at, YTIA_at, YVQH_at
7	OPUBD_at, PHRE_at, SIPS_at, YBFF_at, YDEM_at, YNAB_i_at, YNAC_at, YNEK_at, YOBF_at, YOKG_at, YONX_at, YOPA_at, YOPR_at, YOTL_at, YPBB_at, YQZH_at, YRDA_at, YRKK_at, YRKL_at, YTGB_at, YUXI_at, YWCE_at, YWQK_at, YYDB_at, YYDF_i_at
8	ARGB_at, ARGC_at, ARGD_at, ARGJ_at, CARA_at, CARB_at
9	PROJ_at, RPLF_at, RPLJ_at, RPLL_at, RPSN_at, RPSP_at, YLQC_at

B.0.8 For Riboflavin dataset

The Riboflavin production dataset contains a much more complicated correlation structure than the CCLE data, see Figure 1 in the main manuscript for a visual illustration. As a result, SciDNet has produced a much larger cluster of genes compared to the cluster sizes from the CCLE dataset. For example, the average cluster size for CCLR and Riboflavin datasets is respectively 2.5 and 17.78. The following Table S5 shows the 9 selected clusters of genes selected by SciDNet while the FDR is controlled at $q = 0.15$. Additionally, SciDNet discovered the gene *YCIC_at* as one of the expressive genes related to riboflavin production which was identified by Bühlmann et al. (2014) as a causal gene in this context.

CHAPTER 4

DEEP LEARNING-AIDED FEATURE SELECTION FOR COGNITIVE RESERVE WITH HIGHLY CORRELATED TRACTOGRAPHY DATA

4.1 Introduction

4.1.1 A motivating case study and building the statistical framework

The common observation that some older adults maintain normal levels of cognitive function despite apparent brain pathology is referred to as cognitive reserve (Stern et al. (2020)). However, the neurobiological substrates, mechanisms, and potential neuroimaging markers of such resilience remain poorly understood. The immense system of structural connections between brain regions, which constitutes subcortical white matter (WM) is one promising source of neuroimaging correlates of cognitive reserve (Chang et al. (2021); Wang et al. (2020a)). Recent studies in Alzheimer's disease, inflammatory, and vascular pathologies report that WM pathways are both enhanced by cognitive and environmental enrichment and are among the earliest systems to be negatively impacted by these pathologies (McPhee et al. (2019); Uddin (2021)). This recent evidence highlights the brain's WM fiber pathways as a potential biological correlate of cognitive reserve.

Diffusion magnetic resonance imaging (DMRI) methods characterize microstructural tissue organization based on the differential movement of water molecules in the presence of barriers (Beaulieu (2002)). In the brain, the hydrophobic myelin sheath that surrounds long neuronal axons serves as such a barrier, making it useful for characterizing subcortical WM tissue microstructure. Sampling biologically informative rates of diffusion in multiple spatial directions permits representing diffusion using summary measures, such as a tensor. The tensor eigenvalues can be averaged to provide a measure of mean diffusivity (MD), or used to calculate parameters such as fractional anisotropy (FA), which reflects the uniformity of diffusion. However, diffusion tensor parameters only afford valid representations of WM microstructure in MRI volumetric pixels (i.e., voxels) that include a single spatial orientation of WM fibers. Because an estimated 60-90% of WM voxels include multiple orientations of fiber populations (Vos et al.

(2011)), the standard methods for representing WM microstructure using the metrics estimated from tensor decomposition is at best questionable.

In contrast to voxel-level scalar values like FA, DMRI tractography represents WM organization as continuous paths between adjacent voxels, or streamlines, computed using directional information from tensor or other diffusion models. These tractography streamlines most commonly serve as masks to delineate anatomically specific tracts for sampling from diffusion tensor (e.g., FA) or other image modalities in the same image space. Most extant applications of tractography methods collapse tract segments, having only a single mean diffusion magnetic resonance imaging (DMRI) metric and variance estimate for each tract and each subject, potentially ignoring a rich anatomical variation along the tracts. Consequently, such summarization reduces the sensitivity and specificity of this neuroimaging technique. As a solution, the along-tract workflow which measures the DMRI metrics at several vertices spread evenly along the tract has been proposed (Colby et al., 2012; Wasserthal et al., 2018a); however, these methods can only model variation within individual fiber tracts as the response.

The University of Michigan Memory and Aging Project (UM-MAP) is a clinical cohort study of aging and dementia that includes data from multi-modal MRI neuroimaging, including DMRI scans. These afford a more detailed study of the distribution of estimable DMRI metrics along the WM tracts in whole brain tractography data. In addition, UM-MAP participants include cognitively unimpaired older adults as well as those diagnosed with mild cognitive impairment (MCI) or dementias of the Alzheimer's type and other etiologies. In our study, we are interested in determining *the most effective and representative parts of WM tracts to characterize cognitive reserve*. To achieve this goal, we combine the DMRI tractography (Wasserthal et al., 2018a) with along tract workflow (Colby et al., 2012) to systematically quantify the WM microstructure through specified DMRI metric (e.g. fractional anisotropy (FA), or spherical harmonic peak amplitudes (sh-peaks)) observed at spatially equidistant vertices spread along the major WM tracts. In this work, we mainly considered the spherical harmonic peak values (sh-peaks, Bastiani et al. (2017)) sampled from the fiber orientation distributions which represent the diffusion maxima for a specific orientation of white matter fibers within a voxel. Thus, unlike FA, which integrates the diffusion signal from multiple directions, sh-peak values are not confounded by crossing fibers. Focusing on $p_t = 50$ major WM tracts in human brains, our main objective is to determine which parts of these WM tracts are associated with some neurogenetic phenotype like the cognitive reserve. To fix ideas, suppose, $y \in \mathbb{R}^n$ is a vector of observed cognitive reserve values for n subjects in the cohort study and $X \in \mathbb{R}^{n\times p}, p = p_t p_v$ is, column by column, the matrix of DMRI metrics observed at p_v vertices of each of the p_t tracts. Formally, we are interested in selecting important features to characterize cognitive reserve among the p potential features using the nonparametric regression: $y = g(x_1, x_2, \ldots, x_p) + \epsilon$; where g is an unknown link function and $\epsilon \sim N(0, \sigma^2 I_n)$ is the random noise. The feature selection with DMRI tractography data from UM-MAP faces additional methodological complications beyond those encountered in typical feature selection problems. These issues are further discussed in Section 4.1.2 and empirically evaluated in Section 4.2. Key points of our general analytic strategy are discussed in Section 4.1.3.

4.1.2 Statistical challenges and related literature

The analysis of tractography data from the UM-MAP study is complicated due to the high dimensionality of data sampling possible from streamlines, especially considering the more limited sample size. Furthermore, the strong associations within and between tract-level measurements and the potentially complicated functional relationship between these tract-measured data and phenotype that precludes the linearity assumption create further challenges. The extant methods in the statistical literature are either too restrictive in view of the modeling assumptions or rely on nontrivial extensions to account for the complex association among the high-dimensional predictors. Failure to adequately address these data complications may result in a higher rate of false discoveries. Although Chapter 3 extensively discusses these issues, we reiterate their significance here, focusing on the aforementioned motivating example based on tractography data.

Reproducible high-dimensional non-linear feature screening The feature selection problem, under the linear model assumption, has been extensively studied over the last twenty years under various data setups. Popular algorithms include the Lasso, Elastic net, SCAD, and MCP; a full review of existing methods can be found elsewhere (e.,g., Fan and Lv (2010). Despite their successes, feature selection algorithms under the linear model assumption tend to have poor performance in settings where the underlying functional form deviates from the linearity. To highlight this limitation, we conducted a basic implementation of the prediction optimal elastic net (Zou and Hastie, 2005) and found that this method results in very poor performance ($\sim 20\%$ coefficient of determination), hence necessitating the need to entertain non-linear association approaches. As we discussed in detail in Chapter 3, the Artificial Neural Network (ANN) models, which relax the linearity assumption, are well known for efficiently approximating complicated functions. This key feature of ANN models has motivated the use of Deep Learning (DL) models for feature selection in recent years. Popular examples include Deep Feature Selection (Chen et al., 2021), DeepPink (Lu et al., 2018), and SurvNet (Song and Li, 2021). Despite their popularity, many of the existing DL algorithms can be overly sensitive to noise. Recent works have shown that a small amount of added noise can drastically change the importance of variables in the model (Ghorbani et al., 2019). As a solution, reproducible variable selections with some form of error control have been advocated. In this realm, the control of the False Discovery Rate (FDR), first proposed by Benjamini and Hochberg (1995), has emerged as a major approach due to being less conservative and more powerful than the Family Wise Error rate (FWER), especially in large-scale multiple testing problems. However, estimating the expectation in the FDR poses a unique challenge for the model-free variable selection methods, and has been investigated in various outlets. Much of the existing approaches have focused on p-values as feature importance in a multiple testing context; see Xia et al. (2017); Li and Barber (2018); Lei and Fithian (2018) for more details. However, generating p-values for DL models that are interpretable is proven to be complicated, prompting researchers to seek alternatives. One such alternative is the knockoff method proposed by Candès et al. (2018), which is essentially a model-free variable selection algorithm with provable FDR control, assuming a well-specified predictor's distribution. Hence, to generate the knockoff, it is necessary to specify or estimate a high-dimensional predictor's distribution, which in many practical settings may be an overwhelming hurdle to overcome. Unlike in genetics studies where the linkage disequilibrium assumption is often made to describe a consistent dependence pattern between the alleles at polymorphisms (Sesia et al. (2018)), with tractography data such a prior correlation pattern cannot be imposed. For example, figure 4.1 shows that for some tracts, the correlations among spatially distant vertices along a tract may be much higher reflecting the spatial symmetry of the white matter representation. Hence, for the tractography data, any model-specific knockoff generation algorithm would be highly inefficient. Recently, DL-based flexible knockoff-generating algorithms have also been proposed (Jordon et al., 2019; Romano et al., 2020); however, these methods are often trained with large samples in big-n-small-p data settings. It is unclear how these methods will perform when the sample size n is significantly smaller than the dimension of the covariates p, as in the current tractography dataset with $n = 210$ subjects and $p \simeq 5000$ high-dimensional DMRI metrics. Additionally, in the context of hierarchical testing, several other competing algorithms have also been proposed including *SUSIE* (Wang et al., 2020b), *KnockoffZoom* (Sesia et al., 2019). While the knockoff-based procedures have the limitation of generating knockoffs from an unknown complex distribution with a comparatively small sample size, most of the non-knockoff-based methods lack their applicability in non-linear setups as they typically depend on p-values.

Highly correlated predictors measured intermittently along the tracts Basic exploratory analyses of UM-MAP data show that the DMRI metrics are highly correlated, with most of the local pairwise correlations exceeding 0.95 and even 0.99 for some tracts (see Figure 4.1). This is a well-known complication for feature selection problems in many modern data sets arising in genetics and imaging studies. This extreme association is problematic for a typical variable selection analysis as the highly correlated predictors become almost indistinguishable in view of the phenotype. Ignoring this extreme association and conducting a variable selection that

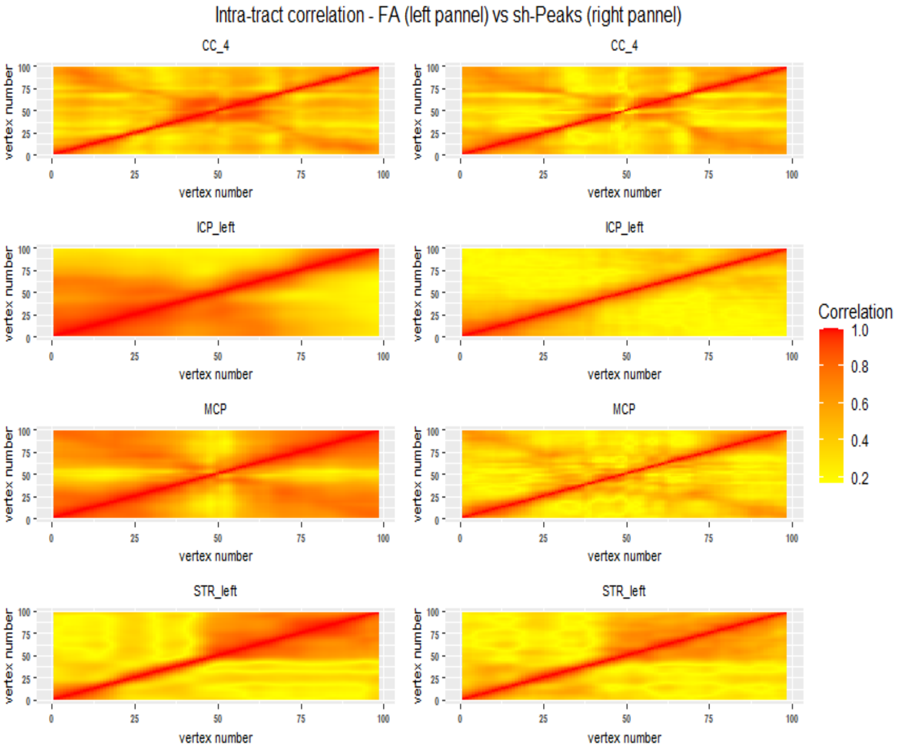

Figure 4.1 Correlation heatmap of the sh-peaks metric observed along six major WM tracts (1) CC-4, (2) ICP-left, (3) MCP, and (4) STR-left (showing distinguishable correlation structure.

focuses solely on individual variables is meaningless as it does not account for the uncertainty due to highly correlated predictors. One can argue that it would be unjustifiable to specifically claim that one of the highly correlated predictors is associated with the response. Alternatively, an approach that accounts for high correlations is the group or cluster selection method, which has been applied in genetics and many other applications. A complication of this approach, however, is the ambiguity on how to define the clusters and subsequently perform the selection for the clusters. In recent works, Candès et al. (2018) considered a heuristic approach where the predictors are clustered according to their pairwise correlations. This method works reasonably well in general but may generate larger clusters, rendering the choice of the representative feature of each cluster nontrivial. And more importantly, larger clusters may not be very informative from a substantive viewpoint. Cluster-based variable selection methods using conditional associations which generate smaller clusters have also been proposed but their application to tractography data is critically lacking.

In addition to the complexities related to high correlation, although the tractography streamlines provide a continuous representation of WM, the DMRI metrics were measured only at a few equidistant locations on each tract. This poses the generic problem of 'missing values' in statistical research as is unclear whether the observed measurements represent the true locations. Without a better understanding of the true DMRI metrics locations, determining the true signal location is at best problematic. This problem is exacerbated in neuroscience research where low sensitivity is real for many feature selection exercises.

4.1.3 The general analytic strategy

The study of WM as a potential marker of cognitive reserve using UM-MAP data is hampered by various methodological and conceptual limitations in WM measurement and statistical modeling. The above discussion highlights the need to develop a methodology capable of accommodating variation across vertices while accounting for the high correlation both between and within tract value measurements. The present study sought to address and overcome these challenges by providing a novel quantitative neuroimaging approach for modeling WM pathways as a neural marker for predicting cognitive reserve. Specifically, it will help determine the most important regions of human WM tracts that are associated with cognitive reserve as well as formulate a general workflow for high-dimensional nonlinear variable selection with highly correlated predictors.

Our methodological contribution relies essentially on a novel screening and cleaning method for the reproducible high-dimensional nonlinear feature selection with highly correlated predictors. As the spatial positions are inherently continuous while the DMRI metrics are observed at some discrete vertices on the tract, we consider here the cluster-level discovery of the positions by leveraging the high association among the vertices. To be concrete, let $C_{causal} = \{Z_1^c, Z_2^c, \ldots, Z_s^c\}$ denote the unknown true set of causal locations along the tracts that harbor the variability which influences the phenotype of interest. We also denote by $C_{observed} = \{Z_1^o, Z_2^o, \ldots, Z_p^o\}$ the locations at which the DMRI metrics are observed. Although there is no guarantee that these metrics are observed at the specific positions in C_{causal}, there are two possibilities for observing some true causal location $Z_j^c \in C_{causal}$: (1) it is observed; i.e. $Z_j^c = Z_k^o$ for some k; or (2) although Z_j^c is not observed, it is fair to assume that a close proxy location Z_k^o is observed for some k due to the strong local dependency along each tract. For this reason, we seek to discover some clusters of locations that can jointly serve as a good proxy for the locations in C_{causal}. With this in mind, we set our target to uncover the subset $S_0 \subset \{1, 2, \dots, p\}$ such that, conditional on features in S_0, the response Y_i is independent of features in the complement set S_0^c. In other words, $S_0 = \{k : f(y|X) \text{ depends on } X_k\}$, where X_k is the realization of the DMRI metric at the observed location Z_k^o and $f(y|X)$ is the conditional density of y given X. The members of the subset S_0 can be interpreted as either the true causal location or the closest proxy of a causal location. To achieve this goal, we implement our proposed method ScIDNet in Chapter 3, where we divide our method into two parts: **Screening** and **Cleaning**. The screening step is a dimension reduction step. We screen out most of the null variables and select an active set of variables $\hat{S}_n$ which will surely contain all the proxy variables needed to cover the causal set C, implying the sure screening property. To reduce the high amount of correlation in the active set, by exploiting their conditional dependency structure, we divide the active variables $\{X_j : j \in \hat{S}_n\}$ into $p_c (<< p)$ spatially connected non-overlapping clusters: $C_1, C_2, \dots, C_{p_c}$ and select an appropriate cluster representative from each cluster. In the cleaning step, we develop an estimate of the number of false discoveries using the resampling technique followed by developing a surrogate of the FDR. Finally, by controlling the surrogate FDR, we select some clusters of highly correlated predictors. Here *true discovery* implies that the selected cluster can serve as a good proxy for at least one element in the true causal set C.

Our comprehensive empirical study utilizing DMRI metrics provides compelling evidence for the effectiveness of the proposed method as a proof of concept, as it achieves higher power and controls the false discovery rate (FDR). While achieving theoretically guaranteed FDR control within a deep learning framework remains an active area of research, our study paves the way for further theoretical investigations into the method's generalizability and broader validity by utilizing SciDNet's theoretical guarantees on FDR control. The proposed approach stands out by not relying on strict modeling assumptions between the response and features and being completely independent of p-values, distinguishing it from other state-of-the-art methods. Consequently, it facilitates a deeper understanding of micro-structural relationships in the human brain within the context of dementia and beyond.

After providing a comprehensive discussion of assumptions and algorithms in Chapter 3, we now proceed directly to the numerical analysis section involving tractography data. The notations used throughout this section remain consistent with those introduced in Chapter 3. We begin by presenting an extensive simulation study that specifically focuses on the tractography data in Section 4.2. Subsequently, we showcase the application of our methodology to the UM-MAP Tractography dataset in Section 4.3. Finally, we conclude with a summary of our findings and outline future research directions in Section 4.4.

4.2 Numerical Illustrations

In this section, we extensively investigate the finite sample performance of the proposed method SciDNet, along with several other state-of-the-art FDR controlling algorithms, using a simulation study. Firstly, Section 4.2.1 verifies the suitability of the nonparanormal transformation for the tractography data, thereby confirming our assumption that the distribution of the features belongs to the non-paranormal family. Secondly, in Section 4.2.2, we demonstrate how major FDR controlling approaches fail in the presence of severe multicollinearity in an ultrahigh-dimensional setting. Additionally, Section 4.2.3 presents the performance evaluation of the proposed method in a synthetic setting, where DMRI-metrics (e.g., sh-peaks) from the tractography data serve as predictors, and the response is generated using a non-linear function. We utilize two metrics to assess the feature selection performance of the algorithms: (1) Power = $\frac{|\hat{D}_n \cap S_0|}{|S_0|}$ and (2) empirical FDR = $\frac{|\hat{D}_n \cap S_0^c|}{|\hat{D}_n|}$.

4.2.1 Checking the non-paranormal transformation on tractography data

The non-paranormal transformation Liu et al. (2009) plays a crucial role in the proposed method SciDnet, particularly in its screening step. This step utilizes clustering to capture the conditional dependency structure after dimensional reduction, ensuring the sure independence

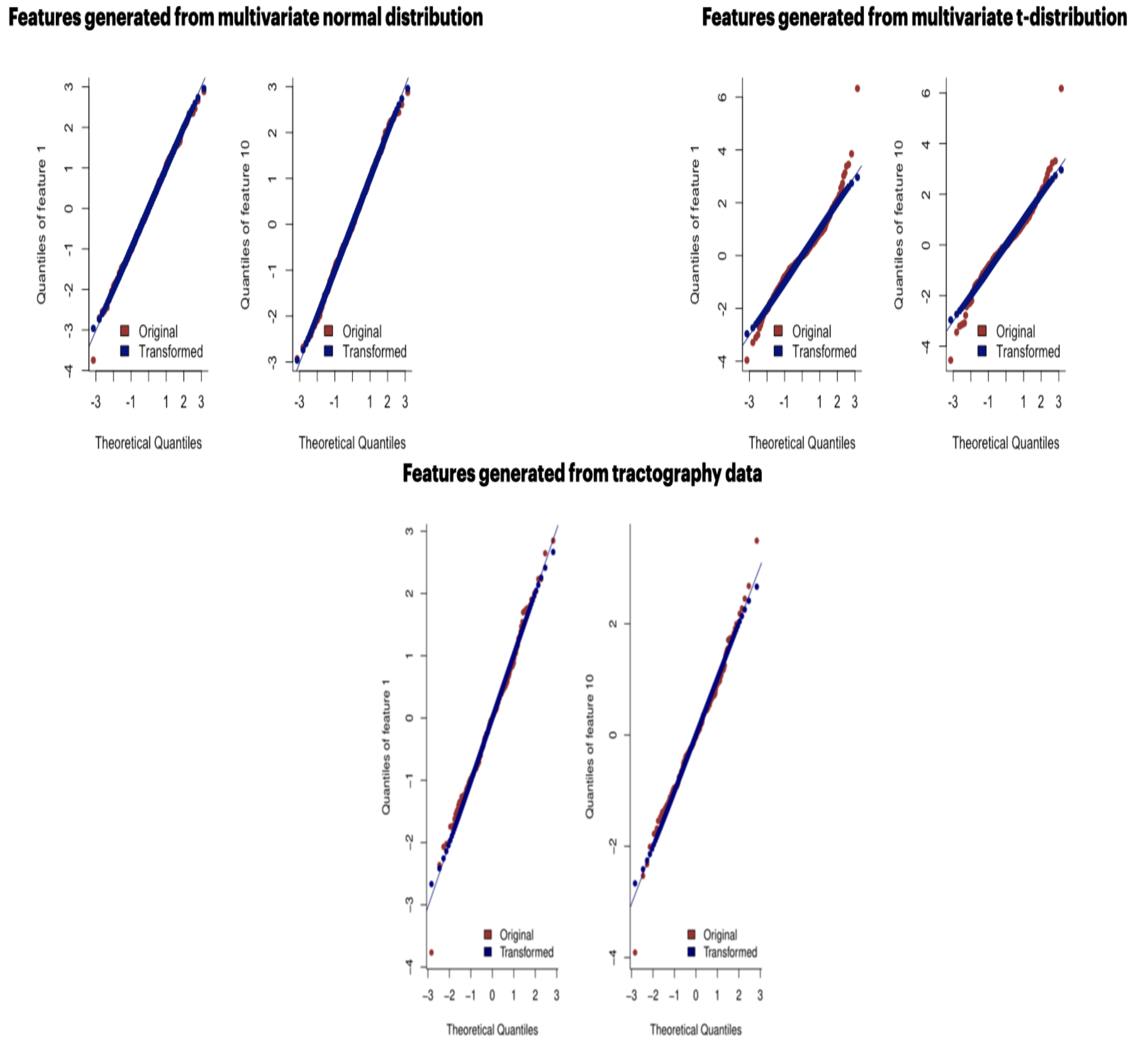

Figure 4.2 Illustration of nonparanormal transformation using simulated and the tractography dataset.

screening property (3.2.2.1) when the features belong to a non-paranormal family. Therefore, before applying SciDnet to the tractography data, it is essential to evaluate the effectiveness of the nonparanormal transformation on this dataset.

To assess the efficiency of the nonparanormal transformation, we employ the following approach. We generate two sets of features: $X_G^{(i)} \sim N(0, \Sigma^{p\times p})$ and $X_t^{(i)} \sim t_p(5)$, where $i = 1, 2, \ldots, n$, from a multivariate Gaussian and a t_5 distribution, respectively, while maintaining a correlation structure defined by $\Sigma_{ij} = 0.95^{|i-j|}$. We set $n = 220$ and $p = 5000$ to mimic the dimensions of the tractography data. Consequently, we obtain the original DMRI tractography data $X_{tract} \in \mathscr{R}^{220\times5000}$. Next, we apply the nonparanormal transformation to the features in X_G, X_t, and X_{tract}.

To evaluate the effectiveness of the transformation, we present Figure 4.2, which displays the QQ-plots of the transformed features compared to the Gaussian distribution. The figure indicates a favorable alignment of the transformed features with the Gaussian distribution, demonstrating a similar pattern to the simulated settings. For ease of demonstration, we only depict the results for the 1st and 10th features. This experimental analysis verifies the validity of employing the nonparanormal transformation in the context of tractography data. Additionally, it confirms that the clusters generated by SciDNet effectively capture the conditional dependencies among the features.

4.2.2 Performance of existing feature selection methods in the presence of high multicollinearity

The predictors are first generated from $X_i \sim N_p(0,\Sigma), i = 1,2,\ldots,n$, for multiple combination of (n, p) and the covariance matric Σ is chosen as a toeplitz matrix with $\Sigma_{ij} = \rho^{|i-j|}, \rho =$ 0.1, 0.5, and 0.9. Under simplistic setting, the response y is generated from $y = x_S\beta_S + \epsilon$, $S = \{5,10,\ldots,50\}, |S| = 10$, with β_S generated from $N(\beta_0, 0.1)$ independently and $\beta_{S^c} = 0$. The random noise $\epsilon \sim N(0,1)$. We focus on the Model-X knockoff (Candès et al., 2018), SurvNet (Song and Li, 2021), and DeepPINK (Lu et al., 2018). For a more rigorous analysis, we consider two different versions of Model-X knockoff - (1) Model-X-Estimated, where the knockoffs are generated using an estimated multivariate Gaussian distribution and (2) Model-X-True, where the knockoffs are generated using the true data generating multivariate gaussian distribution mentioned above. For the knockoff generation, we consider the equicorrelated construction using the R package knockoff: The Knockoff Filter for Controlled Variable Selection. To implement the SurvNet and DeepPINK, we use the codes mentioned in the respective papers Song and Li (2021); Lu et al. (2018). We set $q = 0.15$ as the FDR control threshold.

Table 4.1 reveals several interesting characteristics. Both Model-X-Estimated and Model-

X-True maintain the power-FDR balance under a low correlation setup. However, with higher multicollinearity, Model-X-Estimated fails to control the FDR below the specified threshold while the Model-X-True controls the FDR efficiently. This disparity indicates Model-X procedure induces inflation in false discoveries if the knockoffs are not generated properly under a 'difficult' situation. As expected, the DL-based algorithms, such as SurvNet and DeepPINK work much better in big-n-small-p and low correlation setups but typically fail in other cases, indicating their reduced effectiveness in ultrahigh dimensional data with small sample sizes.

Table 4.1 Empirical power and observed FDR of various feature selection algorithms with standard error in parentheses

			$\beta = 2$			$\beta = 4$		
	(n, p)		ρ =0.1	ρ =0.5	ρ =0.9	ρ =0.1	ρ =0.5	ρ =0.9
Model-X-Estimated	$(400, 1000)$	Power	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
		FDR	0.13 (0.17)	0.12 (0.12)	0.27 (0.18)	0.11 (0.19)	0.20 (0.18)	0.27 (0.20)
Model-X-True	$(400, 1000)$	Power	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
		FDR	0.08 (0.13)	0.09 (0.12)	0.14 (0.17)	0.12 (0.14)	0.11 (0.13)	0.08 (0.12)
SurvNet	$(400, 1000)$	Power	0.27 (0.20)	0.32 (0.22)	0.35 (0.24)	0.49 (0.24)	0.52 (0.28)	0.58 (0.29)
		FDR	0.31 (0.36)	0.53 (0.30)	0.59 (0.23)	0.21 (0.21)	0.53 (0.18)	0.60 (0.17)
	$(10000, 60)$	Power	0.99 (0.05)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
		FDR	0.20 (0.15)	0.80 (0.02)	0.78 (0.07)	0.14 (0.11)	0.80 (0.02)	0.56 (0.32)
DeepPINK	$(400, 1000)$	Power	0.01 (0.02)	0.03 (0.04)	0.00 (0.00)	0.03 (0.04)	0.01 (0.03)	0.02 (0.05)
		FDR	0.23 (0.40)	0.35 (0.42)	0.33 (0.47)	0.45 (0.44)	0.24 (0.41)	0.24 (0.40)
	$(10000, 60)$	Power	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
		FDR	0.18 (0.04)	0.29 (0.13)	0.25 (0.11)	0.17 (0.01)	0.24 (0.12)	0.24 (0.12)

4.2.3 Performance of the proposed method on synthetic data

Prior to implementing the proposed method on the cognitive reserve, we conduct a synthetic study to understand its effects on the DMRI tractography data. As mentioned in the section 4.1.1, we consider $p_t = 50$ major WM tracts for $n = 220$ subjects. We measure the DMRI-metric named spherical harmonic peak amplitudes (sh-peaks) at spatially equidistant $p_v = 98$ vertices on each of the tracts. Hence the feature vector $x^i \in \mathbb{R}^p$, $p = p_t p_v = 5000$, contains all the sh-peaks values measured at each of the vertices of the tracts for the i-th subject, $i = 1, 2, \dots, 220$.

We further simulated the response $y^{(i)}$ from $x^{(i)}$ using some nonlinear model. For this purpose, single Index models are straightforward yet flexible examples of nonlinear models where the response is related to a linear combination of the features through an unknown nonlinear, monotonic link function, i.e. $y^{(i)} = g(x^{(i)\prime}\beta) + \epsilon$. Here we choose the following two link functions as our data-generating process: (1) M_1: $g(x) = \frac{x^3}{10} + 3\frac{x}{10}$ and (2) M_2: $g(x) = log(10 + e^x)$. Similar to Section 4.2.2, the coefficients $\beta \in \mathbb{R}^p$ is sparse with the true nonzero locations $S = \{50, 100, 150, 200, 250\}$, where $\beta_{S^c} = 0, \beta_S \sim N_S(u\beta_0, 0.1)$ with $u = \{\pm 1\}^S$. The value of β_0 is set as $\beta_0 = 1.5$. The random error $\epsilon \sim N(0, \sigma^2)$, maintaining the decreasing signal-to-noise ratio as $snr = 7:3$ and $3:7$. All the performance metrics are based on 50 Monte Carlo replications.

Table 4.2 Power and empirical FDR of the proposed method with standard error in parentheses from the synthetic study

$g(\cdot)$	snr		Screening	Screening + LassoNet	Screening + Model-X Knockoff (Linear)	Screening + Model-X Knockoff (RF)	Screening + proposed Cleaning (SciDNet))
M$_1$	7:3	Power	1.000 (0.00)	0.996 (0.03)	0.870 (0.34)	0.921 (0.39)	0.984 (0.05)
		FDR	0.992 (0.00)	0.525 (0.14)	0.178 (0.19)	0.182 (0.12)	0.0.098 (0.10)
		n_var	624 (0.00)	419.09 (44.56)	286.85 (121.50)	291.23 (37.41)	281.06 (60.59)
		n_clust	43.22 (2.60)	11.96 (5.91)	5.80 (2.87)	6.64 (0.62)	5.50 (0.73)
	3:7	Power	1.000 (0.00)	0.992 (0.04)	0.788 (0.40)	0.841 (0.37)	0.970 (0.08)
		FDR	0.992 (0.00)	0.619 (0.14)	0.174 (0.24)	0.193 (0.21)	0.004 (0.02)
		n_var	624.00 (0.00)	405.56 (53.33)	236.90 (131.70)	248.76 (158.57)	285.54 (53.83)
		n_clust	49.44 (5.19)	14.94 (6.04)	5.78 (4.03)	6.92 (4.89)	4.87 (0.45)
M$_2$	7:3	Power	1.000 (0.00)	1.000 (0.00)	0.776 (0.42)	0.825 (0.29)	0.996 (0.03)
		FDR	0.992 (0.00)	0.656 (0.08)	0.176 (0.19)	0.183 (0.26)	0.021 (0.06)
		n_var	624 (0.00)	441.32 (48.80)	247.22 (140.54)	262.50 (88.01)	274.92 (45.48)
		n_clust	43.58 (4.01)	15.32 (3.89)	5.38 (3.41)	7.98 (2.75)	5.08 (0.34)
	3:7	Power	0.991 (0.04)	0.960 (0.09)	0.427 (0.47)	0.436 (0.25)	0.533 (0.24)
		FDR	0.992 (0.00)	0.762 (0.10)	0.106(0.21)	0.153 (0.17)	0.026 (0.09)
		n_var	624.00 (0.00)	357.73 (92.57)	98.93 (113.80)	99.46 (125.77)	115.04 (62.58)
		n_clust	49.18 (8.63)	24.33 (11.53)	3.16 (4.15)	3.22 (1.43)	2.71 (1.16)

Table 4.2 demonstrates an ablation study where we compare the following four methods: (1) **Screening only**, (2) **Screening + Cleaning with LassoNet**, (3) **Screening + Cleaning with knockoff** with q=20%, and (4) the proposed method **Screening + Cleaning with resampled LassoNet** with q=10% and 20%, where q= the error-control cutoff. Focusing on the cluster-level discoveries, two additional performance metrics are included here: (1) n_var, the total number of features in the selected clusters, and (2) n_clust, the total number of selected clusters. Lower values of n_var and n_clust indicate better support recovery for a group-level feature selection method. In this context, Table 4.2 empirically consolidates several interesting characteristics: (a)

The screening step maintains the sure screening property and thereby selects a slightly bigger set of features resulting in high power and high FDR which necessitates further cleaning; (b) All the cleaning steps aim to reduce the FDR while maintaining the power of the screening step; (c) The proposed method achieves the best performance (in terms of the power-FDR tradeoff) by effectively using the added information from the resampling. Although for a high noise setup, the power is comparatively low, it still maintains the nominal rate of false discoveries. This empirical study sets the stage for the real data analysis in section **??** where the simulated y is replaced by the cognitive reserve values from the UM-Map study as the outcome variable.

4.3 Real Data Analysis - UM-MAP Tractography Data

As discussed in Section 4.1.1, the WM fiber tract segmentation facilitates in rigorous analysis of WM microstructural quantities and their relation to the cognitive performance of the human brain. In order to understand which segments of human WM tracts are associated with the cognitive reserve, the proposed method has been implemented in Michigan Alzheimer's Disease Research Center (MADRC) dataset. We considered 50 major tracts that shield almost 90% of human WM (Wasserthal et al., 2018a); a full list of these tracts is relegated to Appendix 4.4.

4.3.1 Results and Selected Tracts

We consider the cognitive reserve values calculated for $n = 220$ subjects as the outcome variable. Further, for the predictors, the sh-peaks metric is observed at $p_v = 98$ vertices spread in a spatially equidistant fashion along each of the $p_t = 50$ major WM tracts. Hence the feature space becomes of dimension $p = p_t p_v = 4900$. With this setting, we implement the proposed screening and cleaning method. In the screening step, we set the size of the active set, $|\hat{S}_n| = 7\left[\frac{n}{log(n)}\right] = 657$ and the other 4143 vertices are rejected by the HZ-test, thereby discarded from the remaining analysis. The active set $\hat{S}_n$ is further divided into 137 disjoint clusters of spatially connected and highly correlated vertices. From each cluster, the proper representative has been chosen by the maximum value of the HZ-statistic among the cluster members, following which the set of cluster representatives $\tilde{S}_n$ has been created. Next in the cleaning step, the LassoNet has been implemented parallelly for $B = 1000$ bootstrap replications. Each time as described in section

3.2.3, the L_1 penalty parameter λ has been chosen in a sequential manner as $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_r$; where for λ_1 all the active predictors are present in the model. Then with the gradual increase of λ, one after another the representatives will get eliminated from the model and finally, λ_r will produce the null model with no predictors. Thus the proposed error estimate $F\hat{D}R$ is constructed by combining the variable importance measured over all the regularization paths for all the bootstrap replications. Controlling the estimated error rate by the threshold $q = 0.15$, the final set of clusters $\hat{D}_n$ is discovered.

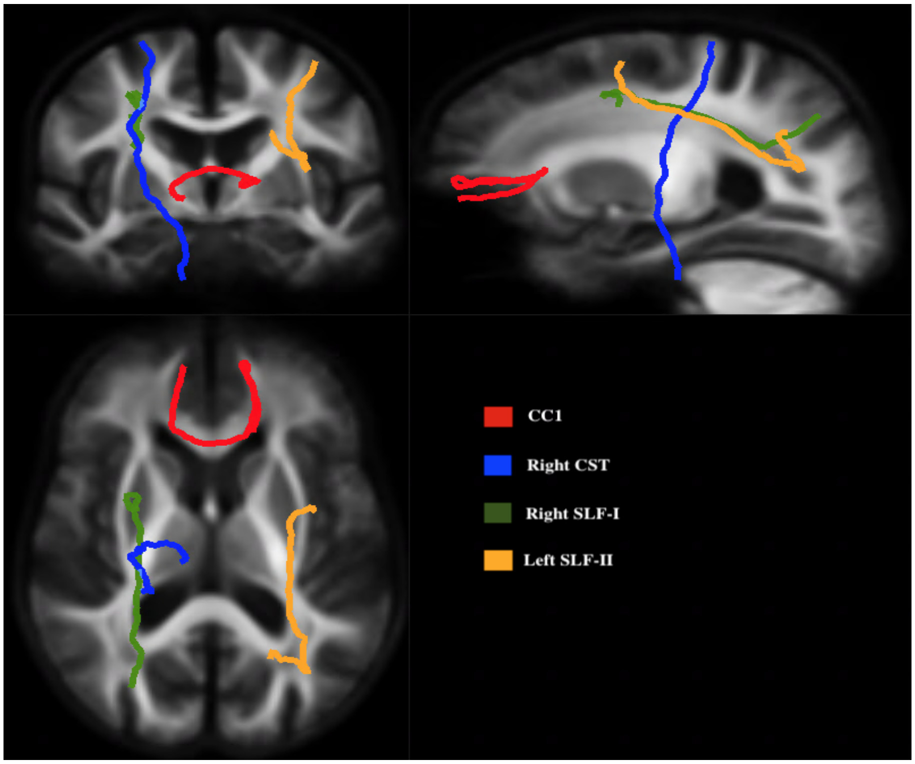

Figure 4.3 Illustration of down-sampled diffusion tractography for streamlines that included significant clusters.

Demonstration of the selected tracts The proposed method discovers several spatially connected parts of the following four major tracts: (1) the first (i.e., ventral genual) corpus callosum fiber bundle that serves as the the inferior aspect of the forceps minor (**CC1**), (2) right cortical spinal tract (**right CST**), (3) First subdivision of the right hemisphere superior longitudinal fasciculus (**right SLF-I**), and (4) Second subdivision of the left hemisphere superior longitudinal fasciculus (**left SLF-II**). Figure 4.3 demonstrates these selected tracts by showing three different orientations of the brain: coronal (top left) or as if facing a mirror, sagittal (top right) or viewed from the side, and axial (bottom left) or viewed from above. In order to display the mean trajectory of the tracts, the TractSeg tractometry algorithm (Wasserthal et al., 2018a) reduces multiple streamlines (e.g., 100 to 1000) originally generated for a given WM fiber tract down to a single representative streamline. Figure 4.3 shows four separate streamlines representing anatomically specific white matter fiber bundles. Each streamline is a curvilinear 3-dimensional object composed of thousands of linked vertices; however, we note that the illustration depicts 3-dimensional streamlines against 2-dimensional cross-sections of the brain. Data are shown in a radiological orientation where the left side of the image corresponds to the anatomical right and vice versa. Additionally, we conducted a sensitivity analysis by implementing the proposed method on several subgroups of the whole data, taking both sh-peaks and FA as the DMRI metric simultaneously. Table 4.3 shows the clusters of the vertices on the selected tracts under different subgroup selections and DMRI metric choices (sh-peaks and FA).

These results reveal clusters of vertices in WM streamlines for the left and right superior longitudinal fasciculus (SLF), and the right hemisphere corticospinal tract (CST) where higher sf-peak value significantly predicted higher cognitive reserve. In contrast, lower sf-peak values in the first (e.g., most anterior and ventral) subdivision of the corpus callosum (CC) predicted greater reserve. The SLF is commonly associated with working memory (Koshiyama et al. (2020)), whereas CST is believed to be more related to motor function, albeit less strongly (Min et al. (2014)). Thus, more maintenance of white matter in SLF-I and right CST predict greater preservation of memory, despite atrophy of medial temporal lobe gray matter regions. The CC1 fiber bundle is the inferior aspect of the genu which connects the left and right ventral prefrontal cortices via the forceps minor. In contrast with the positive effect in the other tracts, the negative association with the reserve in CC1 shows the reduced sh-peak signal in this region predicts better cognitive maintenance. We note that these effects are observed in regions where fibers from the anterior cingulum bundle cross over the CC genu. How these two fiber systems may

Table 4.3 Selected tracts and vertices

Experiment	Cluster	Selected tract	Vertices	Association sign
With all subjects, Metric used: **sh-peaks**	1	CC-1	55,56,57,58 59,60,61,62	-
	2	right SLF-I	98	+
		left SLF-II	1,2,3,4,5	+
	3	right CST	60,61,62,63	+
Excluding AD_MD, Metric used: **sh-peaks**	1	CC_1	54,55,56,57,58,59 60,61,62,63,64	-
	2	left SLF-I	2,3,4,5	+
	3	right CST	60,61,62,63	+
With educ ≥ 15, Metric used: **sh-peaks**	1	CC-1	56,57,58,59,60	-
	2	right SLF-I	98	+
		left SLF-II	1,2,3,4,5	+
With educ ≤ 17, metric Used: **sh-peaks**	1	right SLF-I	98	+
		left SLF-II	1,2,3,4,5	+
	3	left ATR	8,9,10	-
With all subjects, metric Used: **FA**	1	CC-4	39,40,41,42,58,59,60	-
	2	right ST-PREM	1,2,3,4,5,6	-
			7,8,9, 10,11	-
	3	right AF	30,31,32, 33,34,35	-
	4	right SLF-II	33,34,35,36,37,38	-

interact to inform cognitive reserve remains an area for future inquiry. In addition, limiting the analysis to exclude those with dementia diagnoses resulted in a nearly identical pattern of selected vertices, although it was more sensitive to reserve as reflected in a larger number of significant vertices in CC1. Similarly, the original pattern of results was largely maintained in participants with more years of formal education, although the vertices in the right CST were no longer significant predictors of the reserve. Moreover, limiting the sample to those with less educational attainment revealed a negative effect of sh-peak values in anterior thalamic radiation (ATR) in the left hemisphere on reserve. This tract also crosses CC1-2, cingulum, and inferior frontal-occipital fibers, suggesting a similar cause as for CC1 in the overall model.

The use of the sf-peak values estimated from the diffusion FODs provides a more orientationally specific estimate than voxel-level scalars like tensor-based estimates of anisotropy (FA) and diffusivity. We also note that reserve is a counterintuitive construct, as high levels of reserve are apparent in those with indications of neurodegeneration (i.e., smaller brain volumes) combined with higher levels of memory performance than would be predicted from the linear relationship alone. Moreover, results from sampling and modeling along-tract FA values produced a markedly different pattern of results. All identified tract segments in the FA analysis were negatively associated with reserve, and all are in areas where FA values are confounded by crossing fibers (Douaud et al. (2011); Chad et al. (2021)). Because lower FA values in crossing fiber regions are associated with dementia risk, modeling sh-peaks and FA separately provide complementary insights into the WM tractography correlates of cognitive reserve.

4.4 Conclusion

In this work, we proposed a DL-based multi-resolutional feature selection algorithm tailored for highly correlated ultra-high dimensional feature space. The contributions of our work are twofold: (1) **From the statistical perspective**, the proposed method efficiently combines several existing tools in statistics and ML literature to circumvent some of the limitations of current DL-based models in handling complex data similar to that of the UM-MAP study. Specifically, it achieves significant dimension reduction while maintaining type-I and type-II error trade-offs by efficiently combining the added information from resampling. Due to the screening step, our method is scalable and its resampling component can be easily implemented as parallel chains for faster computation. (2) **From the application perspective**, the proposed method addresses at least two critical shortcomings in the extant literature for handling tractography data in answering important questions in cognitive reserve. First, unlike existing approaches that treat diffusion tractography data as responses, the current analysis permits treating diffusion tractography data as the predictor, rather than the response. As prior methods have focused on binary or continuous predictors of differences in streamline-sampled values, the proposed method improves the validity of modeling streamlined WM estimates as a predictor of behavior. Second, other methods for tractometry cannot model multiple streamlines together, much less the large number reported here. By integrating a more robust approach for controlling for multiple tests and variable selection, this method permits simultaneous modeling of whole brain white matter tractography and discovers multiresolution clusters associated with some neurogenetic disorders. Nevertheless, even this report utilized streamlines reduced to centroids with sampling points limited to 100 per WM tract. Future work is needed to capitalize on the considerably higher dimensionality of these data types as predictors of neurocognitive aging outcomes.

The basic intuition, scalability, and exciting empirical results of the proposed method on simulated and real datasets encourage further research in multiple directions. From a **theoretical** aspect, we are actively working on developing a theoretical foundation of this ’screening’ and ’cleaning’ strategy for provable FDR control. It would be worth mentioning that although we used the sure independence screening with HZ-test and LassoNet as the main tools, the proposed method puts forward a more generic framework and can be implemented with any other model-free feature screening method and sparsity-inducing DL algorithms like Feng and Simon (2017). One limitation we mainly focus on here is that the proposed method is developed considering regression setup with a continuous outcome because of the requirements of the Henze–Zirkler sure Independence test used in the screening step. Further research should be conducted for extending the proposed method to classification problems as well. From an **application perspective**, as the specific parts of the tracts selected by the proposed method have shown to have strong predictive ability, more interpretable nonlinear models (e.g. decision trees or random forest) can be entertained for further analysis. This will further provide deeper insights into the association of these selected tracts on the cognitive reserve, neurodegeneration, and beyond.

BIBLIOGRAPHY

Bastiani, M., Cottaar, M., Dikranian, K., Ghosh, A., Zhang, H., Alexander, D. C., Behrens, T. E., Jbabdi, S., and Sotiropoulos, S. N. (2017). Improved tractography using asymmetric fibre orientation distributions. *Neuroimage*, 158:205–218.

Beaulieu, C. (2002). The basis of anisotropic water diffusion in the nervous system–a technical review. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo*, 15(7-8):435–455.

Bender, A. R., Daugherty, A. M., and Raz, N. (2013). Vascular risk moderates associations between hippocampal subfield volumes and memory. *Journal of cognitive neuroscience*, 25(11):1851–1862.

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

Brandt, J. (1991). The hopkins verbal learning test: Development of a new memory test with six equivalent forms. *Clinical Neuropsychologist*, 5(2):125–142.

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(3):pp. 551–577.

Chad, J. A., Pasternak, O., and Chen, J. J. (2021). Orthogonal moment diffusion tensor decomposition reveals age-related degeneration patterns in complex fiber architecture. *Neurobiology of Aging*, 101:150–159.

Chang, Y.-L., Chao, R.-Y., Hsu, Y.-C., Chen, T.-F., and Tseng, W.-Y. I. (2021). White matter network disruption and cognitive correlates underlying impaired memory awareness in mild cognitive impairment. *NeuroImage: Clinical*, 30:102626.

Chen, Y., Gao, Q., Liang, F., and Wang, X. (2021). Nonlinear variable selection via deep neural networks. *Journal of Computational and Graphical Statistics*, 30(2):484–492.

Colby, J. B., Soderberg, L., Lebel, C., Dinov, I. D., Thompson, P. M., and Sowell, E. R. (2012). Along-tract statistics allow for enhanced tractography analysis. *Neuroimage*, 59(4):3227–3242.

Craft, S., Newcomer, J., Kanne, S., Dagogo-Jack, S., Cryer, P., Sheline, Y., Luby, J., Dagogo-Jack, A., and Alderson, A. (1996). Memory improvement following induced hyperinsulinemia in alzheimer's disease. *Neurobiology of aging*, 17(1):123–130.

Dhollander, T., Clemente, A., Singh, M., Boonstra, F., Civier, O., Duque, J. D., Egorova, N., Enticott, P., Fuelscher, I., Gajamange, S., et al. (2021). Fixel-based analysis of diffusion mri: methods, applications, challenges and opportunities. *Neuroimage*, 241:118417.

Dodge, H. H., Goldstein, F. C., Wakim, N. I., Gefen, T., Teylan, M., Chan, K. C. G., Kukull, W. A., Barnes, L. L., Giordani, B., Hughes, T. M., Kramer, J. H., Loewenstein, D. A., Marson, D. C., Mungas, D. M., Mattek, N., Sachs, B. C., Salmon, D. P., Willis-Parker, M., Welsh-Bohmer, K. A., Wild, K. V., Morris, J. C., Weintraub, S., and National Alzheimer's Coordinating Center (NACC) (2020). Differentiating among stages of cognitive impairment in aging: Version 3 of the uniform data set (UDS) neuropsychological test battery and MoCA index scores. *Alzheimers Dement. (N. Y.)*, 6(1):e12103.

Douaud, G., Jbabdi, S., Behrens, T. E., Menke, R. A., Gass, A., Monsch, A. U., Rao, A., Whitcher, B., Kindlmann, G., Matthews, P. M., and Smith, S. (2011). Dti measures in crossing-fibre areas: Increased diffusion anisotropy reveals early white matter alteration in mci and mild alzheimer's disease. *NeuroImage*, 55(3):880–890.

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

Feng, J. and Simon, N. (2017). Sparse-input neural networks for high-dimensional nonparametric regression and classification. *arXiv preprint arXiv:1711.07592*.

Ghorbani, A., Abid, A., and Zou, J. (2019). Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3681–3688.

Jordon, J., Yoon, J., and van der Schaar, M. (2019). KnockoffGAN: Generating knockoffs for feature selection using generative adversarial networks. In *International Conference on Learning Representations*.

Koshiyama, D., Fukunaga, M., Okada, N., Morita, K., Nemoto, K., Yamashita, F., Yamamori, H., Yasuda, Y., Matsumoto, J., Fujimoto, M., Kudo, N., Azechi, H., Watanabe, Y., Kasai, K., and Hashimoto, R. (2020). Association between the superior longitudinal fasciculus and perceptual organization and working memory: A diffusion tensor imaging study. *Neuroscience Letters*, 738:135349.

Lei, L. and Fithian, W. (2018). Adapt: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):649–679.

Li, A. and Barber, R. F. (2018). Multiple testing with the structur... adaptive benjaminihochberg algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(10).

Lu, Y., Fan, Y., Lv, J., and Stafford Noble, W. (2018). Deeppink: reproducible feature selection in deep neural networks. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

McPhee, G. M., Downey, L. A., and Stough, C. (2019). Effects of sustained cognitive activity on white matter microstructure and cognitive outcomes in healthy middle-aged adults: A systematic review. *Ageing Res. Rev.*, 51:35–47.

Min, Z.-G., Rana, N., Niu, C., Ji, H.-M., and Zhang, M. (2014). Does diffusion tensor tractography of the corticospinal tract correctly reflect motor function? *Med. Princ. Pract.*, 23(2):174–176.

Raffelt, D. A., Tournier, J.-D., Smith, R. E., Vaughan, D. N., Jackson, G., Ridgway, G. R., and Connelly, A. (2017). Investigating white matter fibre density and morphology using fixel-based analysis. *Neuroimage*, 144:58–73.

Raz, N., Lindenberger, U., Rodrigue, K. M., Kennedy, K. M., Head, D., Williamson, A., Dahle, C., Gerstorf, D., and Acker, J. D. (2005). Regional brain changes in aging healthy adults: general trends, individual differences and modifiers. *Cerebral cortex*, 15(11):1676–1689.

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

Sesia, M., Katsevich, E., Bates, S., Candès, E., and Sabatti, C. (2019). Multi-resolution localization of causal variants across the genome. *bioRxiv*.

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

Song, Z. and Li, J. (2021). Variable selection with false discovery rate control in deep neural networks. *Nature Machine Intelligence*, 3(5):426–433.

Stern, Y., Arenaza-Urquijo, E. M., Bartrés-Faz, D., Belleville, S., Cantilon, M., Chetelat, G., Ewers, M., Franzmeier, N., Kempermann, G., Kremen, W. S., et al. (2020). Whitepaper: Defining and investigating cognitive reserve, brain reserve, and brain maintenance. *Alzheimer's & Dementia*, 16(9):1305–1311.

Tournier, J.-D., Calamante, F., and Connelly, A. (2012). Mrtrix: diffusion tractography in crossing fiber regions. *International journal of imaging systems and technology*, 22(1):53–66.

Tournier, J.-D., Smith, R., Raffelt, D., Tabbara, R., Dhollander, T., Pietsch, M., Christiaens, D., Jeurissen, B., Yeh, C.-H., and Connelly, A. (2019). Mrtrix3: A fast, flexible and open software framework for medical image processing and visualisation. *NeuroImage*, 202:116137.

Uddin, L. Q. (2021). Cognitive and behavioural flexibility: neural mechanisms and clinical considerations. *Nature Reviews Neuroscience*, 22(3):167–179.

Vos, S. B., Jones, D. K., Viergever, M. A., and Leemans, A. (2011). Partial volume effect as a hidden covariate in DTI analyses. *Neuroimage*, 55(4):1566–1576.

Wang, F., Ren, S.-Y., Chen, J.-F., Liu, K., Li, R.-X., Li, Z.-F., Hu, B., Niu, J.-Q., Xiao, L., Chan, J. R., and Mei, F. (2020a). Myelin degeneration and diminished myelin renewal contribute to age-related deficits in memory. *Nat. Neurosci.*, 23(4):481–486.

Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020b). A simple new approach to variable selection in regression, with application to genetic fine-mapping. *bioRxiv*.

Wasserthal, J., Maier-Hein, K. H., Neher, P. F., Northoff, G., Kubera, K. M., Fritze, S., Harneit, A., Geiger, L. S., Tost, H., Wolf, R. C., et al. (2020). Multiparametric mapping of white matter microstructure in catatonia. *Neuropsychopharmacology*, 45(10):1750–1757.

Wasserthal, J., Neher, P., and Maier-Hein, K. H. (2018a). Tractseg - fast and accurate white matter tract segmentation. *NeuroImage*, 183:239–253.

Wasserthal, J., Neher, P., and Maier-Hein, K. H. (2018b). Tractseg-fast and accurate white matter tract segmentation. *NeuroImage*, 183:239–253.

Wasserthal, J., Neher, P. F., Hirjak, D., and Maier-Hein, K. H. (2019). Combined tract segmentation and orientation mapping for bundle-specific tractography. *Medical image analysis*, 58:101559.

Xia, F., Zhang, M. J., Zou, J. Y., and Tse, D. (2017). Neuralfdr: Learning discovery thresholds from hypothesis features. In *NIPS*.

Xie, L., Wisse, L. E. M., Das, S. R., Wang, H., Wolk, D. A., Manjón, J. V., and Yushkevich, P. A. (2016). Accounting for the confound of meninges in segmenting entorhinal and perirhinal cortices in t1-weighted MRI. *Med. Image Comput. Comput. Assist. Interv.*, 9901:564–571.

Yushkevich, P. A., Pluta, J. B., Wang, H., Xie, L., Ding, S.-L., Gertje, E. C., Mancuso, L., Kliot, D., Das, S. R., and Wolk, D. A. (2015). Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment. *Hum. Brain Mapp.*, 36(1):258–287.

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.

APPENDIX A

MADC TRACTOGRAPHY DATA - MORE INFORMATION

A.1 Study sample.

Data were drawn from the University of Michigan Memory and Aging Project (UM-MAP), the primary clinical cohort at the MADRC. The sample included 221 participants (67% women) from 51 to 89 years of age. A consensus diagnosis was made following neuropsychological evaluation using the National Alzheimer's Coordinating Center (NACC) criteria by neurologists, neuropsychologists, nurses, social workers and other specialists during a consensus conference. The sample was divided into three subgroups based on the last recorded diagnosis for each participant (Table 1): cognitively unimpaired (CU; n=117; 73% women), amnestic or non-amnestic MCI (MCI; n=62; 70% women) and multi-domain amnestic dementia (DAT; n=42; 55% women) consistent with Alzheimer's disease and mixed dementia.

A.1.1 MRI acquisition.

All neuroimaging data were acquired on a 3 Tesla General Electric Discovery Magnetic Resonance System equipped with a 32-channel receiving/transmitting head coil at the University of Michigan's Functional MRI Laboratory. T1-weighted structural images were collected with the following parameters: TR=3173.1 ms; TE=24.0 ms; inversion time=896 ms; flip angle=111○; FOV=220×220 mm; 43 axial slices with thickness=3 mm and no spacing; acquisition time=100 s. A diffusion-weighted 2D dual spin echo pulse sequence was acquired in 81 axial slices with voxel dimensions of 1.7 mm3, repetition time (TR)=4100 ms; echo time (TE)=2.5 ms; field of view (FOV)=240×240 mm. We acquired 96 volumes including 6 without diffusion weighting (b=0), 30 volumes with weightings of b=700 s/mm2, and 60 volumes with b=2000 s/mm2, for a total of 96 gradient encoding directions. In addition, a 2D spin echo field map image was acquired with the same dimensions and was applied to create a reverse phase encoded b0 image for distortion corrections.

A.2 MRI processing.

Image processing utilized the high-performance computing cluster at Michigan State University, with Intel Xeon Gold 6148 CPU cluster nodes. All dMRI data pre-processing and processing closely followed the steps for multi-shell multi-tissue (MSMT) constrained spherical deconvolution and fixel-based analysis published in the MRtrix user manual (available here). These included algorithmic preprocessing steps for mitigation of thermal noise, Gibbs ringing artifacts, nonlinear distortions from motion artifacts and eddy currents, as well as intensity bias (Raffelt et al. (2017); Tournier et al. (2019, 2012)). All pre-processed data were upsampled to a voxel dimension of 1.25 mm3. Individual MSMT response functions were estimated for each upsampled volume using the Dhollander algorithm (Dhollander et al. (2021)). The individual response functions were averaged, and the mean values were used to compute fiber orientation distributions (FODs) in each voxel. A subset of FODs from 59 cases, equally divided between diagnostic groups, was used to compute a sample-specific FOD template and a template mask; all cases were subsequently nonlinearly spatially transformed to the template. Next, we estimated the spherical harmonic peak amplitudes (sh-peaks) in FODs from each template-registered case and the FOD template. We used the dwi2tensor and tensor2metric functions in MRtrix to compute tensors in the upsampled, preprocessed dMRI volumes and estimate individual FA images. We used the transformation functions previously calculated for registering FODs to the template to transform the upsampled FA data to the group FOD template. Using the TractSeg 2.3 algorithm (Wasserthal et al. (2018b, 2019, 2020)), we estimated 72 white matter fiber bundles using the group template sh-peaks; the TractSeg tractometry function was used to resampled each fiber bundle to a single centroid streamline and distributed 100 equally spaced sampling points across the length of the representative streamline. For each representative streamline, we sampled the sh-peak and FA values under each sampling point for each participant.

Automated segmentation of hippocampal subfields (ASHS; Yushkevich et al. (2015)) is a multi-atlas label fusion method for determining the anatomical boundaries and quantifying volumes in specific regions of the brain's medial temporal lobe. We used the ASHS-PMC-T1 atlas (Xie et al. (2016)) developed for segmenting bilateral volumes of the anterior and posterior hippocampi and entorhinal cortices from conventional T1-weighted structural MRI data. All segmentations were inspected by trained laboratory personnel for issues with segmentation quality. The volume of the intra-cranial vault estimated by ASHS was used to statistically adjust output volumes for differences in participants' head sizes using the ANCOVA method (Bender et al. (2013); Raz et al. (2005)).

A.3 Cognitive testing.

Neuropsychological tests from the Uniform Data Set 3 (UDS3) included measures of delayed recall from the Hopkins Verbal Learning Test (HVLT) and Craft Story Test (Brandt (1991); Craft et al. (1996)). Measures of delayed recall in these tests of episodic memory are sensitive to amnestic MCI and preclinical Alzheimer's disease (Dodge et al. (2020)).

A.4 Cognitive reserve.

We standardized and averaged the scores for delayed recall from the Craft Story and HVLT tasks to create a composite of episodic memory. Similarly, we averaged z-scores for volumes of bilateral entorhinal cortices and anterior and posterior hippocampi to generate a brain volume composite score. The memory composite was regressed on the brain volume composite and the residuals served as the measure of cognitive reserve.

APPENDIX B

LIST OF MAJOR WM TRACTS

The list of 50 major WM tracts considered in this current study is given below. More pictorial demonstration of these tracts and related information can be found that Wasserthal et al. (2018a).

AF_left, AF_right, ATR_left, ATR_right, CC_1, CC_2, CC_3, CC_4, CC_5, CC_6, CC_7, CG_left, CG_right, CST_left, CST_right, FPT_left, FPT_right, ICP_left, ICP_right, IFO_left, IFO_right, ILF_left, ILF_right, MCP, OR_left, OR_right, POPT_left, POPT_right, SCP_left, SCP_right, SLF_I_left, SLF_I_right, SLF_II_left, SLF_II_right, SLF_III_left, SLF_III_right, STR_left, STR_right, UF_left, UF_right, T_PREM_left, T_PREM_right, T_PAR_left, T_PAR_right, T_OCC_left, T_OCC_right, ST_FO_left, ST_FO_right, ST_PREM_left, ST_PREM_right

CHAPTER 5

CONCLUSIONS

In this thesis, we have explored advanced statistical methodologies for feature selection in ultra-high dimensional datasets, specifically focusing on the analysis of tractography data in the context of the UM-MAP study. Our research journey has been guided by the need to address the challenges posed by high dimensionality, strong associations within and between tract-level measurements, and the complex functional relationship between these measurements and phenotype.

Through an iterative research process, we started with a linear model-based approach and extended it to develop SciDNet, a deep learning-based method, effectively combining the power of statistical inference and the flexibility of neural networks. We have demonstrated the effectiveness of the proposed methods in achieving higher power and controlled false discovery rate (FDR) compared to other state-of-the-art methods through extensive empirical studies and simulations. We have also examined the assumptions and algorithms underlying SciDNet, providing theoretical justification and practical insights into its application.

Our findings have significant implications for understanding the micro-structural relationships in the human brain, particularly in the context of dementia and beyond. By providing a method that is independent of strict modeling assumptions and p-values, SciDNet contributes to a better understanding of the intricate connections between DMRI metrics and phenotype.

In conclusion, this thesis has addressed critical challenges in the analysis of tractography data by proposing and validating the SciDNet method. The insights gained from this research will pave the way for further theoretical investigations, generalizability studies, and broader applications. We are confident that the methodologies developed in this thesis will contribute to advancements in the field of high-dimensional statistics and enhance our understanding of complex brain connectivity patterns.

We would like to express our gratitude to all the individuals who have supported and encouraged us throughout this research endeavor. Their contributions, guidance, and belief in our abilities have been invaluable. This thesis marks the culmination of years of hard work, dedication, and collaboration, and we are grateful for the opportunity to contribute to the field of statistical analysis in such a meaningful way.

As we conclude this chapter, we look forward to future research and collaborations that will build upon the foundations laid in this thesis. The journey does not end here; it continues with new questions, challenges, and discoveries that will shape the future of statistical analysis in the realm of high-dimensional datasets.

Thank you for joining us on this intellectual exploration and for being a part of this remarkable journey.